UNFOLDING SHARED SEMANTIC REPRESENTATIONS IN VISION AND LANGUAGE

by

KATERINA MARIE SIMKOVA

A thesis submitted to the University of Birmingham for the degree of DOCTOR OF PHILOSOPHY

School of Psychology
College of Life and Environmental Sciences
University of Birmingham
October 2024

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

Abstract

When we interact with language, the brain effortlessly draws vivid mental images that resemble representations of directly perceived visual scenes. Similarly, research shows that the visual system encodes representations of visual scenes in a manner akin to how embeddings from large language models encode the meaning of a sentence. This implies that vision and language draw on a shared representational system for semantic encoding, but the precise nature of this visuo-linguistic overlap in the brain, as well as its implications for behaviour, remain unknown. The present thesis leveraged similarity judgements of natural scene images and their corresponding sentence captions to explore how vision and language converge in their ability to recognise semantic regularities of the visual world. We found that the visual and linguistic similarity judgements not only converge in the behavioural context but both also predict a remarkably similar network of visually evoked patterns along the mid- and high-level visual regions. Furthermore, we established the behavioural relevance by showing that the perceived dissimilarities in both vision and language effectively predict response times to a visual target primed by a sentence. Upon the observation that linguistic similarity judgements exhibit a higher interindividual variability, we used the participants' own perceived dissimilarities to make predictions about their priming effect. This revealed that idiosyncratic linguistic dissimilarity is the strongest predictor of participants' response times. The findings of the present thesis demonstrate that vision and language project their representations onto a shared similarity space, most likely embedded in the representational structure of the visual brain. Along with the observation that language plays a distinct role in eliciting individually relevant representations, these findings may profoundly impact our understanding of the visual system by placing greater emphasis on semantic idiosyncrasies.

Acknowledgements

I would like to express my sincere thanks to my supervisor Dr Ian Charest. Thank you for encouraging me and showing up for me even through lockdowns and long-distance supervision. Ian's help and creative ideas have greatly contributed to this work.

Special thanks to Dr Clayton Hickey for support and genuine concern for my academic and personal wellbeing. Thank you for adopting me into your lab and motivating me throughout the writing up process of my PhD.

I would like to also thank everyone at the School of Psychology, University of Birmingham who helped me to obtain funding for the third year of my PhD, which made it possible for me to complete this project.

Finally, thank you to all my incredible salseros and salseras. Dancing with you has been my escape and a constant reminder to keep moving no matter how difficult the steps. My heartfelt thank you to Jasper for being vision to my language, in research and in life.

About this thesis

This thesis incorporates the three following papers that correspond to the empirical studies:

Simkova, K. M., Doerig, A., Hickey, C., & Charest, I. (in prep.) Representations in vision and language converge in a multidimensional space of perceived dissimilarities.

Simkova, K. M., van den Bosch, J. J. F., Hickey, C., & Charest, I. (in prep.) The semantic distance between a linguistic prime and a natural scene target predicts reaction times in a cross-modal priming experiment.

Simkova, K. M., van den Bosch, J. J. F., Hickey, C., & Charest, I. (in prep). Representational dynamics of vision and language in perception and imagery.

The following conference abstracts were published from the findings in this thesis:

Simkova, K. M., van den Bosch, J. J. F., & Charest, I. (2022). The role of semantics in similarity judgements of scene stimuli. *2022 Conference on Cognitive Computational Neuroscience.*

Simkova, K. M., van den Bosch, J. J. F., Grignolio, D., Hickey, C., & Charest, I. (2023). The semantic distance between a linguistic prime and a natural scene target predicts reaction times in a visual search experiment. *Journal of Vision, 23*(9), 5055-5055.

Simkova, K. M., van den Bosch, J. J. F., Hickey, C., & Charest, I. (2024). Linguistic and visual similarity judgements predict EEG representational dynamics in visual perception and sentence reading. *Journal of Vision, 24*(10), 1198-1198.

Table of contents

Chapter 1 13

General introduction 13

1.1 From retinal projection to semantic representation: how do we know what we see? 15

1.2 The co-occurrence hypothesis: “Neurons that fire together, wire together” 18

1.3 Semantic memory as the underpinning mechanism of the cross-modal integration 22

1.4 Finding the best model of the common representational space 27

1.5 References 30

Chapter 2 43

Representations in vision and language converge in a multidimensional space of perceived dissimilarities 43

2.1 Abstract 44

2.2 Introduction 45

2.3 Results 46

2.4 Discussion 62

2.5 Methods 67

2.6 References 74

2.7 Supplementary material 82

Chapter 3 98

The semantic distance between a linguistic prime and a natural scene target predicts reaction times in a cross-modal priming experiment 98

3.1 Abstract 99
3.2 Introduction 100
3.3 Material and methods.....102
3.4 Results 114
3.5 Discussion 123
3.6 References.....129
3.7 Supplementary material 136
Chapter 4 138
Representational dynamics of vision and language in perception and imagery 138
4.1 Abstract 139
4.2 Introduction 140
4.3 Results 142
4.4 Discussion 159
4.5 Materials and methods.....166
4.6 References.....176
4.7 Supplementary material 187
Chapter 5 190
General discussion 190
5.1 Representations in vision and language converge in their ability to encode regularities of the visual world 192
5.2 Stability of the representational similarity space as a function of task demands.....195
5.3 Is the shared representational space grounded in the visual cortex? 198

5.4 Suggestions for future work 199

5.5 Concluding remarks 201

5.6 References 203

List of figures

Figure 2.1 Experimental design of the MA tasks. 48

Figure 2.2 Representational alignment of multimodal similarity judgements with brain responses to viewing natural scene images. 52

Figure 2.3 Representational alignment of multimodal similarity judgements and LLM representations. 54

Figure 2.4 Behavioural analysis to characterise representational alignment in visual and linguistic similarity judgements. 61

Figure 3.1 Study design in Experiment 1. 105

Figure 3.2 Study design in Experiment 2. 107

Figure 3.3 Cross-modal priming effect predicted by NSD representational dissimilarities. 115

Figure 3.4 Cross-modal priming effect predicted by averaged linguistic and visual dissimilarities. 117

Figure 3.5 Cross-modal priming effect predicted by idiosyncratic dissimilarities. 118

Figure 3.6 Cross-modal priming effect in low and high distractor-target dissimilarity trials. 123

Figure 4.1 EEG experimental design. 143

Figure 4.2 Neural representational dynamics compared to the weighted similarity model of behavioural similarity judgements. 146

Figure 4.3 Representation of mid- and high-level features in neural geometry of multimodal perception and imagery. 151

Figure 4.4 Categorical representations in visual perception and imagery.....155

Figure 4.5 Categorical representations in sentence reading and linguistic imagery. 157

Figure 4.6 Generalisation of categorical representation for people and outdoor scenes across time and modalities. 159

Supplementary Figure 2.1 Experimental design of the control MA task. 82

Supplementary Figure 2.2 Representational alignment of multimodal similarity judgements with brain responses at the participant-level. 83

Supplementary Figure 2.3 Representational alignment between the brain responses and multimodal similarity judgements from session 1. 85

Supplementary Figure 2.4 Representational alignment between the brain responses and multimodal similarity judgements from session 1 at the participant-level.....87

Supplementary Figure 2.5 Representational alignment between the brain responses and multimodal similarity judgements from session 1 using Pearson correlations. 88

Supplementary Figure 2.6 Representational alignment between the brain responses and multimodal similarity judgements from session 1 participant-level using Pearson correlation. 90

Supplementary Figure 2.7 Pairwise difference in RCNN prediction accuracies.. .91

Supplementary Figure 2.8 Pairwise difference in RCNN and Alexnet prediction accuracies. 92

Supplementary Figure 2.9 Cross-session within-modality comparisons. 93

Supplementary Figure 2.10 Interindividual variability tends to display the opposite effect when controlled for prior exposure. 95

Supplementary Figure 2.11 Covert visual modality compared to behaviour, brain, and models. 96

Supplementary Figure 3.1 LME predictions using permuted idiosyncratic linguistic RDMs.. 136

Supplementary Figure 4.1 Multiclass decoding in multimodal perception and imagery. 187

Supplementary Figure 4.2 Generalisation of categorical representations from visual perception and sentence reading.....188

Supplementary Figure 4.3 Generalisation of categorical representations across time and modalities.. 189

List of tables

Table 3.1. The effect of prime-target dissimilarity and its interaction with distractor-target dissimilarity on log-transformed response times.....120

Table 3.2. The effect of prime-target and distractor-target dissimilarity and their interaction in response times. 122

List of abbreviations

DNNs	deep neural networks
EBA	extrastriate body area
EEG	electroencephalography
EVC	early visual cortex
FDR	false discovery rate
FFA	fusiform face area
fMRI	functional magnetic resonance imaging
IT	inferotemporal cortex
ITS	inferior temporal sulcus
LDA	linear discriminant analysis
LGN	lateral geniculate nucleus
LLMs	large language models
LME	linear mixed effects model
MA	multiple arrangements
MDS	multidimensional scaling
MEG	magnetoencephalography
MS COCO	Microsoft Common Objects in Context
MVPA	multivariate pattern analysis
NNLs	non-negative least squares
NSD	Natural Scenes Dataset
PPA	parahippocampal place area
RCNN	recurrent convolutional network
RDM	representational dissimilarity matrix
RSA	representational similarity analysis
RT	response time
STS	superior temporal sulcus
VVIQ	vividness of visual imagery questionnaire
VWFA	visual word form area

Chapter 1

General introduction

In 1970, Shepard and Chipman asked participants to judge the similarity of 15 pictorially displayed shapes, and written names of the U.S. states. They found that the similarity relationships between the imagined representations triggered by words mirrored the relationships between the directly perceived shapes. For this to happen, there must be some resemblance between the imagined representations in language and the corresponding objects. Shepard and Chipman (1970) argued that although it may seem as though we are imagining a green square, it does not imply that a literal green square occurs in our mind. What they proposed instead was that the functional relationships between internal representations reflect the relational structure of corresponding objects in the real world. Their proposal, however, was discredited by authors who were influenced by the early twentieth-century behaviourists (Watson, 1913). The main critique was that the study of mental representation is unscientific (Barsalou, 2008) because conscious introspection is not reflective of the underlying processes that give rise to our thoughts (Pylyshyn, 1973). As a result, the second half of the twentieth century was dominated by the view that cognitive processes, such as language, are better understood through propositional representations that capture cognition in a more abstract, symbolic manner. This view was concurrent to the development of computer-based simulations where such abstract symbols were thought to be a powerful tool to represent logical operations carried out by our mind.

It was not until quite recently that the view pioneered by Shepard and Chipman (1970) received a wide range of support. Research showed that similarity structure derived from the external world is not only predictive of the representational architecture in the brain (Charest et al., 2014; Kriegeskorte et al., 2008b; Mur et al., 2013) but also guides overt behaviour (Cohen et al., 2015; Cohen et al., 2017; Yeh et al., 2024). There is now ample evidence that illustrates how semantic representations derived from language benefit from grounding into

visual information (Andrews et al., 2009; Silberer, & Lapata, 2014; Wang et al., 2023), calling for further research into how vision and language converge onto a shared representational space. The introduction chapter begins with a brief overview of the transformations carried out by the visual ventral pathway. Next, we explore underlying dimensions of similarity, highlighting its role as a key organising principle of the visual cortex. We then examine the parallels between the distributed coding of the visual system and predictive coding in language comprehension. Finally, traditional views of semantic memory as a store of amodal symbols are challenged, leading to a proposal for a more plausible model of visuo-linguistic integration that considers growing evidence from computational cognitive neuroscience.

1.1 From retinal projection to semantic representation: how do we know what we see?

The first task that the visual brain must solve is to recognise that even if each encounter with an object activates an entirely different retinal pattern (DiCarlo et al., 2012), all these responses belong to a representation of the same object. This challenge, known as the view-invariant problem (DiCarlo, & Cox, 2007; DiCarlo et al., 2012; Kravitz et al., 2008), unfolds from the first moment that the changes in light and colour hit our retina. The neural signals initially travel to the contralateral hemisphere through the optic nerve and, from the lateral geniculate nucleus (LGN), the signals are then projected onto the early visual cortex (EVC). The activation of EVC neurons corresponds to their exact location in the visual field – this retinotopic organisation enables the visual cortex to preserve information about the spatial relationships just as they appear in the real world (Felleman, & van Essen, 1991; Hubel, & Wiesel, 1962; Hubel, & Wiesel, 1968; Logothetis et al., 1994). These features are then integrated into a unified representation, which is further propagated from the EVC to the anterior parts of the brain through a

hierarchically organised stream of increasing complexity (DiCarlo et al., 2012; Felleman, & van Essen, 1991). An influential model that describes the hierarchical stream suggests functional segregation in two[1] distinct cortical pathways: the ventral 'what' pathway for object recognition and the dorsal 'where' pathway for spatial information (Goodale, & Milner, 1992; Ungerleider, & Haxby, 1994; Ungerleider, & Mishkin, 1982). Some evidence challenged this functional segregation on the grounds that the dorsal pathway not only interconnects with the ventral pathway (Andersen et al., 1990), but even encodes shape properties (Lehky, & Sereno, 2007). What makes the ventral pathway significantly more powerful in object discrimination, however, is its ability to encode the rapidly changing details of colour, shape, and direction (Goodale et al., 1991; Goodale, & Milner, 1992; Jakobson et al., 1991; Lehky, & Sereno, 2007). The object features are then propagated through the ventral stream and projected to the inferotemporal (IT) cortex where an ensemble of response patterns gives rise to a semantic representation that enables us to understand and reason about what we see (DiCarlo et al., 2012; Kravitz et al., 2008). Despite the complexity of these neural computations, in less than 200 ms (Kirchner, & Thorpe, 2006; Potter, 1976; Thorpe et al., 1996), the brain arrives at an accurate conclusion that the differential retinal projections all belong to a representation of a single object.

1.1.1 Functional organisation of the visual ventral pathway

It is now well-established that object representations in the visual ventral pathway emerge from an ensemble of response patterns in regions that represent features of preferential category knowledge. This was first demonstrated in case studies where individuals with isolated lesions along the visual ventral pathway displayed selective impairments in object recognition. For

[1] Recent work provides evidence for a third visual pathway specialised for social perception. See Pitcher, & Ungerleider, 2022 for a review.

example, McNeil and Warrington (1993) described a case of patient WJ with lesions in the occipital and temporal lobes. WJ exhibited severe prosopagnosia (i.e., face recognition impairment) while retaining their ability to recognise other visually ambiguous objects. Others reported cases of landmark agnosia where individuals had difficulty recognising familiar landmarks, but their ability to orient their bodies to those landmarks remained intact (Aguirre et al., 1998; Levine et al., 1985; Whiteley, & Warrington, 1978). This converging evidence led subsequent research to investigate the possibility that the visual ventral pathway consists of modules specialised for processing distinct category knowledge. Brain imaging research in healthy individuals, indeed, identified several discrete modules including the parahippocampal place area (PPA), fusiform face area (FFA), and extrastriate body area (EBA) for processing of scenes, faces, and body parts respectively (Aguirre et al., 1998; Epstein & Kanwisher, 1998; Kanwisher et al., 1997; McCarthy et al., 1997). Although the existence of discrete modules provided strong justification as to why isolated lesions cause selective impairments to category recognition, it sparked many controversies. Some authors raised concerns that the findings arose from a presumed category structure rather than data-driven methods (Kiani et al., 2007) while others argued that the number of object categories needed to be represented exceeds the available cortical space (Ishai et al., 1999). Cell recordings from the monkey visual cortex revealed that cells that cluster together also commonly respond to similar features, but their selectivity is not identical (Tanaka, 1996). This observation triggered broader questions about whether the visual encoding along the visual pathway is truly modular, or whether object representations emerge from a distributed pattern of neural activations.

The advent of multivariate analysis techniques has made it possible to study the granularity of object representations and challenged the modular view of the visual pathway. Ishai and colleagues (1999) demonstrated that many of the discrete modules are also responsive to visual inputs from non-preferential

categories. These modules were found to carry sufficient information about the non-preferential categories for them to be classified using multivariate pattern analysis (MVPA) techniques. Object representations in the brain are therefore not restricted to discrete modules but rather emerge from a distributed pattern of neural activations where distinct channels encode various features of the attended object (Ishai et al., 1999; Haxby et al., 2001; Kiani et al., 2007). As suggested by Haxby and colleagues (2001), such features are not necessarily related to the visual appearance of the objects because category information was detectable even though participants attended to stimuli stripped of colour (e.g., furry texture is more representative of a cat than a shoe). In fact, Ishai and colleagues (1999) suggested that the increasing complexity of neural computations in the ventral pathway transforms the visual representation into primitives of interpretable attributes that "embody distinctions between object representations". Such feature coding tackles the previously raised concerns about the limited cortical space (Ishai et al., 1999) because each combination of active channels – each responsive to selective attributes – produces an unlimited variety of unique object representations (Ishai et al., 1999; Hanson et al., 2004; Haxby et al., 2001). This raises an important question: what aspects are responsible for such a distributed coding of the visual cortex?

1.2 The co-occurrence hypothesis: "Neurons that fire together, wire together"

It was long believed that specialised functional areas for object recognition have a strong evolutionary origin where the ability to efficiently distinguish between animate and inanimate objects was essential for survival (Caramazza, & Shelton, 1998). However, several authors introduced an intriguing twist when they identified the visual word form area (VWFA) located in the occipitotemporal cortex, which preferentially responds to letters and digits (Allison et al., 1994; Polk,

& Farah, 1998; Shallice, & Saffran, 1986; Warrington, & Shallice, 1980). Aguirre and colleagues (1998) argued that the "development of an area specialised for processing culturally developed symbols cannot be attributed to a similar evolutionary origin as it is the case with the animacy distinction". Instead, as Aguirre and colleagues (1998) noted, the VWFA would develop because letters and digits commonly co-occur in natural environments. Motivated by the correlation learning theory, neural cells that are involved in processing these symbols would therefore become strongly interconnected after a time of repeated co-activation by similar symbols (Hebb, 1949; Polk, & Farah, 1995). In fact, the entire visual world follows certain rules where the brain's adaptation to them must play a key role in how we understand and navigate natural environments (Biederman et al., 1982). One of such regularities is in reference to physical space, indicative of an object's typical position in the visual field (Kaiser et al., 2019). Research shows that typical retinal locations create certain expectations about where objects are through neural tuning (Golomb, & Kanwisher, 2012; Gomez et al., 2018; Gomez et al., 2019; Kaiser, & Cichy, 2018a). If these expectations are violated, visual search in such scenes results in reduced accuracy and slower performance (de Haas et al., 2016; Kaiser, & Cichy, 2018b). It also matters how objects semantically relate to one another with several authors reporting that object recognition and categorisation are substantially faster and more accurate if a visual scene is contextually consistent (Biederman et al., 1982; Davenport, & Potter, 2004; Palmer, 1975). Such regularities of the visual world could be encapsulated by a more general mental concept: similarity.

1.2.1 Similarity as a key organising principle of the brain and thought

Similarity is a mental construct that describes the degree to which objects share features, encompassing natural associations between objects such as shape, colour, texture, function, or general semantics (Cichy et al., 2019; Contier

et al., 2024; Hebart et al., 2020; Tversky, 1977). Several studies have expanded on the distributed view of the visual pathway by using formal similarity measures to model the geometry of the neural representational space. From this perspective, each neural representation is represented as a point in space where the distance between these points indicates the dissimilarity of their neural response patterns (Kriegeskorte et al., 2008a). It was found that if neural response patterns represent objects with many similar features, their representations in the geometry tend to cluster together, indicating lower levels of dissimilarity between them. However, the more distinct the object features become, the more dissimilar neural patterns represent them and the further apart they are from each other (Kriegeskorte et al., 2008b; Kiani et al., 2007; Mur et al., 2013). The representational geometry of the visual pathway was also found to intrinsically predict visual search behaviour: if participants are asked to search for a car among phones, then processing of the car is impaired because both object categories compete for a similar neural representation (Carlson et al., 2014; Cohen et al., 2014; Cohen et al., 2015; Cohen et al., 2017; Ritchie et al., 2015; Yeh et al., 2024). Searching for a car among faces is, however, substantially faster because each of the objects is represented by a distinct neural response pattern, enabling simultaneous processing. Mur and colleagues (2013) identified almost identical representational geometry after asking human participants to arrange images on a two-dimensional plane according to their similarity. These findings collectively support what Shepard and Chipman (1970) initially proposed: it seems apparent that internal representations mirror the similarity relations of the external world. What remains less clear, however, is whether the internal concept of similarity transfers across modalities. Shepard and Chipman (1970) used names of the states to compare similarity relations of internal representations to relations of directly perceived stimuli. A problem with this approach, however, is that the

internal representation was most likely elicited by a clear visual memory of the states' shapes, leaving the complexity of language per se unaccounted for.

1.2.2 How do the regularities of the visual world affect visuo-linguistic integration?

Draschkow and Võ (2017) are the main protagonists in describing the structure of the visual world through linguistic jargon: syntactic rules refer to physical constraints where objects have strictly defined spatial relations, while semantic rules refer to typical context any object belongs to. They showed that a model of word frequency predicts visual scenes significantly better if the objects in them are arranged according to such rules. For example, the model performed better if a laptop was placed in an office rather than a bathroom, which would go against the semantic rules of the visual world. Because language, similarly to vision, derives from rule-governed elements (Draschkow, & Võ, 2017), it is an effective tool in encoding regularities of stable and predictable environments. Bonner, & Epstein (2021) showed that models that learn distributional statistics of language are excellent predictors of object-evoked responses in functional magnetic resonance imaging (fMRI) and magnetoencephalography (MEG). More recently, Doerig and colleagues (2024) showed that sentence embeddings from large language models (LLMs) outperform traditional models of human vision at predicting brain response patterns to viewing natural scene images.

As we already established, human vision is efficient at complex neural computations that go beyond object categorisation to capturing relations of one object to another. What this suggests is that there may be some correspondence between the linguistic form as a culturally developed construct to describe the regularities of the visual world and the way the visual pathway transforms its sensory inputs. Many questions, however, remain unanswered: does the internal

geometry of visual representational space generalise to other modalities, such as natural language processing? How does the representational geometry enable cross-modal integration? What is its behavioural relevance?

1.3 Semantic memory as the underpinning mechanism of the cross-modal integration

1.3.1 Language processing

Arriving at a semantic representation in language involves a series of combinatorial and predictive processes that allow us to recognise words as components of a meaningful structure, grounded in our non-linguistic knowledge of the world (Kamide et al., 2003; Perfetti, 2000). This process begins with lexical access, as prosodic (e.g., stress, intonation, and pauses in speech) and orthographic (e.g., the visual appearance of words in reading) cues signal word and phrase changes critical for syntactic processing (Cutler, & Clifton, 2000; Kjelgaard, & Speer, 1999; Luo, & Zhou, 2010). Once these cues converge, according to the concurrent activation model (Cutler, & Clifton, 2000), the brain activates potential word candidates from memory, continuously updating its hypothesis of the sentence's meaning with each incoming word. The most influential model of semantic integration – the predictive processing framework – posits that our understanding of language is based on informed predictions of sentence meaning, relying on the semantic relatedness and associations between words (Federmeier and Kutas, 2001; Heilbron et al., 2022; Kuperberg, & Jaeger, 2016; Schrimpf et al., 2021; Levy, 2008). For example, if a text says that an alarm went off in the building, a sentence ending with "everyone immediately rushed towards the exit" is semantically more congruent than a sentence ending with "everyone slowly walked towards the exit". Studies show that words and phrases in a semantically congruent context are recognised and remembered much faster,

and visually fixated for longer compared to semantically incongruent discourse (Ehrlich, & Rayner, 1981; Kutas, & Hillyard, 1980; Kamide et al., 2003; Levy et al., 2009; Schwanenflugel, & Shoben, 1985; Stanovich, & West, 1983; van Berkum et al., 1999). Semantically incongruent discourse, on the other hand, was associated with specific changes in electrophysiological markers (Hagoort et al., 2004; Kutas, & Hillyard, 1980; van Berkum et al., 1999): it was found that phrases that violate expectations elicit a larger amplitude of the N400 component (i.e., a negative-going potential occurring around 400 milliseconds after stimulus onset). Most importantly, Federmeier and Kutas (1999) revealed graded changes in the N400 component where between-category violations (e.g., "a chef prepared the salad with a hammer") elicited a larger N400 amplitude compared to within-category violations (e.g., "a chef prepared the salad with a whisk"). It has also been recently discovered that such predictive processing in language can be even interpreted using similar LLMs that were used to predict sensory transformations in the visual pathway (Schrimpf et al., 2021). Together, this evidence clearly demonstrates that years of experience with regularities of the visual world shape the way the brain interprets inputs in vision and language. A key organising principle seems to be the kind of semantic similarity that was found to underlie visual processing (Kutas, & Federmeier, 2000; Newman et al., 2010). A question proposed by Kutas and Federmeier (2000), however, still holds: is there a single, amodal system that encodes these relations and is accessed equally by vision and language, or does language draw on a representational geometry identified in vision?

1.3.2 Semantic memory for amodal representations

The amodal view of the semantic system suggests that concepts are represented independently of sensory modalities (Fodor, 1975; Pylyshyn, 1986). This means that regardless of whether you read the word "cat" or see a picture of a cat, both experiences access a shared, modality-independent representation in

the brain. Such representation would consist of amodal symbols that comprise various characteristics of the cat, including physical or taxonomic features (Kemmerer, 2023). Evidence that supports the amodal semantic system largely comes from lesion studies that show that specific brain damage causes progressive loss of semantic knowledge regardless of the sensory modality (Snowden et al., 1989). Additionally, neuroimaging studies in healthy individuals show it is possible to generalise semantic representations from language to vision and vice versa (Shinkareva et al., 2011; Simanova et al., 2014). While these findings demonstrate some semantic overlap that may be related to amodal symbols, an entirely amodal view of the semantic system faces several critiques.

A major controversy of the amodal semantic system revolves around the symbol grounding problem. It is believed that without an explicit connection to the sensory world, the system could never reach an accurate representation of its environment (Kemmerer, 2023). Moreover, efforts to identify a stable neuroanatomical system dedicated exclusively to storing amodal symbols have produced mixed results. For example, Martin and colleagues (2018) found modality-invariant semantic processing in the perirhinal cortex, but Devereux and colleagues (2013) observed common semantic coding of words and objects in the left-lateralized network. Others propose a shared semantic architecture in the anterior temporal lobe (ATL; Patterson et al., 2007; Lambon Ralph et al., 2010). However, if these areas were indeed entirely amodal stores, then damage to modality-specific regions would not lead to impairments in semantic knowledge in that particular modality. Many authors, however, report cases of individuals who lost their ability to categorise objects in one modality, but categorising in another modality remained intact (Hodges et al., 1992; Hodges et al., 1995; Warrington, 1975). Along with evidence that the visual cortex participates in the retrieval of semantic representations in naming (Chao et al., 1999; Martin et al., 1996), categorisation (Ritchie et al., 2015), or similarity judgements tasks (Mur et

al., 2013; Perani et al., 1999), it is apparent that sensory modalities play a significant role in how semantics is represented.

1.3.3 Grounding language in sensorimotor information

An alternative to the amodal semantic system is the grounded model of cognition that assumes that semantic representations are grounded directly in modality-specific systems (Barsalou, 2008; Boroditsky, & Ramscar, 2002; Gibbs, 2006; Kemmerer, 2023). Consider the example of reading the word "cat". According to the grounded model of cognition, the word is not stored in a purely abstract, amodal format, but instead, it retrieves sensory information from the visual cortex (or any other relevant sensory systems) to represent the concept of the cat. There is now substantial evidence illustrating how language draws on the representational system involved in visual processing. First, studies into metaphorical representations show that abstract concepts in language are often based on mappings to more experiential domains such as vision (Boroditsky, 2000). For example, if we want to express whether we are happy or sad, we frequently describe these feelings in terms of verticality (Barsalou, 2008): we say that we are feeling on top of the world if we are happy, but we say that we feel down or that we hit rock bottom if we are sad. Evidence even shows that these metaphorical mappings to spatial information trigger visual illusions that lead participants to judge a target's motion consistent with the direction implied by the metaphor (Dils, & Boroditsky, 2010). Studies have provided a neural basis to such visual illusions showing that language activates a widely distributed semantic network overlapping with regions engaged during vision (Huth et al., 2016; Popham et al., 2021; Tang et al., 2021; Tang et al., 2023), which often also exhibits in behaviour (Harder, & Tylén, 2019; Lupyan, & Spivey, 2010; Lupyan, & Swingley, 2012; Meteyard et al., 2007; Meyer et al., 2007; Sperber et al., 1979; Stanfield, & Zwaan, 2001). Because these findings show that abstract concepts are naturally

grounded in sensory systems, it poses a fundamental challenge to the concept of amodal semantic system.

This grounding most likely originates in early representation learning where visual categories develop in parallel with language. Research shows that if a set of visual objects is consistently associated with a label, it highlights similarities between those objects, allowing for them to be represented as members of one category (Bergelson, & Swingley; 2012; LaTourrette, & Waxman, 2020; LaTourrette et al., 2023; Tincoff, & Jusczyk, 1999). Although one of the most controversial views in contemporary cognitive science, Kemmerer (2023) argues that cross-linguistic diversity in semantic conventions must have significant consequences for grounded models of cognition as a result of such early representation learning. For example, Winawer and colleagues (2007) demonstrate that cross-linguistic differences in colour categorisation affect performance in a perceptual colour discrimination task. Similarly, Athanasopoulos and colleagues (2015) showed an effect of cross-linguistic grammatical constraints on the categorisation of motion events. Despite its controversy, the theory of linguistic relativity sheds further light on the interconnected nature of language and vision, raising additional questions about the stability and nature of the visual representational space.

1.3.4 Evidence from artificial models of human cognition

Artificial models of human cognition illustrate how language benefits from visual grounding, as some attributes are difficult to convey exclusively through language. As Collell and Moens (2016) point out, “everyone can easily tell from an image whether a face is attractive, yet if one has to describe with language what properties make a face attractive will certainly struggle”. This further demonstrates the practical consequences of the symbol grounding problem

where traditional distributional models of semantics have been limited by building semantic representations entirely in terms of connections of one symbol to another (Bruni et al., 2014). There is now evidence that semantic representations benefit from a combination of visual and linguistic data because each data type provides different but complementary information (Andrews et al., 2009; Lazaridou et al., 2015; Kiela, & Bottou, 2014; Silberer, & Lapata, 2014; Vong et al., 2024). For example, Collell and Moens (2016) demonstrated that while visual models perform better in describing attributes such as motion, surface type, or colour, linguistic models outperform visual models in capturing encyclopaedic knowledge and information about an object's function. Recent evidence shows that models that leverage joint representations of vision and language perform significantly better in many computational tasks including semantic similarity judgements (Kiela, & Bottou, 2014; Silberer, & Lapata, 2014), labelling (Lazaridou et al., 2015), or common-sense reasoning (Lin, & Parikh, 2015). Similar models were found to be extraordinarily good at predicting representations in the visual cortex, outperforming their unimodal counterparts (Tang et al., 2023; Wang et al., 2023).

1.4 Finding the best model of the common representational space

The complex transformations carried out by the visual ventral pathway enable the sensory inputs to be broken down into primitives (Ishai et al., 1999) that comprise interpretable properties of the representation. This includes low-level visual features such as location in the visual field (Golomb, & Kanwisher, 2012; Gomez et al., 2018; Gomez et al., 2019; Kaiser, & Cichy, 2018a) as well as semantic relations of one object to another (Charest et al., 2014; Draschkow, & Võ, 2017; Kriegeskorte et al., 2008b; Kiani et al., 2007; Mur et al., 2013). Therefore, in alignment with the early proposal of Shepard and Chipman (1970), the internal representational space does indeed correspond to relationships and regularities

of the visual world. Similarly, we established that language comprehension is characterised by predictive processing that, through associative learning (Bergelson, & Swingley, 2012; LaTourrette et al., 2023; Smith et al., 2014), captures the stable arrangement of the visual world (Draschkow, & Võ, 2017). A possibility therefore is that the inherent ability to recognise the regularities of the visual world is what makes vision and language converge onto a shared representational space. Such a model of visual-linguistic integration is somewhat related to the work of Patterson and colleagues (2007) who proposed the 'spokes-plus-hub' model that enables inputs from multiple sensory modalities to be transformed and integrated within a single 'hub'. There is also some alignment with the perceptual symbol systems (PSS) model where amodal symbols are rather defined as the dynamic system of connections between the primitives (Barsalou, 1999). Because many studies are showing that the visual cortex is central to language comprehension (Huth et al., 2016; Popham et al., 2021; Tang et al., 2021; Tang et al., 2023), it seems possible that the shared representational space is modal in that it preserves modality-specific features, but amodal in the dynamic connections that represent the relational structure.

1.4.1 Aims of the thesis

In this thesis, I will leverage a powerful tool in comparing representational space – representational similarity analysis (RSA) – to explore the plausibility and behavioural relevance of such a shared representational space in vision and language. Chapter 2 will tackle the shortcomings of the original study by Shepard and Chipman (1970) by asking participants to judge the similarity of natural scene images and sentence captions. This paradigm will enable us to compare the internal simulations of natural scenes with directly perceived scenes. The representational geometry obtained in the visual and linguistic similarity judgements will then be tested for its stability across individuals and sessions, and

relevance to visual representations in the brain. Chapter 3 will use the representational geometry of perceived dissimilarities to investigate how regularities of such a space predict visual search behaviour in a cross-modal priming experiment. This chapter will provide answers to questions including whether language can prime visual targets and whether this relationship is predictable by dissimilarities traditionally associated with vision. Motivated by theories of differential semantic typology, Chapter 3 will also explore whether idiosyncratic representational geometries constrain predictions of the cross-modal priming effect compared to predictions using the representational geometry of a representative participant. Finally, Chapter 4 will solidify the neural underpinnings of shared representations in a mental imagery task while electroencephalography (EEG) is recorded. This chapter will explore whether neural representations simulated in vision and language project onto a shared representational space and what features are responsible for the visuo-linguistic overlap through comparison to mid- and high-level feature-based models.

1.5 References

Aguirre, G. K., Zarahn, E., & D'Esposite, M. (1998). An area within human ventral cortex sensitive to "building" stimuli: evidence and implications. *Neuron, 21*(2), 373-383.

Allison, T., McCarthy, G., Nobre, A., Puce, A., & Belger, A. (1994). Human extrastriate visual cortex and the perception of faces, words, numbers, and colors. *Cerebral Cortex, 4*(5), 544-554.

Andersen, R. A., Asanuma, C., Essick, G., & Siegel, R. M. (1990). Corticocortical connections of anatomically and physiologically defined subdivisions within the inferior parietal lobule. *The Journal of Comparative Neurology, 296*(1), 65-113.

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review, 11*(3), 463-498.

Athanasopoulos, P., Bylund, E., Montero-Melis, G., Damjanovic, L., Schartner, A., Kibbe, A., Riches, N., & Thierry, G. (2015). Two languages, two minds: flexible cognitive processing driven by language of operation. *Psychological Science, 26*(4), 518-526.

Barsalou, L. W. (1999). Perceptual symbol systems. *The Behavioural and Brain Sciences, 22*(4), 577-609.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59*, 617-645.

Bergelson, E., & Swingley, D. (2012). At 6-9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences of the United States of America, 109*(9), 3253-3258.

Biederman, I., Messanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: detecting and judging objects undergoing relational violations. *Cognitive Psychology, 14*(2), 143-177.

Bonner, M. F., & Epstein, R. A. (2021). Object representations in the human brain reflect the co-occurence statistics of vision and language. *Nature Communications, 12*(1).

Boroditsky, L. (2000). Metaphoric structuring: understanding time through spatial metaphors. *Cognition, 75*(1), 1-28.

Boroditsky, L., & Ramscar, M. (2002). The roles of body and mind in abstract thought. *Psychological Science, 13*(2), 185-189.

Bruni, E., Tran, N. K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research, 49*(1), 1-47.

Caramazza, A., & Shelton, J. R. (1998). Domain-specific knowledge systems in the brain: the animate-inanimate distinction. *Journal of Cognitive Neuroscience, 10*(1), 1-34.

Charest, I., Kievit, R. A., Schmitz, T. W., Deca, D., & Kriegeskorte, N. (2014). Unique semantic space in the brain of each beholder predicts perceived similarity. *Proceedings of the National Academy of Sciences of the United States of America, 111*(40), 14565-14570.

Carlson, T. A., Ritchie, J. B., Kriegeskorte, N., Durvasula, S., & Ma, J. (2014). Reaction time for object categorization is predicted by representational distance. *Journal of Cognitive Neuroscience, 26*(1), 132-142.

Cichy, R. M., Kriegeskorte, N., Jozwik, K. M., van den Bosch, J. J. F., & Charest, I. (2019). The spatiotemporal neural dynamics underlying perceived similarity for real-world objects. *NeuroImage, 194*, 12-24.

Chao, L. L., Haxby, J. V., & Martin, A. (1999). Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nature Neuroscience, 2*(10), 913-919.

Cohen, M. A., Alvarez, G. A., Nakayama, K., & Konkle, T. (2017). Visual search for object categories is predicted by the representational architecture of high-level visual cortex. *Journal of Neurophysiology, 117*(1), 388-402.

Cohen, M. A., Konkle, T., Rhee, J. Y., Nakayama, K., & Alvarez, G. A. (2014). Processing multiple visual objects is limited by overlap in neural channels. *Proceedings of the National Academy of Sciences of the United States of America, 111*(24), 8955-8960.

Cohen, M. A., Nakayama, K., Konkle, T., & Stantic, M. (2015). Visual awareness is limited by the representational architecture of the visual system. *Journal of Cognitive Neuroscience, 27*(11), 2240-2252.

Collell, G., & Moens, M. F. (2016). Is an image worth more than a thousand words? On the fine-grain semantic differences between visual and linguistic representations. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2807-2817.

Contier, O., Baker, C. I., & Hebart, M. N. (2024). Distributed representations of behaviour-derived object dimensions in the human visual system. *Nature Human Behaviou*r, 1-15.

Cutler, A., & Clifton, C. (2000). *Comprehending spoken language: a blueprint of the listener.* In Brown, C. M., & Hagoort, P. The Neurocognition of Language. Oxford University Press.

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science, 15*(8), 559-564.

De Haas, B., Schwarzkopf, D. S., Alvarez, I., Lawson, R. P., Henriksson, L., Kriegeskorte, N., & Rees, G. (2016). Perception and processing of faces in the human brain is tuned to typical feature locations. *The Journal of Neuroscience, 36*(36), 9289-9302.

Devereux, B. J., Clarke, A., Marouchos, A., & Tyler, L. K. (2013). Representational similarity analysis reveals commonalities and differences in the semantic processing of words and objects. *The Journal of Neuroscience, 33*(48), 18906-18916.

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences, 11*(8), 333-341.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron, 73*(3), 415-434.

Dils, A. T., & Boroditsky, L. (2010). Visual motion aftereffect from understanding motion language. *Proceedings of the National Academy of Sciences of the United States of America, 107*(37), 16396-16400.

Doerig, A., Kietzmann, T. C., Allen, E., Wu, Y., Naselaris, T., Kay, K., & Charest, I. (2024). Visual representations in the human brain are aligned with large language models. *bioRxiv*.

Draschkow, D., & Vo, M. L. (2017). Scene grammar shapes the way we interact with objects, strengthens memories, and speeds search. *Scientific Reports, 7*(1), 16471.

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior, 20*(6), 641-655.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature, 392*(6676), 598-601.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: long-term memory structure and sentence processing. *Journal of Memory and Language, 41*(4), 469-496.

Federmeier, K. D., & Kutas, M. (2001). Meaning and modality: influences of context, semantic memory organization, and perceptual predictability on picture processing. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 27*(1), 202-224.

Felleman, D. J., & van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex, 1*(1), 1-47.

Fodor, J. A. (1975). *The Language of Thought*. Harvard University Press.

Gibbs, R. W. (2006). *Embodiment and Cognitive Science*, Cambridge University Press.

Golob, J. D., & Kanwisher, N. (2012). Higher level visual cortex represents retinotopic, not spatiotopic object location. *Cerebral Cortex, 22*(12), 2794-2810.

Gomez, J., Drain, A., Jeska, B., Natu, V. S., Barnett, M., & Grill-Spector, K. (2019). Development of population receptive fields in the lateral visual stream improves spatial coding amid stable structural-functional coupling. *Neuroimage, 188*, 59-69.

Gomez, J., Natu, V., Jeska, B., Barnett, M., & Grill-Spector, K. (2018). Development differentially sculpts receptive fields across early and high-level human visual cortex. *Nature Communications, 9*(1), 788.

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences, 15*(1), 20-25.

Goodale, M. A., Milner, A. D., Jakobson, L. S., & Carey, D. P. (1991). A neurological dissociation between perceiving objects and grasping them. *Nature, 349*(6305), 154-156.

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science, 304*(5669), 4388-441.

Hanson, S. J., Matuska, T., & Haxby, J. V. (2004). Combinatorial codes in ventral temporal lobe for object recognition: Hasby (2001) revisited: is there a 'face' area? *Neuroimage, 23*(1), 156-166.

Harder, M., & Tylen, K. (2019). What is up and down in embodied language processing? An experimental study on semantic priming of visual perception. *Cognitive Semiotics, 12*(2).

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293*(5539), 2425-2430.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour, 4*(11), 1173-1185.

Hebb, D. O. (1949). *The Organization of Behavior. A Neuropsychological Theory*. John Wiley & Sons, Inc, London.

Heilbron, M., Armeni, K., Schoffelen, J. M., Hagoort, P., & de Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences of the United States of America, 119*(32).

Hodges, J. R., Graham, N., & Patterson, K. (1995). Charting the progression in semantic dementia: implications for the organisation of semantic memory. *Memory, 3*(3-4), 463-495.

Hodges, J. R., Patterson, K., Oxbury, S., & Funnell, E. (1992). Semantic dementia. Progressive fluent aphasia with temporal lobe atrophy. *Brain, 115*(6), 1783-1806.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology, 160*(1), 106-154.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology, 195*(1), 215-243.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature, 532*(7600), 453-458.

Ishai, A., Ungerleider, L. G., Martin, A., Schouten, J. L., & Haxby J. V. (1999). Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Sciences of the United States of America, 96*(16), 9379-9384.

Jakobson, L. S., Archibald, Y. M., Carey, D. P., & Goodale, M. A. (1991). A kinematic analysis of reaching and grasping movements in a patient recovering from optic ataxia. *Neuropsychologia, 29*(8), 803-809.

Kaiser, D., & Cichy, R. M. (2018a). Typical visual-field locations enhance processing in object-selective channels of human occipital cortex. *Journal of Neurophysiology, 120*(2), 848-853.

Kaiser, D., & Cichy, R. M. (2018b). Typical visual-field locations facilitate access to awareness for everyday objects. *Cognition, 180*, 118-122.

Kaiser, D., Quek, G. L., Cichy, R. M., & Peelen, M. V. (2019). Object vision in a structured world. *Trends in Cognitive Sciences, 23*(8), 672-685.

Kamide, Y., Altmann, G. T., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: evidence from anticipatory eye movements. *Journal of Memory and Language, 49*(1), 133-156.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience, 17*(11), 4302-4311.

Kemmerer, D. (2023). Grounded cognition entails linguistic relativity: a neglected implication of a major semantic theory. *Topics in Cognitive Science, 15*(4), 615-647.

Kiani, R., Esteky, H., Mirpour, K., & Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology, 97*(6), 4296-4309.

Kiela, D., & Bottou, L. (2014). Learning image embeddings using convolutional neural networks for improved multi-modal semantics. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 36-45.

Kirchner, H., & Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: visual processing speed revisited. *Vision Research, 46*(11), 1762-1776.

Kjelgaard, M. M., & Speer, S. R. (1999). Prosodic facilitation and interference in the resolution of temporary syntactic closure ambiguity. *Journal of Memory and Language, 40*(2), 153-194.

Kravitz, D. J., Vinson, L. D., & Baker, C. I. (2008). How position dependent is visual object recognition? *Trends in Cognitive Sciences, 12*(3), 114-122.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008a). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2.*

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P. A. (2008b). Matching categorical representations in inferior temporal cortex of man and monkey. *Neuron, 60*(6), 1126-1141.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition, and Neuroscience, 31*(1), 32-59.

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences, 4*(12), 463-470.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science, 207*(4427), 203-205.

Lambon Ralph, M. A., Sage, K., Jones, R. W., & Mayberry, E. J. (2010). Coherent concepts are computed in the anterior temporal lobes. *Proceedings of the National Academy of Sciences of the United States of America, 107*(6), 2717-2722.

LaTourrette, A. S., & Waxman, S. R. (2020). Naming guides how 12-month-old infants encode and remember objects. *Proceedings of the National Academy of Sciences of the United States of America, 117*(35), 21230-21234.

LaTourrette, A., Chan, D. M., & Waxman, S. R. (2023). A principled link between object naming and representation is available to infants by seven months of age. *Scientific Reports, 13*(14328).

Lazaridou, A., Nghia The Pham, & Baroni, M. (2015). Combining language and vision with a multimodal skip-gram model. *Proceedings of the 2015 Conference of the North America Chapter of the Association for Computational Linguistics: Human Language Technologies,* 153-163.

Lehky, S. R., & Serano, A. B. (2007). Comparison of shape encoding in primate dorsal and ventral visual pathways. *Journal of Neurophysiology, 97*(1), 307-319.

Levine, D. N., Warach, J., & Farah, M. (1985). Two visual systems in mental imagery: dissociation of 'what' and 'where' in imagery disorders due to bilateral posterior cerebral lesions. *Neurology, 35*(7), 1010-1018.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126-1177.

Lin, X., & Parikh, D. (2015). Don't just listen, use your imagination: Leveraging visual common sense for non-visual tasks. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR*), 2984-2993.

Logothetis, N. K., Pauls, J., Bulthoff, H. H., & Poggio, T. (1994). View-dependent object recognition by monkeys. *Current Biology, 4*(5), 401-414.

Luo, Y., & Zhou, X. (2010). ERP evidence for the online processing of rhythmic pattern during Chinese sentence reading. *Neuroimage, 49*(3), 2836-2849.

Lupyan, G., & Spivey, M. J. (2010). Making the invisible visible: verbal but not visual cues enhance visual detection. *PLoS One, 5*(7).

Lupyan, G., & Swingley, D. (2012). Self-directed speech affects visual search performance. *Quarterly Journal of Experimental Psychology, 65*(6), 1068-1085.

Martin, C. N., Douglas, D., Newsome, R. N., Man, L. L., & Barense, M. D. (2018). Integrative and distinctive coding of visual and conceptual object features in the ventral visual stream. *Elife, 7.*

Martin, A., Wiggs, C. L., Ungerleider, L. G., & Haxby, J. V. (1996). Neural correlates of category-specific knowledge. *Nature, 379*(6566), 649-652.

McCarthy, G., Pruce, A., Gore, J. C., & Allison, T. (1997). Face-specific processing in the human fusiform gyrus*. Journal of Cognitive Neuroscience, 9*(5), 605-610.

McNeil, J. E., & Warrington, E. K. (1993). Prosopagnosia: a face-specific disorder. *The Quarterly Journal of Experimental Psychology. Human Experimental Psychology, 46*(1), 1-10.

Meteyard, L., Bahrami, B., & Vigliocco, G. (2007). Motion detection and motion verbs: language affects low-level visual perception. *Psychological Science, 18*(11), 1007-1013.

Meyer, A. S., Belke, E., Telling, A. L., & Humphreys, G. W. (2007). Early activation of object names in visual search. *Psychonomic Bulletin & Review, 14*(4). 710-716.

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A., & Kriegeskorte, N. (2013). Human object-similarity judgements reflect and transcend the primate-IT object representation. *Frontiers in Psychology, 4*.

Newman, S. D., Ikuta, T., & Burns, Jr T. (2010). The effect of semantic relatedness on syntactic analysis: an fMRI study. *Brain and Language, 113*(2), 51-58.

Palmer, T. E. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition, 3*(5), 519-526.

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews, 8*(12), 976-987.

Perani, D., Schnur, T., Tettamanti, M., Gorno-Tempini, M., Cappa, S. F., & Fazio, F. (1999). Word and picture matching: a PET study of semantic category effects. *Neuropsychologia, 37*(3), 293-306.

Perfetti, C. A. (2000). *Comprehending written language: a blueprint of the reader*. In Brown, C. M., & Hagoort, P. The Neurocognition of Language. Oxford University Press.

Pitcher, D., & Ungerleider, L. G. (2022). Evidence for a third visual pathway specialised for social perception. *Trends in Cognitive Sciences, 25*(2), 100-110.

Polk, T. A., & Farah, M. J. (1995). Brain localization for arbitrary stimulus categories: a simple account based on Hebbian learning. *Proceedings of the National Academy of Sciences of the United States of America, 92*(26), 12350-12373.

Polk, T. A., & Farah, M. L. (1998). The neural development and organization of letter recognition: evidence from functional neuroimaging, computational modelling, and behavioural studies. *Proceedings of the National Academy of Sciences of the United States of America, 95*(3), 847-852.

Popham, S. F., Huth, A. G., Bilenko, N. Y., Deniz, F., Gao, J. S., Nunez-Elizalde, A. O., & Gallant, J. L. (2021). Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature Neuroscience, 24*(11), 1628-1636.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology. Human Learning and Memory, 2*(5), 509-522.

Pylyshyn, Z. W. (1973). What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological Bulletin, 80*(1), 1-24.

Pylyshyn, Z. W. (1986). *Computation and cognition: toward a foundation for cognitive science*. The MIT Press.

Ritchie, J. B., Tovar, D. A., & Carlson, T. A., (2015). Emerging object representations in the visual system predict reaction times for categorization. *PLoS Computational Biology, 11*(6).

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: integrative modelling converges in predictive processing. *Proceedings of the National Academy of Sciences of the United States of America, 118*(45).

Schwanenflugel, P. J., & Shoben, E. J., (1985). The influence of sentence constraint on the scope of facilitation for upcoming words. *Journal of Memory and Language, 24*(2), 232-252.

Shallice, T., & Saffran, E. (1986). Lexical processing in the absence of explicit word identification: evidence from a letter-by-letter reader. *Cognitive Neuropsychology, 3*(4), 429-458.

Shepard, R. N., & Chipman, S. (1970). Second-order isomorphism of internal representations: shapes of states. *Cognitive Psychology, 1*(1), 1-17.

Shinkareva, S. V., Malave, V. L., Mason, R. A., Mitchell, T. M., & Just, M. A. (2011)., Commonality of neural representations of words and pictures. *NeuroImage, 54*(3), 2418-2425.

Silberer, C., & Lapata, M. (2014). Learning grounded meaning representations with autoencoders. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 1,* 721-732.

Simanova, I., Hagoort, P., Oostenveld, R., & van Gerven, M. A. (2014). Modality-independent decoding of semantic information from the human brain. *Cerebral Cortex, 24*(2), 426-434.

Smith, L. B., Suanda, S. H., & Yu, C. (2014). The unrealized promise of infant statistical word-referent learning. *Trends in Cognitive Sciences, 18*(5), 251-258.

Snowden, J., Goulding, P. J., & Neary, D. (1989). Semantic dementia: a form of circumscribed cerebral atrophy. *Behavioural Neurology, 2*(3), 167-182.

Sperber, R. D., McCauley, C., Ragain, R. D., & Weil, C. M. (1979). Semantic priming effects on picture and word processing. *Memory & Cognition, 7*(5), 339-345.

Stanfield, R. A., & Zwaan, R. A. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science, 12*(2), 153-156.

Stanovich, K. E., & West, R. F. (1983). On priming by a sentence context. *Journal of Experimental Psychology, 112*(1), 1-36.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience, 19,* 109-139.

Tang, J., LeBel, A., & Huth, A. G. (2021). Cortical representations of concrete and abstract concepts in language combine visual and linguistic representations. *bioRxiv*.

Tang, J., Du, M., Vo, V. A., Lal, V., & Huth, A. G. (2023). Brain encoding models based on multimodal transformers can transfer across language and vision. *arXiv*.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*(6582), 520-522.

Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science, 10*(2), 172-175.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*(4), 327-352.

Ungerleider, L. G., & Haxby, J. V. (1994). 'What' and 'where' in the human brain. *Current Opinion in Neurobiology, 4*(2), 157-165.

Ungerleider, L. G., & Mishkin, M. (1982*). Two cortical visual systems*. In Goodale, M., Ingle, D. J., & Mansfield, R. J. W. (Eds). Analysis of visual behaviour. The MIT Press.

Van Berkum, J. J., Hagoort, P., & Brown, C. M. (1999). Semantic integration in sentences and discourse: evidence from the N400. *Journal of Cognitive Neuroscience, 11*(6), 657-671.

Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. M. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science, 383*(6682), 504-511.

Wang, A. Y., Kay, K., Naselaris, T., Tarr, M. J., & Wehbe, L. (2023). Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence, 5*, 1415-1426.

Warrington, E. K. (1975). The selective impairment of semantic memory. *The Quarterly Journal of Experimental Psychology, 27*(4), 635-657.

Warrington, E. K., & Shallice, T. (1980). Word-form dyslexia. *Brain, 103*(1), 99-112.

Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review, 20*(2), 158-177.

Whiteley, A. M., & Warrington, E. K. (1978). Selective impairment of topographical memory: a single case study. *Journal of Neurology, Neurosurgery, and Psychiatry, 41*(6), 575-578.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences of the United States of America, 104*(19), 7780-7785.

Yeh, L. C., Throat, S., & Peelen, M. V. (2024). Predicting cued and oddball visual search performance from fMRI, MEG, and DNN neural representational similarity*. The Journal of Neuroscience, 44*(12).

Chapter 2

Representations in vision and language converge in a multidimensional space of perceived dissimilarities

2.1 Abstract

Emerging evidence suggests that representations in vision and language converge in their ability to transform their inputs into a representational format well captured by large language models (LLMs). Any answer to how vision and language converge into a shared representational format also requires an understanding of how these shared representations are behaviourally relevant. We collected visual and linguistic similarity judgements from 63 participants who performed the multiple arrangements (MA) task on 100 natural scene images and 100 sentence captions describing the images. To describe how the multimodal representations derive from a common representational space, we used the behavioural representational geometries to predict functional magnetic resonance imaging (fMRI) activity of 8 participants from the Natural Scenes Dataset (NSD) while they viewed the image set in a visual recognition task. Representational similarity analysis (RSA) revealed strikingly similar predictions along the mid- and high-level visual regions for both visual and linguistic similarity judgements. Importantly, we observed reliable predictions even though the linguistic similarity judgements were made by participants who had never seen the images before. Furthermore, we used recurrent convolutional neural networks trained to predict LLM embeddings of scene captions to investigate whether LMM representations explain the multidimensional similarity space. We found that LMM embeddings of the sentence captions offer an outstanding training objective in predicting the multidimensional space obtained in behaviour. Together, these results demonstrate that language and vision converge in their ability to recognise and encode regularities of the visual world. We anticipate our results to be the first step towards establishing a shared multidimensional similarity space that encodes behaviourally relevant representations about the sensory experience.

2.2 Introduction

How can we model high-level representations in response to complex natural scenes? Emerging evidence in cognitive computational neuroscience suggests that large language model (LLM) embeddings of natural scene captions are well aligned with brain responses to viewing corresponding natural scene images (Doerig et al., 2024; Lahner et al., 2024; Tang et al., 2023). In fact, training visual Deep Neural Networks (DNNs) to predict LLM embeddings leads to state-of-the-art models of visually evoked neural activities (Doerig et al., 2024). In addition, aligning visual and linguistic representations when modelling natural scenes has been shown to provide even better predictors of visual responses in the brain (Conwell et al., 2022, Conwell et al., 2023, Wang et al., 2023), and improve performance in multiple computational tasks (Lu et al. 2019; Tan and Bansal 2019; Pramanick et al. 2022; Radford et al. 2021). One possibility to explain the success of LLMs in predicting the brain's visual responses is that, through training on vast linguistic corpora, LLMs learn to capture rich statistical world knowledge (Kaplan et al. 2020; Hernandez et al. 2021). Similarly, the brain needs to extract rich representations infused with world knowledge from visual inputs. For this reason, both systems converge onto a similar representational format that accounts for the complex information conveyed in natural scenes.

These recent developments in the cognitive computational neuroscience of vision are parallelled with breakthroughs in language research where growing evidence shows that LLMs are good predictors of natural language representations in the brain (Russo et al. 2022; Caucheteux, & King, 2022; Dong, & Toneva, 2023; Goldstein et al., 2021; Jain, & Huth, 2018; Toneva, & Wehbe, 2019; Wehbe et al., 2014). Some evidence using LLMs to predict brain responses to language would even suggest that the human language system is optimised for predictive processing (Schrimpf et al., 2021 see Tuckute et al., 2024 for a

systematic review), and high-level language comprehension of narratives (Caucheteux et al., 2022). This set of converging evidence from language and vision leads to the intriguing possibility that sensory systems are geared to transform their low-level sensory inputs into a representational format well captured by large language models. However, it remains unclear how such representations could be useful for behaviour.

Here, to further understand whether the representational format of the visual system that is accounted for by the LLMs is relevant to behaviour, we measured similarity judgments for a subset of the natural scenes in the Natural Scenes Dataset (NSD; Allen et al., 2022). The judgments were performed on the natural scenes (visual modality) and the scene captions (sentences describing the natural scene images; linguistic modality). Both resulting sets of similarity judgments were used to model visual representations in the human brain and compared with state-of-the-art recurrent convolutional neural network models trained either to predict category-level representations or LLM-embeddings of scene captions (Doerig et al. 2024). Our results show that similarity judgements performed on sentences and images similarly predict responses in the human ventral visual stream and can be similarly modelled with great accuracy using artificial neural networks that transform visual inputs into LLM embeddings of scene captions.

2.3 Results

Sixty-three participants attended three lab-based sessions in which they completed three multi-arrangement (MA; Kriegeskorte, & Mur 2012) tasks each presenting either visual or linguistic stimuli. Participants were counterbalanced across the behavioural sessions using the Latin square design to control for task order effects. This design allowed us to retain the unique similarity structure of

each modality in the first participants' sessions and to investigate how the similarity structure changed as participants progressed through the experiment. The stimulus set was obtained from the Natural Scenes Dataset (NSD; Allen et al., 2022) and consisted of 100 natural scene images and 100 sentence captions describing the images. Participants viewed the entire set of images on a computer screen in the MA task and were instructed to use a computer mouse to arrange the images according to their similarity such that images perceived as more similar were placed closer together while images perceived as less similar were placed farther apart (Figure 2.1A left). The linguistic MA task followed the same procedure but presented the set of 100 sentences. Every sentence was initially presented as an asterisk in the MA arena, with text appearing either when the participant hovered the mouse cursor over the asterisk or when the item was selected to be moved (Figure 2.1A right). To control for the cognitive load between the visual and linguistic modalities, we also introduced a control condition in which participants completed the MA task on 'hidden images' (Supplementary Figure 2.1).

First, we investigated the overlap between the representational geometry of vision and language. For every participant, on each trial of the MA task, we computed the Euclidean distance between every pair of stimuli to determine the perceived dissimilarity. An adaptive procedure was used to select stimuli for subsequent trials using the 'lift the weakest' algorithm that maximises evidence for each pair. The partial trial dissimilarity matrices were aligned using an iterative weighted average procedure (Kriegeskorte, & Mur 2012). This was done independently for each task modality. The dissimilarities were then assembled in two modality-specific representational dissimilarity matrices (RDMs) and averaged across all participants (Figure 2.1B). A Spearman correlation of the upper-triangular part of these two RDMs revealed a significant relationship between the visual and linguistic RDMs (Spearman's ρ = .781), demonstrating a

shared representational structure between the similarity judgements of visual and linguistic stimuli.

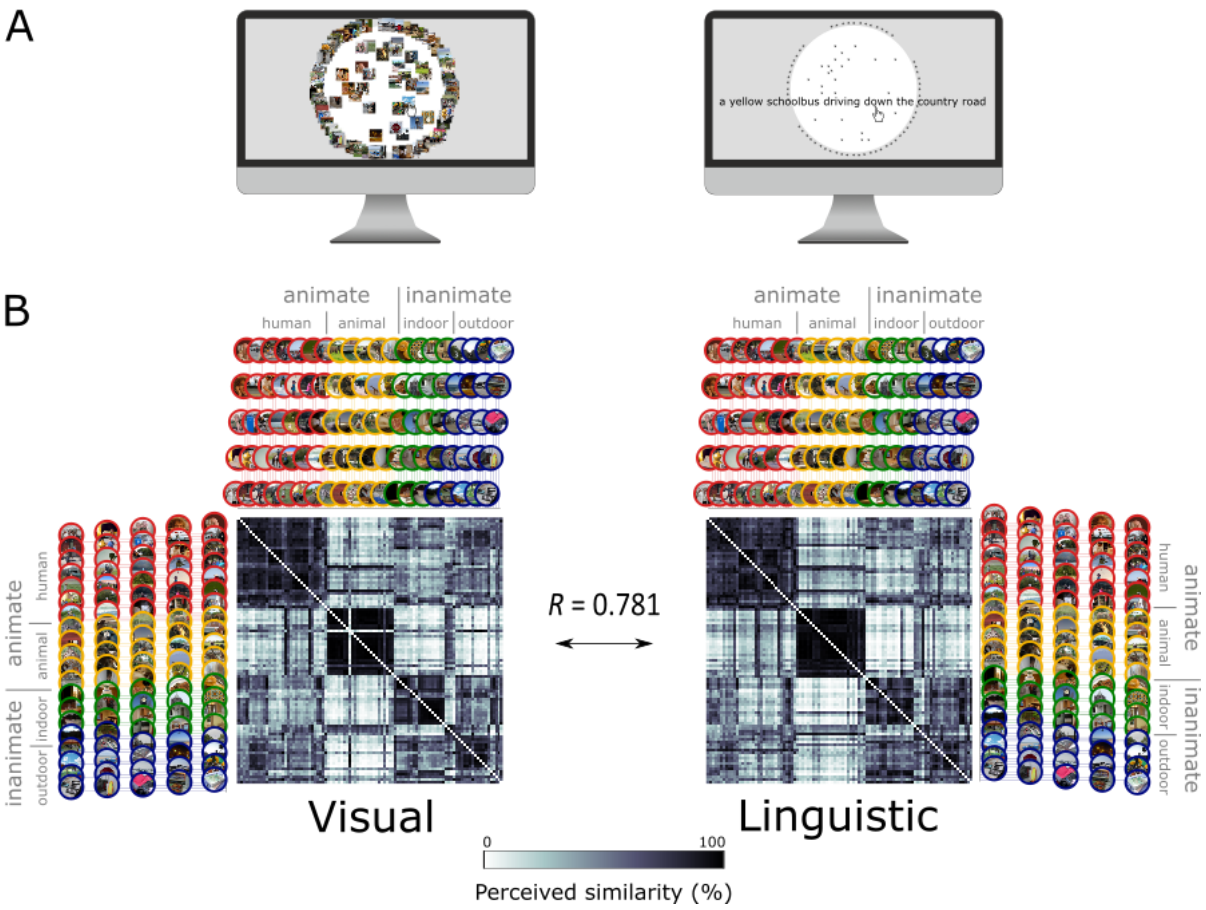
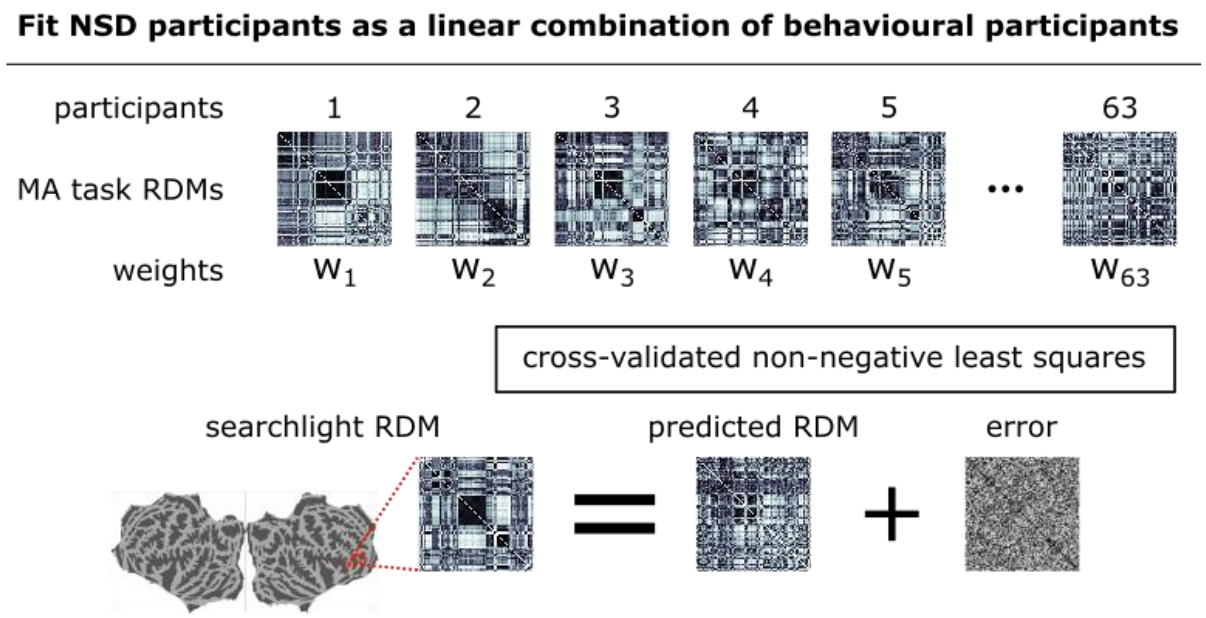


Figure 2.1 | Experimental design of the MA tasks. (A) In separate lab sessions each scheduled at least a week apart, participants completed the MA task either on the visual (left) or linguistic (right) modality; the order of the sessions was counterbalanced across participants. To handle the large linguistic dataset, we adapted the MA method such that every caption was depicted as an asterisk and only the item currently being dragged was displayed. (B) The rank-transformed averaged modality-specific RDMs were correlated to investigate the representational overlap between the visual and linguistic modalities. The fixed-effects between-modality RDM correlation for the full dataset was 0.781 suggesting a strong level of representational overlap.

2.3.1 Visual and linguistic similarity judgements predict visually evoked brain activity

Next, we explored whether the behavioural representational geometry shares a representational structure with the visually evoked brain representational space. To this end, we used the publicly available NSD fMRI dataset (Allen et al., 2022), which consists of high-field (7T) functional MRI data from 8 participants who viewed ~73,000 natural scene images in a visual recognition task. In each NSD participant, we extracted response patterns for our subset of 100 images. We used an fMRI searchlight procedure to compare local brain representational geometries to behavioural RDMs across the whole cortex. For each searchlight sphere, we trained a model to predict the brain representational geometries based on a linear combination of the 63 behavioural RDMs. This was performed using a 10-fold cross-validation procedure. In each fold, weights were fitted using a non-negative least squares (NNLs) regression using a subset of images for training, and the learned weights were then projected on distances from the remaining test set of images to obtain a predicted searchlight RDM. The representational alignment was then tested as the correlation between the predicted brain RDM and observed brain RDM. The correlation values for test set predictions were then averaged across folds. This was performed independently for the visual and linguistic behavioural RDMs. We observed a significant overlap between the representational geometry predicted by the visual modality and the observed brain RDMs with Pearson correlations peaking at 0.363 at the group level (see Supplementary Figure 2.2 for participant-level analyses). Whole brain searchlight correlations are shown on a flattened cortical surface map (Figure 2.2B top panel) revealing significant alignment in regions located bilaterally along the mid- and high-level visual regions including the inferior temporal sulcus (ITS) and the superior temporal sulcus (STS).

We followed the same procedure to investigate how well the linguistic representational geometry predicts the geometry of brain responses to viewing natural scene images. We found strong evidence for representational alignment between the observed brain RDMs and RDMs predicted by the linguistic modality with Pearson correlations peaking at 0.353 at the group level (see Supplementary Figure 2.2 for participant-level analyses). Most importantly, the flattened cortical surface map (Figure 2.2B bottom panel) revealed a strikingly similar network of significant relationships to the ones predicted by the visual modality located bilaterally along the occipitotemporal cortex. These results remained even when using RDMs from linguistic similarity judgements made by participants who had never seen the natural scene images before (Supplementary Figure 2.3 and Supplementary Figure 2.4). Our results demonstrate that representations in behaviour and afferent vision project onto a shared multidimensional space of perceived dissimilarities. Note that this holds even when we compare brain and behavioural RDMs using simple Pearson correlation (Supplementary Figure 2.5 and Supplementary Figure 2.6) demonstrating that the effect cannot be attributed to the cross-validated procedure where flexible weights were assigned to each behavioural RDM when creating the predicted brain RDMs.


Fit NSD participants as a linear combination of behavioural participants
participants
1
2
3
4
5
63
MA task RDMs
weights
w_1
w_2
w_3
w_4
w_5
w_{63}
cross-validated non-negative least squares
searchlight RDM
predicted RDM
error
Group-level statistics

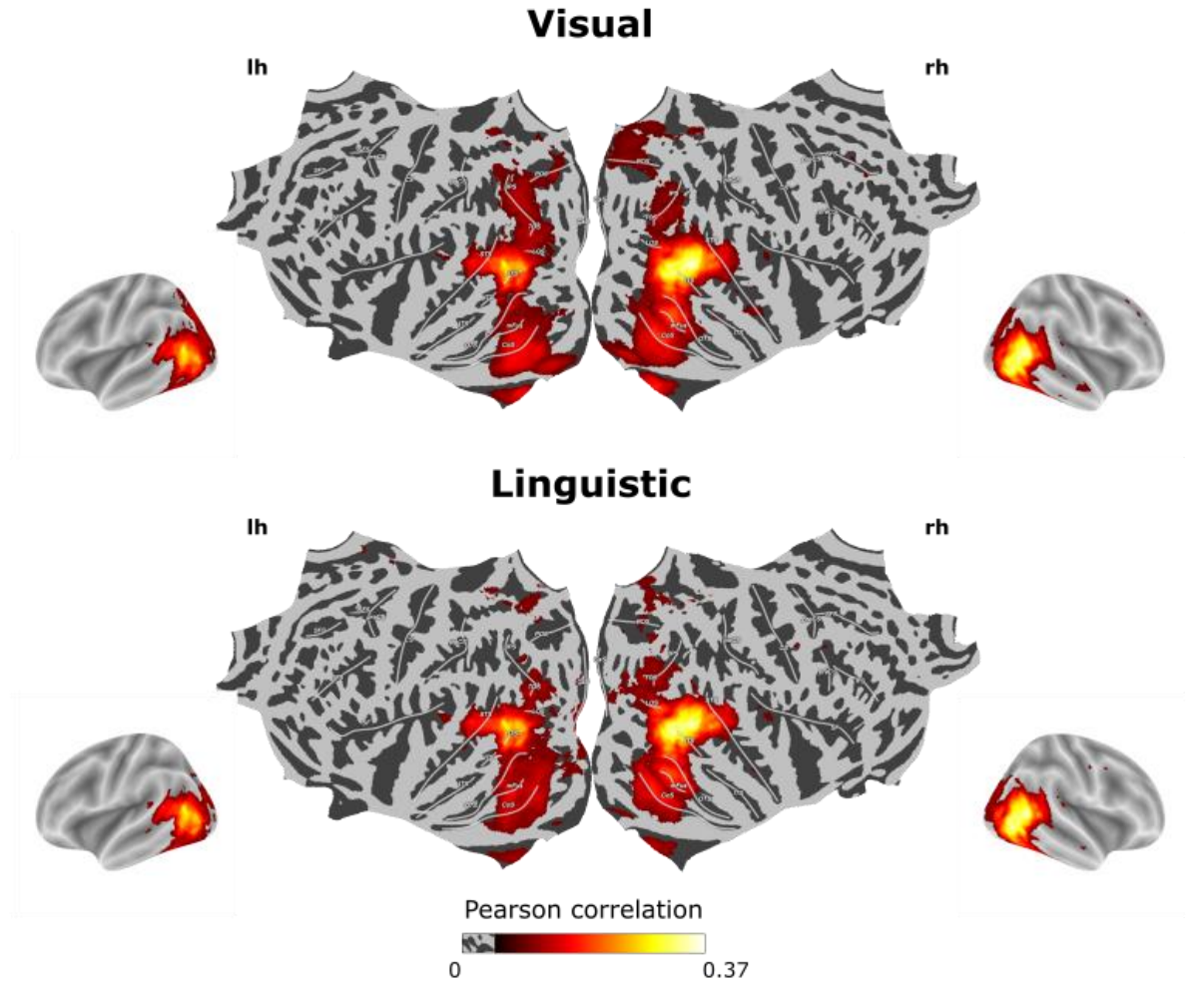

Visual
lh
rh
Linguistic
lh
rh
Pearson correlation
0
0.37

Figure 2.2 | Representational alignment of multimodal similarity judgements with brain responses to viewing natural scene images. (A) Non-negative least squares regression was used to assign flexible weights to each behavioural RDM to represent their relative fit to the searchlight brain RDM. NNL weights were then used to create behaviour-predicted brain RDMs, which were then compared to brain test RDMs at every searchlight. (B) Pearson correlations were averaged across folds in each NSD participant. Significance was determined across NSD participants using a one-sided test against zero and thresholded at the FDR corrected $p < 0.05$ level. Comparison of the behaviour-predicted and observed brain RDMs revealed strong correlations located bilaterally along the occipitotemporal cortex with correlations peaking at 0.363 and 0.353 for the visual (top panel) and linguistic (bottom panel) modalities respectively.

2.3.2 The similarity structure of the shared multidimensional space aligns with LMM representations

Previous research has revealed that sentence embeddings from large language models (LLMs) are well-suited for explaining brain responses to viewing natural scene images (Doerig et al., 2024; Tang et al., 2023) and language comprehension (Russo et al. 2022; Caucheteux, & King, 2022; Dong, & Toneva, 2023). Do representations from LLMs align with representational geometry of behavioural similarity judgements just as well as they align with representations in the human brain? To answer this question, we passed the set of 100 natural scene images through a recurrent convolutional neural network (RCNN) introduced by Doerig and colleagues (2024), which was trained to predict sentence embeddings and category labels of the images. We constructed model RDMs where each RDM represented the activity of the last layer and time step from each of the sentence- and category-computable RCNNs. We then created a behaviour-predicted RDM using the cross-validated non-negative least squares procedure and evaluated it on the RCNN RDMs to investigate the extent to which the geometry of behavioural RDMs fits the representational structure of the recurrent activity of the model. We

found that both visual and linguistic RDMs generally fit better early time steps in the first six layers of both RCNNs (Figure 2.3A - Figure 2.3D; see Supplementary Figure 2.7 for statistical comparison within RCNN). This trend, however, reversed from layer 7 onwards where behavioural RDMs yielded a better fit with the last time step within a layer. Overall, late layers in both RCNNs were found to be a better representation of the behavioural RDMs (Supplementary Figure 2.7). What stands out in Figure 2.3E and Figure 2.3D is that starting from layer 4 onward, both visual and linguistic RDMs produced a constantly better fit with the sentence-computable RCNN. Most importantly, behavioural RDMs predicted the similarity structure of sentence-computable RCNN significantly better than the structure of AlexNet (Supplementary Figure 2.8A; Supplementary Figure 2.8B), which is a model that has been at the forefront of explaining representational spaces in the human brain (Jozwik et al., 2018; Khalig-Razavi, & Kriegeskorte, 2014). These findings hold even when we control for potential task order effects in the behavioural RDMs (Supplementary Figure 2.8C; Supplementary Figure 2.8D).

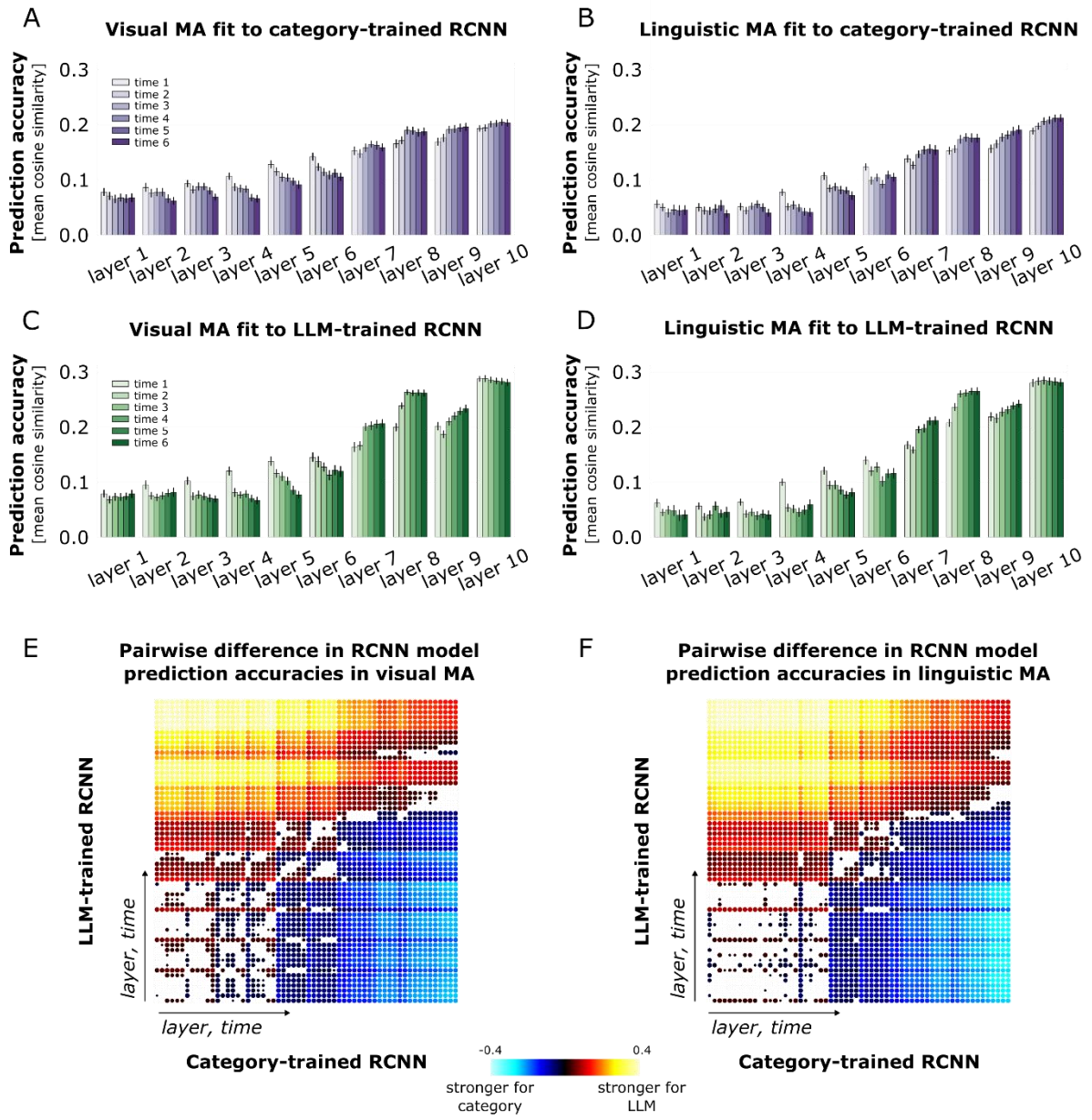


Figure 2.3 | Representational alignment of multimodal similarity judgements and LLM representations. Cross-validated non-negative least squares procedure was used to create a behaviour-predicted RDM. The bars show average cosine similarity between the behaviour-predicted RDM and observed RDMs for a given layer and time step across 10 seeds. The error bars are standard errors of the mean estimated across 10 cross-validation folds for each seed. Prediction accuracy between the category-trained RCNN RDMs and behaviour-predicted RDMs in vision and language is depicted in panels (A) and (B) respectively. Panels (C) and (D) show fit with the LLM-trained RCNN RDMs. The pairwise difference in prediction accuracies between LLM- and category-trained RCNNs was

calculated to identify a better-fitting RCNN in the visual modality (E) and linguistic modality (F). Two-sided t-tests were performed on cosine similarities across seeds for each layer and time step to determine whether there is a significant difference in prediction accuracies between RCNNs. Only significant differences in cosine similarities are depicted. The radius of the points marks the size of the p-values ($p < 0.05$ FDR corrected). The colour gradient depicts the size of the difference where the mean prediction accuracies of the category-trained RCNN were subtracted from the mean prediction accuracies of the LLM-trained RCNN. The layer and time complexity increases from the origin.

2.3.3 Alignment in representational geometry of vision and language

Since every participant completed both task modalities and the stimuli covaried between them, we could assume that the task exposure in early sessions could increase the representational alignment between participants in later sessions due to strategic recall of previously encountered properties and arrangement tactics. For this reason, we compared between-modality correlations in the first and final experimental sessions to explore whether the representational alignment increased once participants were exposed to the full stimulus set. First, we correlated RDMs from participants who completed the visual MA task in their first session and participants who completed the linguistic MA task in their first session. We found an average Spearman correlation of the visual and linguistic RDMs of 0.156 (Figure 2.4B - dark grey bar, one-sided test: $t(18) = 18.354$, $p < 0.0001$) suggesting that there is a significant overlap between the representational geometry of the directly perceived visual stimuli and the internally reconstructed representations of the linguistic stimuli. It is worth noting that this result is based on between-subject comparisons where neither of the groups was familiar with the content of the other modality, which demonstrates that participants generally agreed on the configuration of the arrangements whether they started the experiment with images or captions (Figure 2.4A). We followed the same procedure to correlate RDMs between-modality in the final session for the same

subset of participants. This analysis is orthogonal to the previous one as both groups considered here were familiar with the content of the other modality. We found a between-modality correlation of 0.195 (Figure 2.4B - light grey bar, one-sided test: $t(18) = 27.648$, $p < 0.0001$). We then performed a one-tailed paired samples t-test to determine whether the task exposure affected the arrangements in later sessions. We hypothesised that if the task exposure affected the way participants arranged the items in later sessions, we would observe an increase in the between-modality correlations in the final session. The test revealed a significant increase in the visual and linguistic RDM correlations in the final session (Figure 2.4B, $t(18) = 3.964$, $p < 0.0001$). This suggests that experience with either modality in early sessions causes participants to configure later arrangements of the other modality similarly.

2.3.4 Preceding visual modality increases representational alignment between participants

What are the prominent features that lead to the increase in representational alignment between participants? Providing a participant started with the visual modality, are their linguistic arrangements in session 2 more likely to resemble the visual arrangements of other participants? To explore whether seeing visual stimuli first causes participants to rely on visual features more later on in the experiment, we compared between-modality between-participant correlations across sessions 1 and 2 depending on whether the visual MA preceded the linguistic MA or vice versa. We found a mean between-modality between-participant correlation of 0.184 (Figure 2.4C - dark grey bar, one-sided test: $t(21) = 32.25$, $p < 0.0001$) when the visual MA in session 1 preceded the linguistic MA in session 2. The mean between-modality between-participant correlation when the linguistic MA preceded the visual MA was 0.131 (Figure 2.4C - light grey bar, one-sided test: $t(19) = 14.11$, $p < 0.0001$). One-tailed independent samples t-test

revealed that the between-modality correlations are significantly higher when the visual MA in session 1 preceded the linguistic MA in session 2 (Figure 2.4C, $t(42) = 5.053$, $p < 0.0001$). This result shows that the representational alignment between participants is affected by the task order where the preceding visual MA task seems to affect the linguistic similarity judgements in session 2 more than the preceding linguistic MA task affects the visual similarity judgements in session 2. The disparity in the cross-modal effect rules out the possibility that participants were actively attempting to remember their representational geometries from previous sessions. Rather, it suggests that there are different qualitative projections between the visual and linguistic geometries.

2.3.5 Difference in qualitative projections between visual and linguistic representational geometries

The previous analysis compared representational geometries between participants and task modalities. This uncovered the possibility that having been exposed to the visual modality in the first session somewhat distorts the geometry of the linguistic modality in the following session. To investigate the overall stability of the representational structure after the cross-modal contamination, we compared the within-modality correlations in session 1 (where no cross-modal effect of the similarity judgements was possible from prior exposure to the other modality) to within-modality correlations in session 2 (potential cross-modal effect from prior exposure to the other task modality). In the visual similarity judgements, we observed significant within-modality between-participant correlations of 0.193 in session 1 (one-sided test: $t(11) = 29.344$, $p < 0.0001$) and 0.146 in session 2 (one-sided test: $t(10) = 14.984$, $p < 0.0001$). Interestingly, the within-modality correlations when participants had not yet taken part in the linguistic MA (session 1) were significantly stronger compared to their second session (Figure 2.4D blue bars; independent t-test: $t(19) = 4.094$, $p < 0.0001$) when

they had prior exposure to the linguistic modality. This reduction in within-modality correlations suggests that seeding the representations from the linguistic modality introduced larger variability in the visual geometry between participants; perhaps reflecting differences in mental imagery or related processes between people.

This was further supported by a reduced within-modality between-participant correlation (Figure 2.4D green bars; Spearman's $\rho = 0.157$; one-sided test: $t(10) = 18.092$, $p < 0.0001$) when participants completed the linguistic MA in their first session (without having seen the visual stimuli before). It is possible that reading a sentence description of a visual scene elicits a broader range of representations and these vary between people depending on their prior experience (e.g., reading "Three women playing frisbee in a field" would likely elicit very different mental images across individuals). Importantly, in the second session of the linguistic similarity judgements, the representational geometries were significantly more aligned between participants (independent t-test: $t(21) = -7.851$, $p < 0.0001$) with the mean within-modality correlation of 0.239 (one-sided t-test: $t(11) = 38.810$, $p < 0.0001$). This suggests that prior exposure to the visual modality significantly constrained the representational geometry between participants. Critically, participants were not informed that they would be presented with stimuli from the other modality before completing the MA tasks. These results show a clear interaction where prior exposure to vision narrows down the potential representational candidates in language, and opposedly, prior exposure to language expands the space of representational candidates in vision. This explanation was further supported by cross-session within-modality comparisons where we found that the linguistic representational geometry was more stable across sessions and participants compared to the visual geometry, which displayed a higher variability between people (Supplementary Figure 2.9).

2.3.6 Individual differences outweigh within-modality similarity in vision but not in language

The next part of the analysis is concerned with individual differences in the representational geometry of vision and language. In the literature, individual differences have been associated with variability in similarity judgements across modalities (Shepard & Chipman, 1970) and brain activity patterns (Charest et al., 2014). Therefore, we expect to observe higher within-subject correlations than between-subject correlations. We obtained within-subject correlations where a participant's visual RDM was correlated with their own linguistic RDM. We then performed between-subject correlations where a participant's visual RDM was correlated with other participants' linguistic RDMs. The between-subject between-modality correlations were averaged, which provided us with one correlation per participant indicating how one's visual RDM compares to other participants' linguistic RDMs on average. This procedure was repeated for all participants who completed the visual and linguistic MA in their first or second sessions. The average within-subject correlation was 0.220 (one-sided test: $t(22) = 8.524$, $p < 0.0001$) and between-subject correlation was 0.160 (one-sided test: $t(22) = 14.453$, $p < 0.0001$). Consistent with our expectation, the paired samples t-test confirmed that a participant's representational geometry from the visual MA is more likely to resemble their geometry from the linguistic MA rather than the linguistic geometry of other participants (Figure 2.4E; $t(22) = 2.881$, $p = 0.004$).

To investigate the stability of individual differences and how they change as a function of the cross-modal projection, we also compared within-subject correlations to between-subject within-modality correlations. If individual differences weigh any of the task modalities more than the within-modality alignment between participants does, within-subject correlations should always be higher than within-modality between-subject correlations. To investigate this,

we correlated a participant's visual RDM with the visual RDMs of other participants. The between-subject correlations were averaged to indicate how one's visual RDM compares to the visual RDMs of other participants on average. This was repeated for all participants who completed the visual and linguistic MA in their first or second session. We found significantly higher within-subject correlations compared to the visual within-modality correlations between participants (mean Spearman's $\rho = 0.175$, $t(22) = 2.063$, $p = 0.026$). This result indicates that the visual representational geometry reflects individual differences (even if it is between modalities) rather than aligning with the visual geometries of other participants, likely because of the projection from session 1 linguistic judgements. We followed the same procedure to compare within-subject correlations to linguistic within-modality correlations. Interestingly, the paired samples t-test did not reveal a significant difference between the within-subject and between-subject within-modality correlations (mean Spearman's $\rho = 0.200$, $t(22) = 0.850$, $p = 0.202$). Note that in this analysis we did not take the task order effect into account, which further exposes the significance of cross-modal projection where linguistic geometry is more likely to be aligned between participants after exposure to the visual modality. Additional analysis where we controlled for prior exposure revealed a trend to the opposite effect suggesting that linguistic geometry has more variability before exposure (Supplementary Figure 2.10).

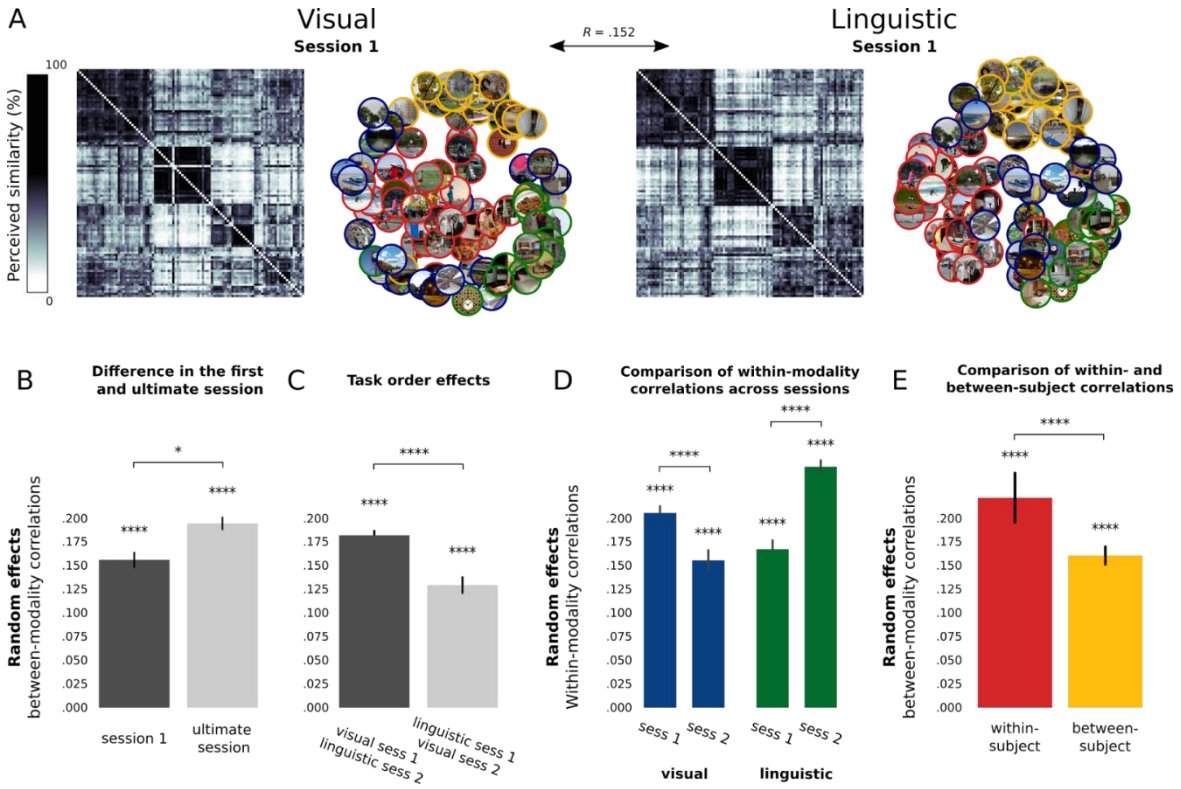


Figure 2.4 | Behavioural analysis to characterise representational alignment in visual and linguistic similarity judgements. (A) Rank-transformed averaged RDMs and multidimensional scaling (MDS) were used to visualise the configuration of the visual (left) and linguistic (right) modalities in session 1. Visual inspection of both reveals that participants generally agreed on the similarity judgements whether they started the experiment with visual or linguistic MA. Note that the captions in the linguistic MA are displayed as images only for this figure (displaying all sentences would result in excessive crowding). The images in the MDS are colour-coded to differentiate between the categories (red- human, yellow - animal, green - indoor, blue - outdoor) and the distance between them depicts their perceived dissimilarity (for comparison to the control condition, see Supplementary Figure 2.11). (B) Comparison of between-modality correlations in the first and ultimate session. (C) Comparison of between-modality correlations across sessions 1 and 2 depending on whether the visual MA preceded the linguistic MA or vice versa. (D) Comparison of within-modality within-session correlations to investigate the stability of visual and linguistic representation. (E) Comparison of within- and between-subject between-modality correlations to investigate the role of

individual differences. The stars denote p-values with: *$p < .05$, ** $p < .01$, *** $p < .001$, **** $p < .0001$.

2.4 Discussion

In this study, we collected multimodal similarity judgements to investigate whether representations in vision and language draw on a shared representational system. We showed that similarity judgements in vision reliably modelled brain responses to viewing natural scene images that were previously accounted for by sentence embeddings from LLMs (Doerig et al., 2024; Lahner et al., 2023; Tang et al., 2023). Perhaps the most compelling finding was that we reliably modelled the same brain responses using similarity judgements of sentence captions. This relationship is even more significant given that the similarity judgements were made by participants who had never seen the images before and were found to display a higher inter-individual variability compared to visual similarity judgements. These findings provide support to prior research that showed that language taps into non-linguistic regions (Fedorenko et al., 2011; Wehbe et al., 2014), and raise a question as to what extent the visual brain is language-aligned (Doerig et al., 2024). Combined with the fact that we found the LLMs to provide an excellent objective of predicting behavioural similarity judgements, we speculate that vision and language converge in their ability to recognise and encode statistical regularities of the visual world into a shared multidimensional similarity space.

Sentence reading relies on the ability to understand the syntactic structure (Humphries et al., 2006; Cutler, & Clifton, 2000; Kjelgaard, & Speer, 1999; Luo, & Zhou, 2010), interpret its semantic meaning (Hagoort et al., 2004; Heilbron et al., 2022; Kutas, & Hillyard, 1980; Schrimpf et al., 2021), and visually reconstruct the sentence's utterances (Dils, & Boroditsky, 2010; Meyer et al., 2007; Stanfield, &

Zwaan, 2001). We showed that participants agreed on the configuration of the arrangements regardless of whether they started the experiment with visual or linguistic similarity judgements. This is an intriguing finding because it suggests that there is some similarity in the way the brain encodes the gist of a scene in sentence reading and image viewing. We showed that both geometries align with many of the same regions that were previously associated with a distributed semantic network in vision (Aguirre et al., 1998; Charest et al., 2014; Cichy et al., 2017; Cichy et al., 2019; Epstein, & Kanwisher, 1998; Haxby et al., 2001; Ishai et al., 1999; Kanwisher et al., 1997; Kiani et al., 2007; McCarthy et al., 1997; Mur et al. 2013). Similarly, prior evidence suggested that even language is supported by a distributed semantic network (Cox et al., 2024; Huth et al., 2016; Popham et al., 2021; Tang et al., 2021; Tang et al., 2023). Consistent with Doerig and colleagues (2024), it is likely that vision and language converge in their inherent ability to recognise and encode regularities of the visual world onto a shared multidimensional space that is functionally organised along multiple dimensions (Bracci, & Op de Beeck 2016; Bracci et al., 2017; Martin, 2007; Martin et al., 2018) including dimensions of perceived dissimilarity (Cichy et al., 2019; Hebart et al., 2020; Magri, & Konkle, 2019). The behavioural predictions of visual brain responses clearly show that the visuo-linguistic alignment is embedded by underlying neural mechanisms of the mid- and high-level visual regions. This alignment also provides behavioural relevance to the representational structure of LLM embeddings because it shows that, in addition to neural mechanisms of visual (Doerig et al., 2024) and linguistic (Russo et al. 2022; Caucheteux, & King, 2022; Dong, & Toneva, 2023; Goldstein et al., 2021; Jain, & Huth, 2018; Toneva, & Wehbe, 2019; Wehbe et al., 2014) processing, they are also effective in capturing processes of human cognition, such as similarity judgements.

The qualities of the shared multidimensional similarity space are further demonstrated by the difference in qualitative projection between the visual and

linguistic representational geometry. We found that the linguistic within-modality similarity increased if participants performed visual similarity judgements first. The main assumption of the embodied model of cognition is that language (and semantics in general) requires support from sensorimotor systems to retrieve the true representations of the world (Barsalou, 2008; Binder, & Desai, 2011; Fedorenko et al., 2011; Wehbe et al., 2014). Consistent with this, it is possible that the semantic demands of linguistic similarity judgements forced participants to re-use modality-specific representations that derived from prior experience in the visual modality. A similar effect has been described in artificial representation learning where an adequate visuo-linguistic alignment was found to significantly improve models' performance in various computational and predictive tasks (Conwell et al., 2022; Conwell et al., 2023; Kiela, & Bottou, 2014; Silberer, & Lapata, 2014; Wang et al., 2023). Similarly, our findings indicate that the multidimensional space is characterised by its ability to flexibly embed language in previous experience even when it is not explicitly required for the task in hand, likely as a result of the visual brain's enormous storage capacity (Brady et al., 2008). It is worth noting here that participants had never been informed that the stimuli covary in content therefore it is unlikely that this effect was due to participants actively attempting to memorise the appearance of the scenes in the first experimental session.

How does the multidimensional space support similarity judgements if the visual cue is absent? A surprising result was that we found the opposite effect if participants performed linguistic similarity judgements first. We observed that the interindividual variability in the visual representational geometry increased if it was preceded by the linguistic modality. As demonstrated by the higher inter-individual variability in linguistic geometries from session 1, it is likely that linguistic similarity judgements accentuated individually relevant representations of one's language, which then modulated the configuration of visual

arrangements in the following session. This is an intriguing finding because it suggests that prior exposure to linguistic semantics shapes the perceived dissimilarity of visual representations in real-time, which is a characteristic previously only described in developmental research into early representation learning (LaTourrette, & Waxman, 2020; LaTourrette et al., 2023; Waxman, & Markow, 1995). Together with the fact that we found the behavioural geometries to align significantly better with the representational structure of LLM embeddings rather than the structure of vision- and category-computable models, language may provide the best representational format to describe the internal geometry of the common multidimensional space.

The question regarding the extent to which the visual brain is language-aligned has been at the forefront of cognitive neuroscience in recent years (Doerig et al., 2024; Tang et al., 2021; Tang et al., 2023). Here, we show that vision and language converge onto a multidimensional space of perceived dissimilarities. The space is language-aligned in that language not only maps into this space but also seems to shape its internal geometry by warping visual representations based on previous experience with the linguistic stimuli. Providing that the behavioural geometries aligned with the visual brain responses, the findings provide some support to recent studies that suggest that the human visual system relies on complex neural computations that enable sensory inputs to be transformed into a format that is retrievable by language (Doerig et al., 2024; Draschkow, & Võ, 2017). We speculate that language may provide the best means of describing the shared multidimensional space for several reasons. First, language activates individually unique representations that provide an insight into how one's language shapes the visual brain (Althaus, & Mareschal, 2014; Ferry et al., 2013; LaTourrette, & Waxman, 2020; LaTourrette et al., 2023; Waxman, & Markow, 1995). Second, language displays the ability to create a representation that is universal enough to warp many possible representations (Lupyan, &

Spivey, 2010; Lupyan, 2012; Winawer et al., 2007). Finally, we found that introducing the covert characteristic to the visual modality (Supplementary Figure 2.1) did not improve the representational alignment with the linguistic geometry (Supplementary Figure 2.11). This further shows that the unique role language plays in drawing on individually relevant representations cannot be attributed to the covert characteristic but instead is specific to language itself. Future research should explore how linguistic semantics intersects with vision, and how the multimodal representations converge in the brain representational space. Together, we anticipate our results to be the first step towards establishing a shared multidimensional semantic space that encodes behaviourally relevant representations about sensory experience.

2.5 Methods

2.5.1 Participants

63 participants (23 ± 4 years of age) with normal or corrected-to-normal vision participated in the experiment. Participants provided informed consent as part of a protocol approved by the University of Birmingham Research Ethics Committee and were awarded £8/h for participation.

2.5.2 Stimuli

The stimulus set was obtained from the Natural Scenes Dataset (NSD; Allen et al., 2022) and consisted of 100 natural scene images and 100 sentence captions describing the images. A greedy algorithm (Allen et al., 2022) selected the set of 100 images from a special set of 515 images such that they maximally span the high-level semantic space, which was defined by embeddings of the images' sentence captions. The sentence captions were collected from five human annotators as part of the Microsoft Common Objects in Context database (MS COCO; Lin et al., 2014). We identified the most frequent words and phrases across the five annotators and combined these to generate one representative sentence caption for every image.

2.5.3 Procedure

Participants completed the visual, linguistic, and control multiple arrangement tasks in three in-person sessions scheduled at least 7 days apart (average time between sessions 8.3 ± 3.3 days). We used the Latin square design to counterbalance the task order across participants. The experiment was hosted by the web-based platform Meadows (https://meadows-research.com/) and administered on a PC with a 25" monitor. The visual MA task (Figure 2.1A left)

followed the standard procedure as introduced by Kriegeskorte and Mur (2012). Briefly, 100 natural scene images were displayed around a white circular arena. Participants were then instructed to arrange the images inside the white arena according to their similarity using mouse drag-and-drop operations such that images perceived as more similar are placed closer together while images perceived as less similar are placed further apart. The linguistic MA task (Figure 2.1A, right) was conceptually similar, but involved the arrangement of sentence captions which allowed us to recover the representational geometry of the images' linguistic referents. To reduce the clutter introduced by the display of numerous lengthy sentences around the MA arena, we adapted the MA method such that every sentence was initially displayed as an asterisk. The sentence appeared only when the participant hovered the mouse cursor over the asterisk and sustained as the participant moved the location of this item. Other than that, the procedure of the linguistic MA was identical to the visual MA.

While the high-level representation is instantly accessible through direct perception in the visual modality, the representation in the linguistic modality is covert and relies on internal simulation of the objects and the relationships between them. This introduced the possibility that the linguistic MA is cognitively more demanding than the visual MA task. To control for the cognitive load of the overt visual and covert linguistic modality, we introduced a control condition in which the natural scene images were initially displayed as white rectangles (Supplementary Figure 2.1). As was the case for captions in the linguistic MA task, in the control task the images appeared only when participants hovered the mouse cursor over the white rectangle and as they moved the item in the arena. This method allowed us to employ the covert characteristic of the linguistic modality where participants are required to mentally simulate and maintain the representation. Other than that, the procedure was the same as in the remaining two tasks.

2.5.3.1 Estimating the representational dissimilarity matrices

The MA method used the weighted-averaging approach described by Kriegeskorte and Mur (2012) to obtain the representational dissimilarities. Following the first MA trial in which the full set of 100 stimuli was displayed, an algorithm computed Euclidean distances between every pair of stimuli ((100 items2 - 100)/2 = 4950 pairs) inside the MA arena and assembled them in an initial RDM symmetric about a diagonal of zeros. Subsets of stimuli that showed the weakest weighted evidence within the RDM were then presented in subsequent MA trials. With every new trial, the representational dissimilarities in the initial RDM were iteratively adjusted by the weighted-averaging approach reducing potential placement errors. The MA trials terminated either when enough evidence was obtained for every pair of stimuli or once 45 minutes elapsed. This yielded three modality-specific RDMs per participant.

2.5.4 Construction of category- and sentence-computable RCNN RDMs

The set of 100 natural scene images was passed through two recurrent convolutional neural networks (RCNNs) whose objective was to produce either a sentence caption or a category label of the image. The architecture of the network was introduced in detail by Doerig and colleagues (2024). Briefly, the network was derived from vNet and was adapted to include both lateral and top-down recurrent connections (Kietzmann et al., 2019). As described by Doerig and colleagues (2024), both RCNNs shared identical architecture and were trained with 10 random seeds on the MS COCO dataset with the NSD stimuli excluded. The only difference between the sentence- and category-computable RCNN is in the objective of the output layer. The sentence-computable RCNN minimised the cosine distance between the predicted and the target sentence embedding obtained from MPNet. The category-computable RCNN, on the other hand, used

multi-hot encoding to predict the category label. After evaluating both RCNNs on the set of 100 natural scene images, we extracted sentence embeddings and category vectors for the last layer and time step from each of 10 random seeds. We calculated correlation distance between the output vectors from each RCNN and assembled the dissimilarity estimates in model RDMs.

2.5.5 Construction of searchlight brain RDMs

Searchlight brain RDMs were constructed using NSD fMRI that were collected from 8 participants who viewed the set of 100 natural scene images in a recognition task (Allen et al., 2022). Activity patterns were extracted in a sphere with a radius of 6 voxels from each participants' single-trial beta weights. We then calculated pairwise correlation distances between the activity patterns and assembled them in a searchlight brain RDM.

2.5.6 Cross-validated non-negative least squares

We created a behaviour-predicted RDM to investigate how well the behavioural geometry in vision and language aligns with patterns of visually evoked activity, and the representational structure of category- and sentence-computable RCNNs. Behavioural RDMs across all sessions were used unless stated otherwise. The NSD stimulus set was first sampled into train and test sets at 70:30 ratio using 10-fold cross-validation. For each fold, pairwise dissimilarities in behavioural and observed RDMs (either brain RDM or RCNN RDM) were subsetted using the train and test sets. The train subset of behavioural RDMs was then fitted on the train subset of the observed RDM using non-negative least squares regression. A major advantage of this approach is that it allows us to maximise information from the behavioural RDMs by assigning flexible weights each representing a participant's relative fit to the train subset of the observed RDM. The NNLs weights were then

used to create a behaviour-predicted RDM that is the best possible representation of the observed representational geometry. This was done using the dot product between the NNLs weights and the test subset of behavioural RDMs. We then evaluated the fit between the behaviour-predicted RDM and observed RDM using Pearson correlation.

For searchlight brain RDMs, this procedure was repeated for each searchlight RDM and each NSD participant separately. Pearson correlations were averaged across folds for each NSD participant and projected from the original funct1pt8mm space to fsaverage using nsdcode (https://github.com/cvnlab/nsdcode/tree/master). For group statistics, Pearson correlations were averaged across NSD participants and activation maps were plotted using the pycortex (Gao et al., 2015) and nilearn (https://github.com/nilearn/nilearn) libraries. The significance of correlations was tested using one-sided t-test across participants and corrected for multiple comparisons at FDR $p < 0.05$. For RCNN comparisons, Pearson correlations were averaged across 10 seeds for each layer and time step and tested for significance using one-sided t-test corrected for multiple comparisons using FDR $p < 0.05$. To test the difference in correlations of category- and sentence-computable models with behaviour-predicted RDMs, Pearson correlations were subtracted and tested for significance using one-sided t-test corrected at FDR $p < 0.05$.

2.5.7 Random effects analysis on correlations of behavioural RDMs

We used the rsatoolbox (van den Bosch, in preparation; https://github.com/rsagroup/rsatoolbox/tree/main) for all operations and comparisons of the RDMs. Behavioural RDMs were correlated across any two conditions using Spearman's correlation from the rsatoolbox, which returned a single matrix of all pairwise correlations.

2.5.7.1 Between-modality correlations compared within- and across-sessions

Pairwise between-modality correlations for visual and linguistic RDMs were obtained by correlating every possible pair of participant RDMs within-session. Only participants who completed the visual (respective linguistic) MA task in the first session and the linguistic (respective visual) MA task in the last session were included in the analysis. We compared all pairwise between-modality correlations in session 1 to all pairwise between-modality correlations in session 1 using one-tailed paired samples t-test to investigate whether the representational alignment increases as a participant progresses through the experiment. To investigate how the task order affects representational alignment between participants, we correlated RDMs between-modality across session 1 and 2 (e.g., visual RDMs in session 1 correlated with linguistic RDMs in session 2 and vice versa). This analysis was restricted to participants who completed the task modalities in the same order such that the modality in session 1 and 2 was identical across participants. Within-subject correlations were excluded from the pairwise correlations by only including the off-diagonal elements of the correlation matrix. One-tailed independent samples t-test was used to investigate whether the between-modality across-session correlations are higher when the visual MA precedes the linguistic MA.

2.5.7.2 Investigating the stability of representational geometry within-modality

To investigate the stability of the representational geometry across participants and sessions, we correlated RDMs from one modality with all RDMs from within the same modality. The correlations were performed either within-session (e.g., visual RDMs from session 1 correlated with visual RDMs from the same session), or across sessions (e.g., visual RDMs from session 1 correlated with visual RDMs from session 2). For comparisons within-session, we only included the off-diagonal elements of the pairwise correlation matrix because elements on

the diagonal were correlations between identical pairs of RDMs. Independent samples t-test was used to compare the within-modality correlations either across sessions or between task modalities.

2.5.7.3 Comparison of individual differences and within-modality similarity

To investigate the influence of individual differences in between-modality similarity, we compared within- and between-participant correlations. For a given pair of task modalities, we correlated RDMs across all participants and compared the within-participant correlations on the diagonal of the pairwise correlation matrix to between-participant correlations off-diagonal of the pairwise correlation matrix. The between-participant correlations were averaged to obtain a between-participant correlation value per participant. One-tailed paired samples t-test was used to determine whether the between-modality correlation is higher within-participants. We also compared the within-participant correlations to between-participant within-modality correlations to determine the relative weighing of individual differences and within-modality similarity. We correlated RDMs for a given session within the same modality. Elements on the diagonal of the pairwise correlation matrix were excluded as these were correlations between identical pairs of RDMs. The between-participant correlations off-diagonal of the correlation matrix were then averaged to obtain an average between-participant within-modality correlation per participant. The between-participant within-modality correlations were then compared to within-participant between-modality correlations for a given pair of task modalities (e.g., correlations within-modality in visual RDMs compared to within-participant between-modality correlations of visual and linguistic RDMs) using paired samples t-test.

2.6 References

Aguirre, G. K., Zarahn, E., & D'Esposite, M. (1998). An area within human ventral cortex sensitive to "building" stimuli: evidence and implications. *Neuron, 21*(2), 373-383.

Allen, J. E., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., Hutchinson, J. B., Naselaris, T., & Kay, K. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience, 25*, 116-126.

Althaus, N., & Mareschal, D. (2014). Labels direct infants' attention to commonalities during novel category learning. *PLoS One, 9*(7).

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59*, 617-645.

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences, 15*(11), 527-536.

Bracci, S., & Op de Beeck, H. (2016). Dissociations and associations between shape and category representations in the two visual pathways. *The Journal of Neuroscience, 36*(2), 432-444.

Bracci, S., Ritchie, J. B., & Op de Beeck, H. (2017). On the partnership between neural representations of object categories and visual features in the ventral visual pathway. *Neuropsychologia, 105*, 153-164.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences of the United States of America, 105*(3), 14325-15329.

Charest, I., Kievit, R. A., Schmitz, T. W., Deca, D., & Kriegeskorte, N. (2014). Unique semantic space in the brain of each beholder predicts perceived similarity. *Proceedings of the National Academy of Sciences of the United States of America, 111*(40), 14565-14570.

Caucheteux, C., & King, J. R. (2022). Brains and algorithms partially converge in natural language processing. *Communitations Biology, 5*(1).

Cichy, R. M., Kriegeskorte, N., Jozwik, K. M., van den Bosch, J. J. F., & Charest, I. (2017) Neural dynamics of real-world object vision that guide behaviour. *bioRxiv*.

Cichy, R. M., Kriegeskorte, N., Jozwik, K. M., van den Bosch, J. J. F., & Charest, I. (2019). The spatiotemporal neural dynamics underlying perceived similarity for real-world objects. *NeuroImage, 194*, 12-24.

Conwell, C., Prince, J. S., Hamblin, C. J., & Alvarez, G. A. (2023). Controlled assessment of CLIP-style language-aligned vision models in prediction of brain & behavioural data. *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Conwell, C., Prince, J. S., Kay, K., Alvarez, G. A., & Konkle, T. (2022). What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *bioRxiv*.

Cox, C. R., Rogers, T. T., Shimotake, A., Kikuchi, T., Kunieda, T., Miyamoto, S., Takahashi, R., Matsumoto, R., Ikeda, A., & Lambon Ralph, R. A. (2024). Representational similarity learning reveals a graded multidimensional semantic space in the human anterior temporal cortex. *Imaging Neuroscience, 2*, 1-22.

Cutler, A., & Clifton, C. (2000). *Comprehending spoken language: a blueprint of the listener*. In Brown, C. M., & Hagoort, P. The Neurocognition of Language. Oxford University Press.

Dils, A. T., & Boroditsky, L. (2010). Visual motion aftereffect from understanding motion language. *Proceedings of the National Academy of Sciences of the United States of America, 107*(37), 16396-16400.

Doerig, A., Kietzmann, T. C., Allen, E., Wu, Y., Naselaris, T., Kay, K., & Charest, I. (2024). Visual representations in the human brain are aligned with large language models. *bioRxiv*.

Dong, D. T., & Toneva, M. (2023). Vision-language integration in multimodal video transformers (partially) aligns with the brain. *arXiv*.

Draschkow, D., & Vo, M. L. (2017). Scene grammar shapes the way we interact with objects, strengthens memories, and speeds search. *Scientific Reports, 7*(1), 16471.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. Nature, 392(6676), 598-601.

Fedorenko, E., Behr, M. K., & Kanwisher, N. (2011). Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences of the United States of America, 108*(39), 16428-16433.

Ferry, A. L., Hespos, S. J., & Waxman, S. R. (2013). Nonhuman primate vocalizations support categorization in very young human infants. *Proceedings of the National Academy of Sciences of the United States of America, 110*(38), 15231-15235.

Gao, J. S., Huth, A. G., Lescroart, M. D., & Gallant, J. L. (2015). Pycortex: an interactive surface visualizer for fMRI. *Frontiers in Neuroinformatics, 9*.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., *et al.* (2021). Thinking ahead: spontaneous prediction in context as a keystone of language in humans and machines. *bioRxiv.*

Hagoort, P., Hald, L., Bastiaansen, M., Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science, 304*(5669), 438-441.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293*(5539), 2425-2430.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour, 4*(11), 1173-1185.

Heilbron, M., Armeni, K., Schoffelen, J. M., Hagoort, P., & de Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences of the United States of America, 119*(32).

Hernandez, D., Kaplan, J., Henighan, T., & McCandlish, S. (2021). Scaling laws for transfer. *arXiv*.

Humphries, C., Binder, J. R., Medler, D. A., & Liebenthal, E. (2006). Syntactic and semantic modulation of neural activity during auditory sentence comprehension. *Journal of Cognitive Neuroscience, 18*(4), 665-679.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature, 532*(7600), 453-458.

Ishai, A., Ungerleider, L. G., Martin, A., Schouten, J. L., & Haxby J. V. (1999). Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Sciences of the United States of America, 96*(16), 9379-9384.

Jain, S., & Huth, A. G. (2018). Incorporating context into language encoding models for fMRI. *bioRxiv.*

Jozwik, K. M., Kriegeskorte, N., Cichy, R., & Mur, M. (2018). Deep convolutional neural networks, features, and categories perform similarly at explaining primate high-level visual representations. *2018 Conference on Cognitive Computational Neuroscience.*

Kaplan, J., McCandlidh, S., Heinghan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. *arXiv*.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience, 17*(11), 4302-4311.

Khalig-Razavi, S. M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology, 10*(11).

Kiani, R., Esteky, H., Mirpour, K., & Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. Journal of Neurophysiology, 97(6), 4296-4309.

Kiela, D., & Bottou, L. (2014). Learning image embeddings using convolutional neural networks for improved multi-modal semantics. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 36-45.

Kietzmann, T. C., Spoerer, C. J., Sorensen, L. K. A., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the*

National Academy of Sciences of the United States of America, 116(43), 21854-21863.

Kjelgaard, M. M., & Speer, S. R. (1999). Prosodic facilitation and interference in the resolution of temporary syntactic closure ambiguity. *Journal of Memory and Language, 40*(2), 153-194.

Kriegeskorte, N., & Mur, M. (2012). Inverse MDS: inferring dissimilarity structure from multiple item arrangements*. Frontiers in Psychology, 3*.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science, 207*(4427), 203-205.

Lahner, B., Dqivedi, K., Iamshchinina, P., Graumann, M., Lascelles, A., Roig, G., Gifford, A. T., Pan, B., Jin, S., Murty, N. A. R., Kay, K., Oliva, A., & Cichy, R. (2024). Modeling short visual events through the BOLD moments video fMRI dataset and metadata. *Nature Communications, 15*(1).

LaTourrette, A. S., & Waxman, S. R. (2020). Naming guides how 12-month-old infants encode and remember objects. *Proceedings of the National Academy of Sciences of the United States of America, 117*(35), 21230-21234.

LaTourrette, A., Chan, D. M., & Waxman, S. R. (2023). A principled link between object naming and representation is available to infants by seven months of age. *Scientific Reports, 13*(14328).

Lin, T. Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Lawrence Zitnick, C., & Dollar, P. (2014). Microsoft COCO: common objects in context. *arXiv*.

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining task-agnostic visuolinguistic representations for vision-and-langauge tasks. *arXiv*.

Luo, Y., & Zhou, X. (2010). ERP evidence for the online processing of rhythmic pattern during Chinese sentence reading. *Neuroimage, 49*(3), 2836-2849.

Lupyan, G., & Spivey, M. J. (2010). Making the invisible visible: verbal but not visual cues enhance visual detection. *PLoS One, 5*(7).

Lupyan, G., & Swingley, D. (2012). Self-directed speech affects visual search performance. *Quarterly Journal of Experimental Psychology, 65*(6), 1068-1085.

Magri, C., & Konkle, T. (2019). Comparing facets of behavioral object representation: implicit perceptual similarity matches brains and models. *2019 Conference on Cognitive Computational Neuroscience*, 90-93.

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology, 58(*1), 25-45.

Martin, C. B., Douglas, D., Newsome, R. N., Man, L. L., & Barense, M. D. (2018). Integrative and distinctive coding of visual and conceptual object features in the ventral visual stream. *eLife, 7.*

McCarthy, G., Pruce, A., Gore, J. C., & Allison, T. (1997). Face-specific processing in the human fusiform gyrus. *Journal of Cognitive Neuroscience, 9*(5), 605-610.

Meyer, A. S., Belke, E., Telling, A. L., & Humphreys, G. W. (2007). Early activation of object names in visual search. *Psychonomic Bulletin & Review, 14*(4). 710-716.

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A., & Kriegeskorte, N. (2013). Human object-similarity judgements reflect and transcend the primate-IT object representation. *Frontiers in Psychology, 4.*

Popham, S. F., Huth, A. G., Bilenko, N. Y., Deniz, F., Gao, J. S., Nunez-Elizalde, A. O., & Gallant, J. L. (2021). Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature Neuroscience, 24*(11), 1628-1636.

Pramanick, S., Jing, L., Nag, S., Zhu, J., Shah, H., LeCun, Y., & Chellappa, R. (2022). VoLTA: Vision-Language Transformer with Weakly-Supervised Local-Feature Alignment. *arXiv*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *arXiv*.

Russo, A. G., Ciarlo, A., Ponticorvo, S., Di Salle, F., Tedeschi, G., & Esposito, F. (2022). Explaining neural activity in human listeners with deep learning via natural language processing of narrative text. *Scientific Reports, 12*(1).

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: integrative modelling converges in predictive processing.

Proceedings of the National Academy of Sciences of the United States of America, 118(45).

Shepard, R. N., & Chipman, S. (1970). Second-order isomorphism of internal representations: shapes of states. *Cognitive Psychology, 1*(1), 1-17.

Silberer, C., & Lapata, M. (2014). Learning grounded meaning representations with autoencoders. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 1*, 721-732.

Stanfield, R. A., & Zwaan, R. A. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science, 12*(2), 153-156.

Tan, H., & Bansal, M. (2019). LXMERT: Learning cross-modality encoder representations from transformers. *arXiv*.

Tang, J., LeBel, A., & Huth, A. G. (2021). Cortical representations of concrete and abstract concepts in language combine visual and linguistic representations. *bioRxiv*.

Tang, J., Du, M., Vo, V. A., Lal, V., & Huth, A. G. (2023). Brain encoding models based on multimodal transformers can transfer across language and vision. *arXiv*.

Toneva, M., & Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural language processing (in the brain). *arXiv.*

Tuckute, G., Sathe, A., Srikant, S., Taliaferro, M., Wang, M., Schrimpf, M., Kay, K., & Fedorenko, E. (2024). Driving and suppressing the human language network using large language models. *Nature Human Behaviour, 8*(3), 544-561.

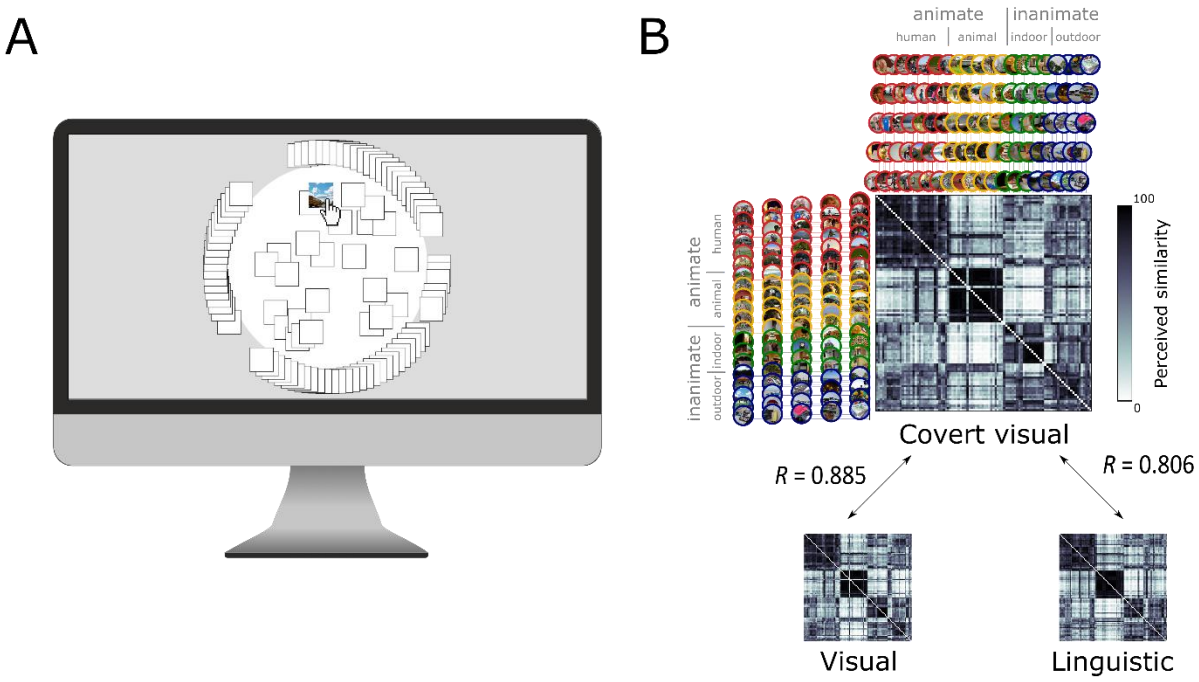
Wang, A. Y., Kay, K., Naselaris, T., Tarr, M. J., & Wehbe, L. (2023). Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence, 5*, 1415-1426.

Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: evidence from 12- to 13-month-old infants. *Cognitive Psychology, 23*(3), 257-302.

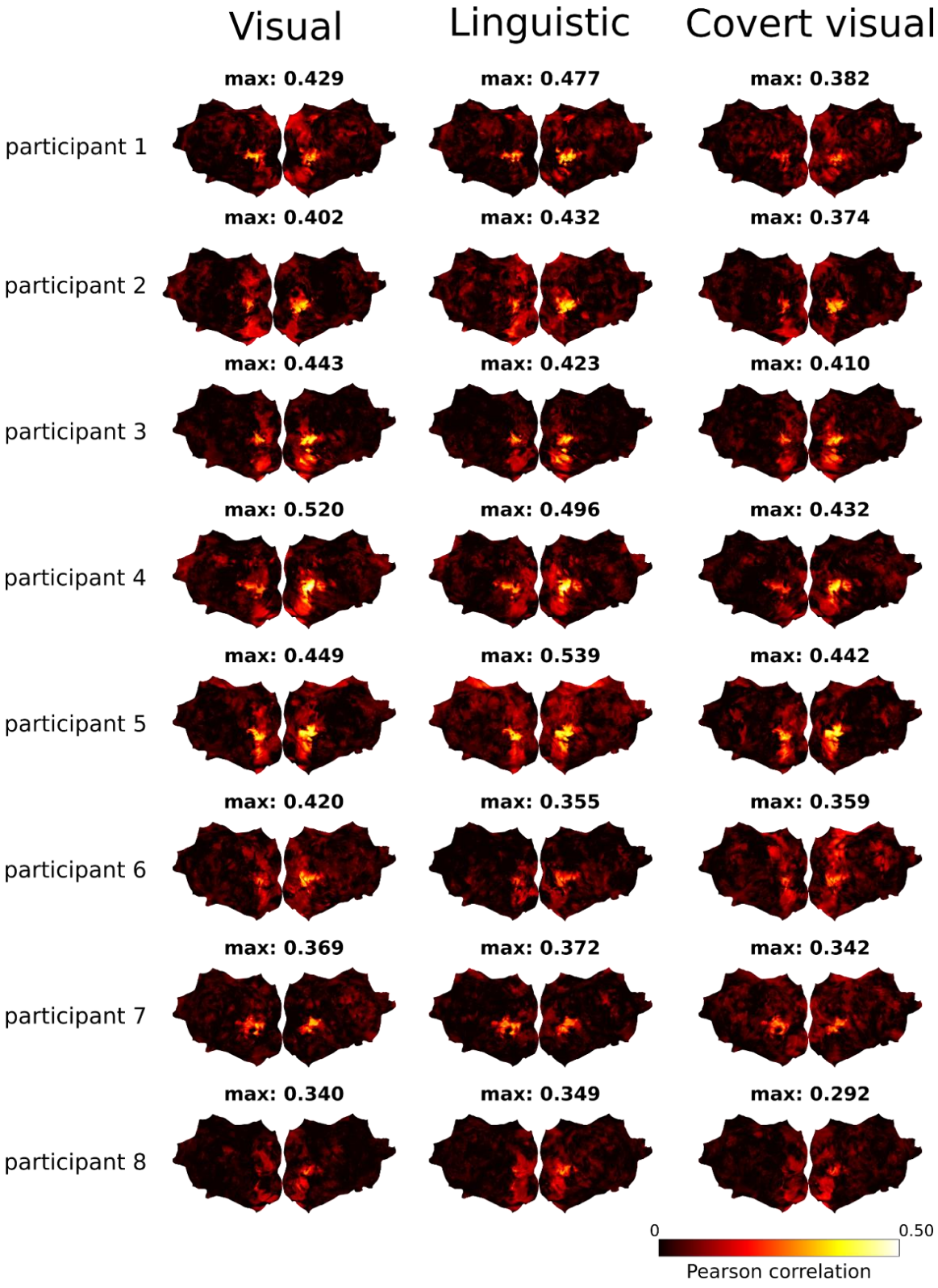
Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., & Mitchell, T. (2014). Simultaneously uncovering the patterns of brain regions involved in different story reading suprocesses. *PLoS One, 9*(11).

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences of the United States of America, 104*(19), 7780-7785.

2.7 Supplementary material

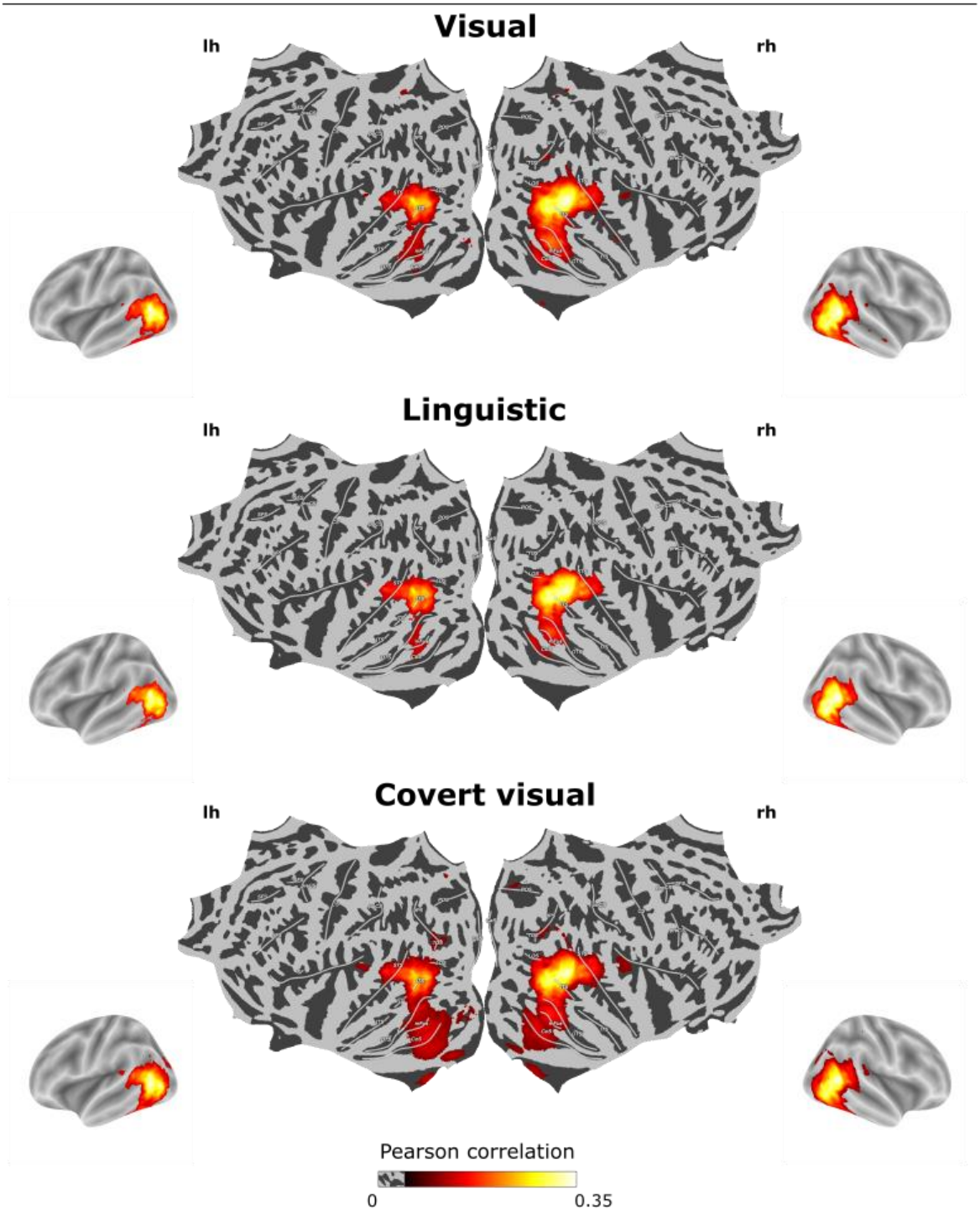


Supplementary Figure 2.1 | Experimental design of the control MA task. (A) We included an additional condition to control for the cognitive load between visual and linguistic modalities. Participants were presented with the same set of 100 images; only these were displayed as white rectangles. The images appeared one at a time as the participant dragged or hovered over them with the mouse cursor. The same procedure was followed as in the other MA tasks. (B) RDMs across all sessions from the control condition were averaged and rank transformed. The control RDM significantly correlated with both modalities but the correlation is slightly higher with the visual RDM (Spearman's ρ = .885).



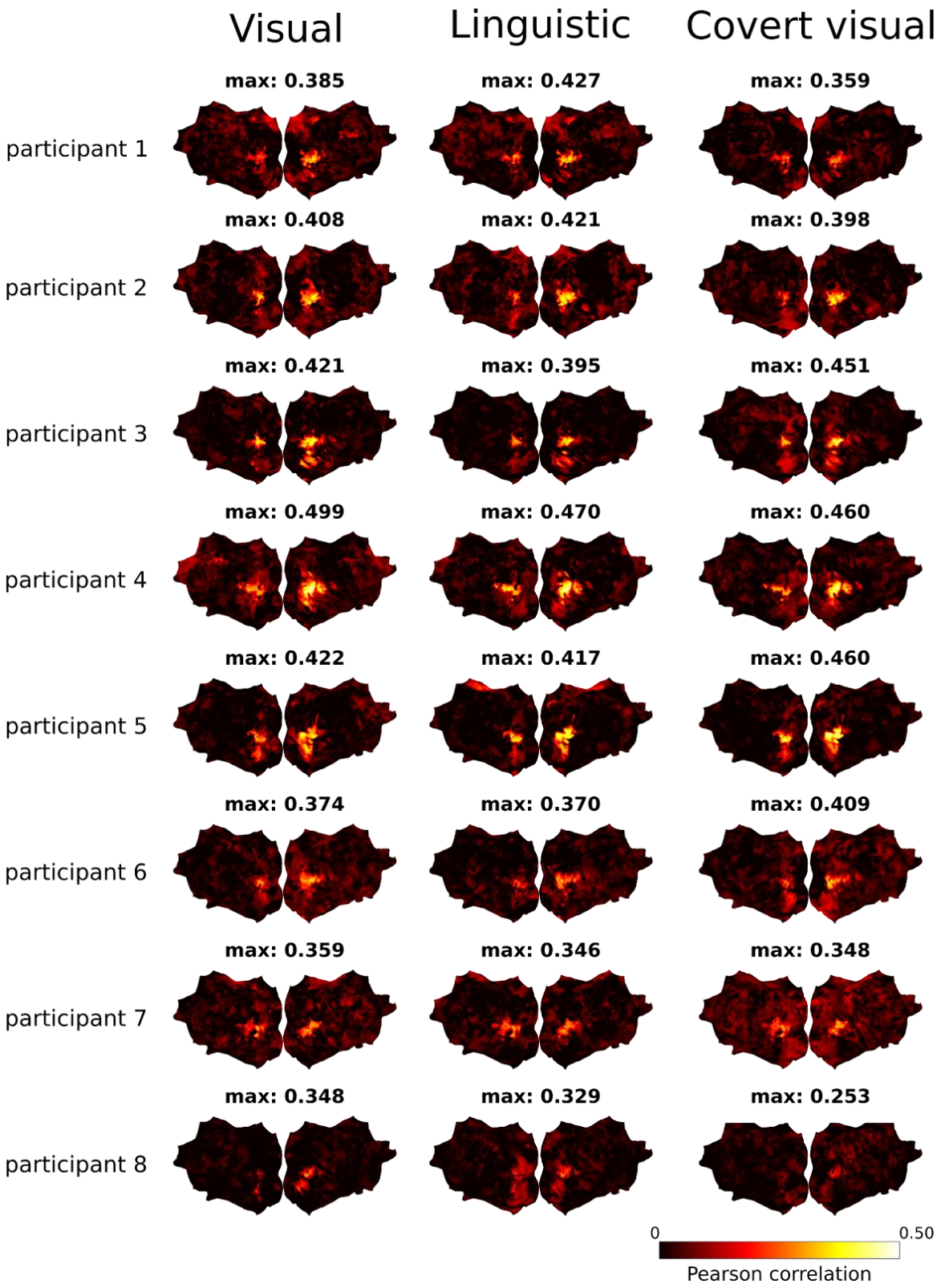
Supplementary Figure 2.2 | Representational alignment of multimodal similarity judgements with brain responses at the participant-level. For each NSD participant, Pearson correlations between the behaviour-predicted brain and observed brain RDMs

were averaged across 10 cross-validation folds. Representational alignment is reported for all three task modalities. Similar to group-level comparisons, the surface maps reveal strong and relatively stable correlations across all NSD participants and tasks bilaterally along the occipitotemporal cortex.



Supplementary Figure 2.3 | Representational alignment between the brain responses and multimodal similarity judgements from session 1. To demonstrate that representational alignment remains even before behavioural participants were familiar with stimuli from the other task modality, only behavioural data from session 1 were used to create the behaviour-predicted brain RDMs. The figure shows group-

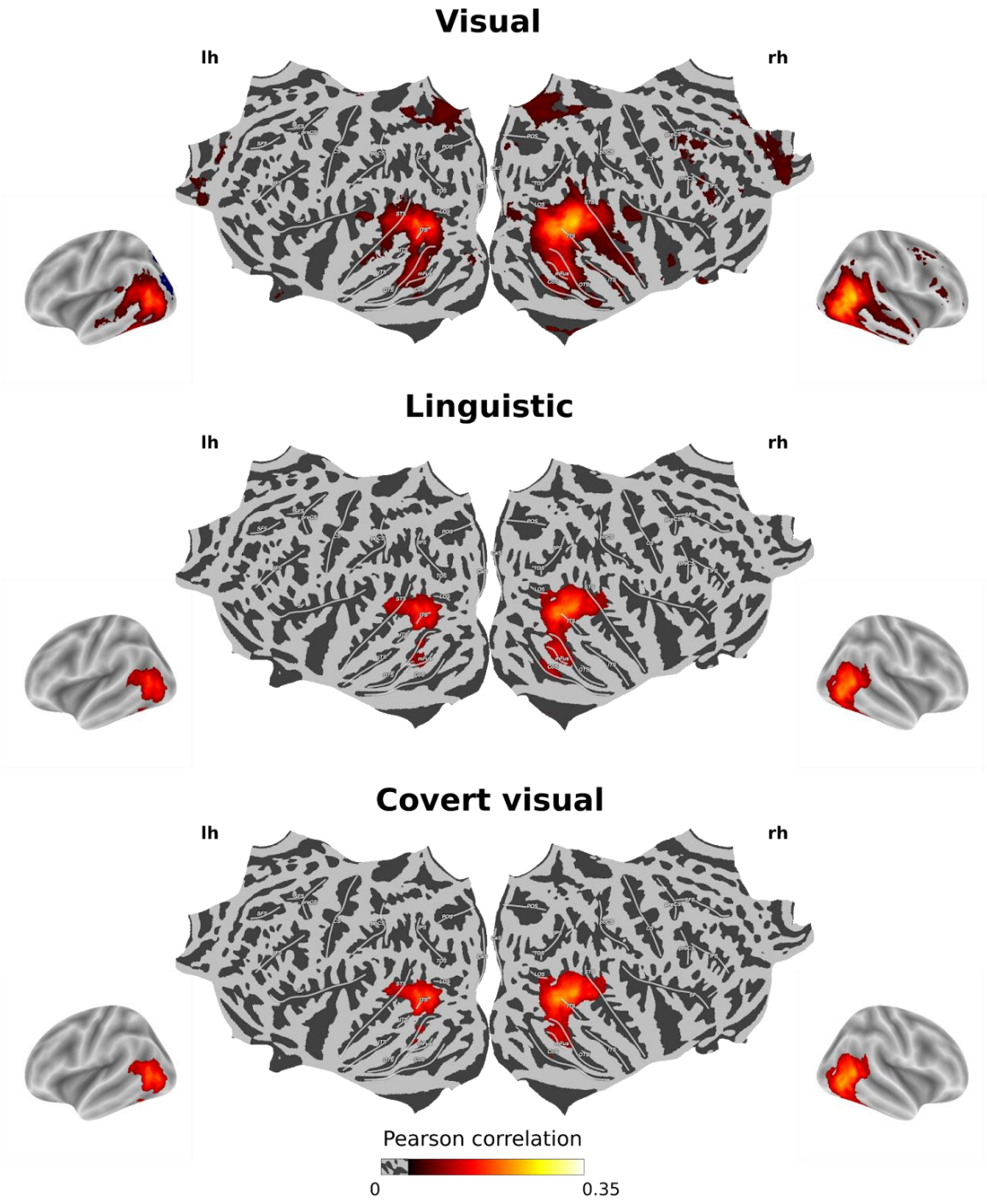
averaged Pearson correlations between the behaviour-predicted and observed brain RDMs ($p < 0.05$, FDR-corrected, one-sided test). The surface maps reveal a significant representational overlap between the RDMs spanning the occipitotemporal cortex similar to what was shown in Figure 2.2B. The Pearson correlations peak at 0.347, 0.338, and 0.346 for visual, linguistic, and covert visual modality respectively. This result is reassuring because it demonstrates that even though behavioural participants in the linguistic judgements had never seen the visual scenes before, they still configured a representational geometry that is predictive of brain responses to viewing the corresponding natural scene images.



Supplementary Figure 2.4 | Representational alignment between the brain responses and multimodal similarity judgements from session 1 at the participant-level. The figure shows Pearson correlations between the behaviour-predicted and

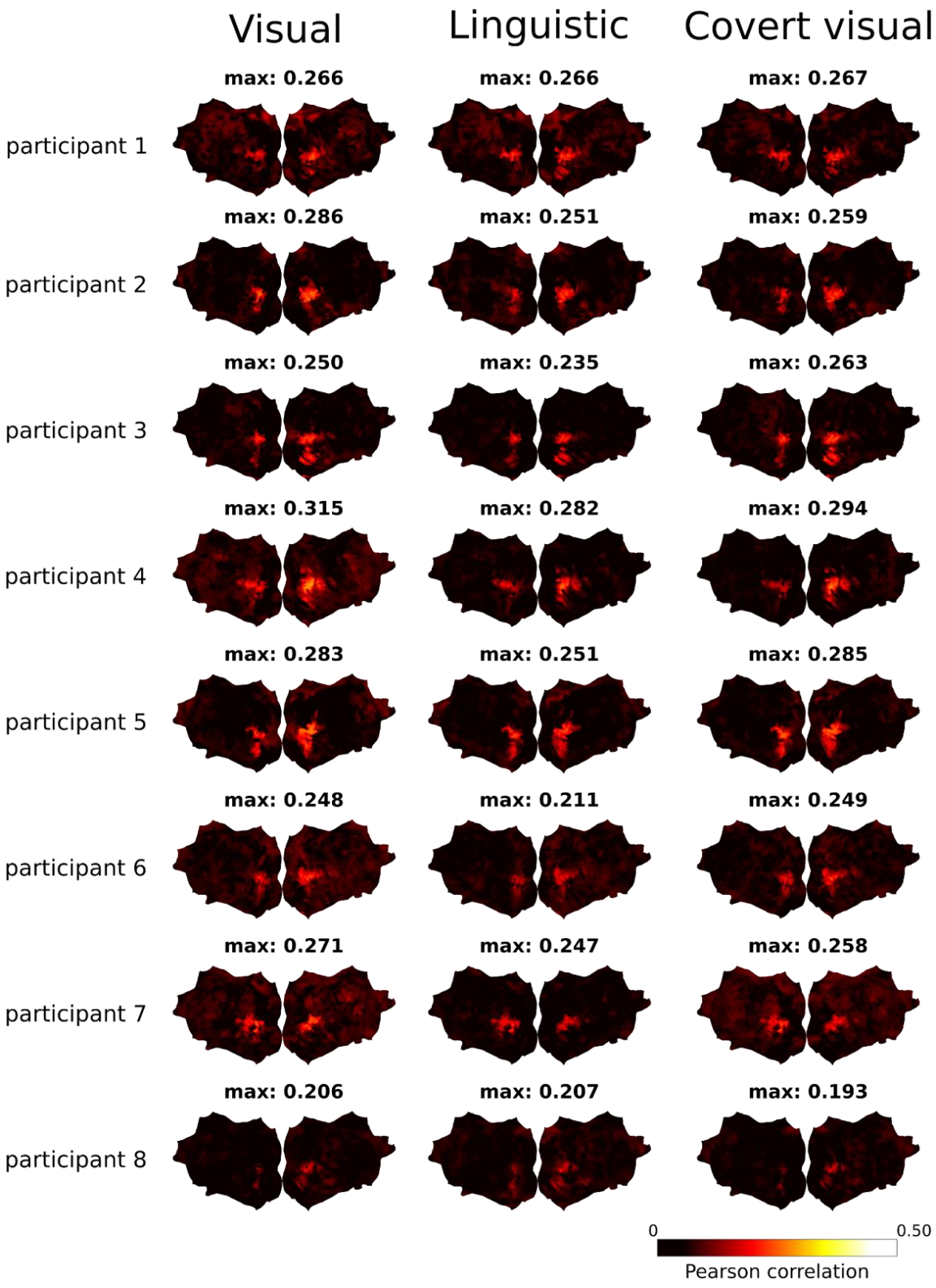
observed brain RDMs averaged across folds for each NSD participant. Representational alignment is obvious across all three modalities despite only behavioural RDMs from session 1 being considered here.

Group-level statistics



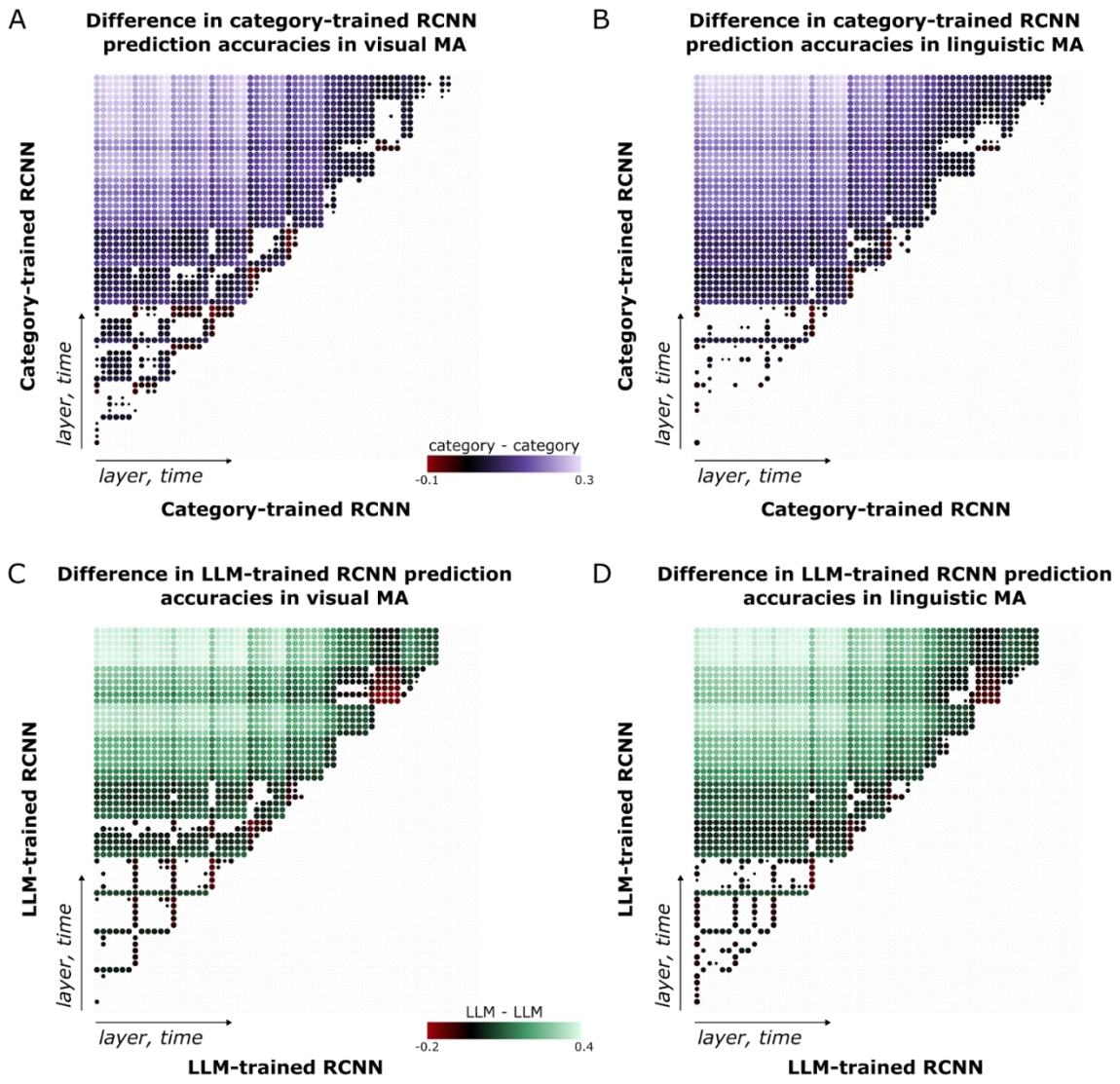
Supplementary Figure 2.5 | Representational alignment between the brain responses and multimodal similarity judgements from session 1 using Pearson

correlations. To demonstrate that the representational alignment is reproducible even without creating a best-fitting behaviour-predicted RDM, we correlated the averaged modality-specific behavioural RDMs from session 1 with searchlight RDMs at every searchlight for each NSD participant. The surface maps show averaged Pearson correlations across the NSD participants ($p < 0.05$, FDR-corrected, one-sided test). The correlations peaked at 0.239, 0.204, and 0.220 for visual, linguistic, and covert visual modality respectively. The spatial distributions of the correlations on the cortex are similar, albeit more isolated, to the ones reported in Figure 2.2B. This finding shows that the similar spatial distribution on the cortex observed with the linguistic and visual behavioural RDMs cannot simply be explained by our re-weighting of the individual subjects in the cross-validated non-negative least squares analysis. This is even more significant considering that we used behavioural RDMs from session 1 where no cross-modal projection played a role.



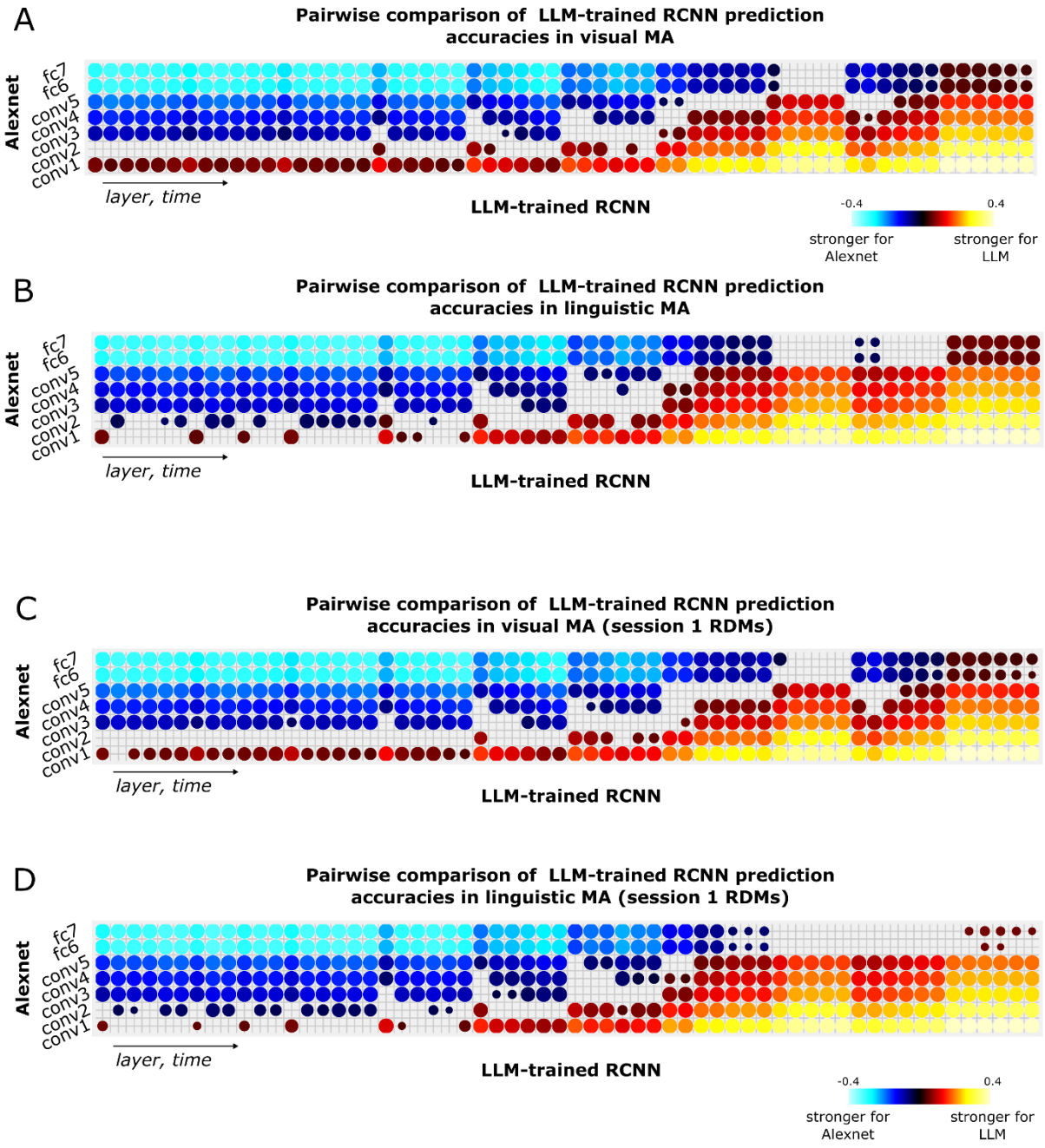
Supplementary Figure 2.6 | Representational alignment between the brain responses and multimodal similarity judgements from session 1 participant-level using Pearson correlation. The figure shows representational alignment at the participant-level where we averaged Pearson correlations across folds in each NSD

participant. Pearson correlations were performed between the averaged modality-specific behavioural RDMs and brain RDMs at every searchlight and only behavioural RDMs from session 1 were used.



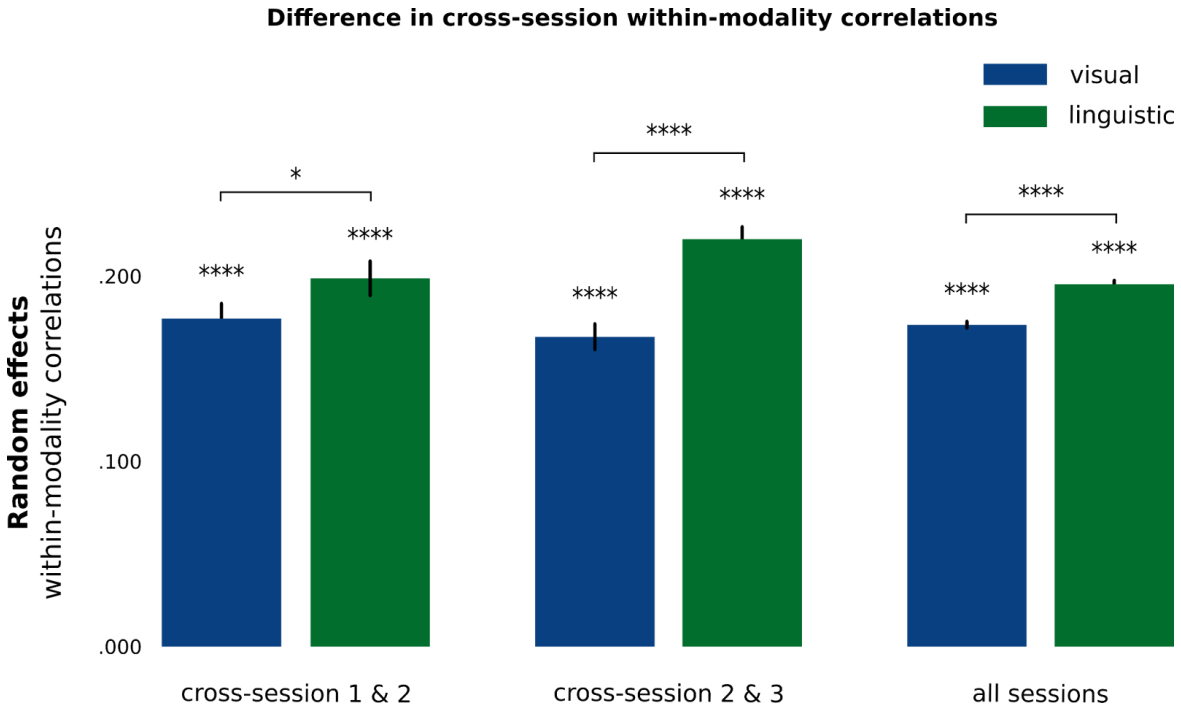
Supplementary Figure 2.7 | Pairwise difference in RCNN prediction accuracies. Differences in prediction accuracies were obtained by subtracting cosine similarities for all pairwise combinations of layers and time steps within an RCNN. The colour gradient depicts the size of the difference. The significance of prediction accuracies across seeds within each layer was determined using a one-sided t-test against zero ($p < 0.05$ FDR corrected). Only significant cosine similarities are depicted. The radius of the points marks the size of p-values. Comparison between category-computable RCNN RDMs and behaviour-predicted RDMs in vision and language is depicted in panels (A) and (B)

respectively. Comparisons between sentence-computable RCNN RDMs and behaviour-predicted RDMs in vision and language are depicted in panels (C) and (D). The layer and time complexity increases from the origin.



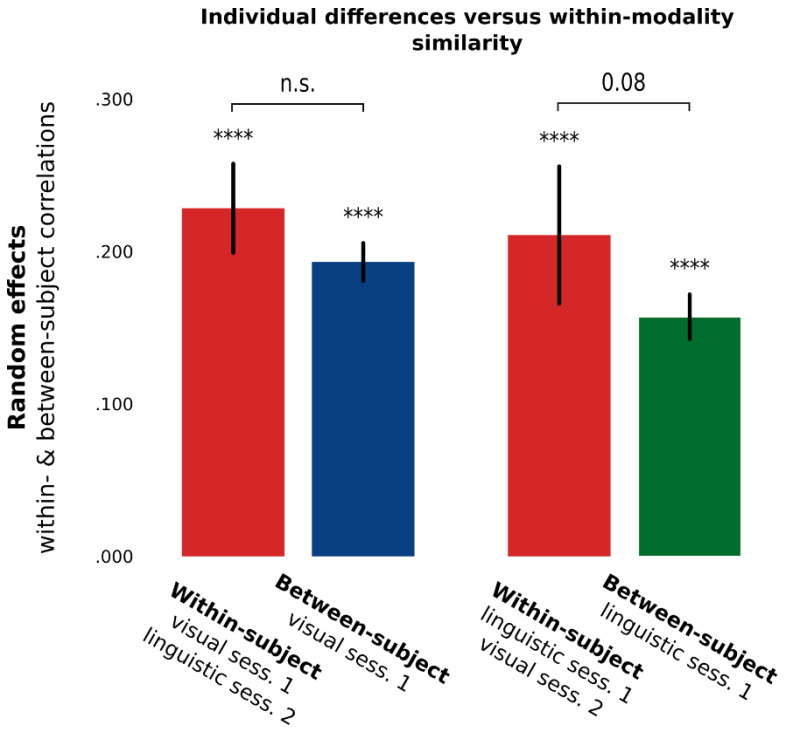
Supplementary Figure 2.8 | Pairwise difference in RCNN and Alexnet prediction accuracies. The image set was passed through the first five convolutional and two fully-connected layers of Alexnet. The behaviour-predicted RDMs were then evaluated on each Alexnet layer RDM to determine their alignment. The prediction accuracies were then

compared to prediction accuracies in the sentence-computable RCNN in visual (A) and linguistic (B) modalities across all sessions. To control for potential task order effects, we performed the same analysis with behavioural RDMs from the participants' first session in visual (C) and linguistic (D) similarity judgements. A two-sided t-test was performed on cosine similarities across seeds (RCNN) and folds (RCNN, Alexnet) to determine whether there is a significant difference in prediction accuracies. The radius of the points marks the size of the p-values (p < 0.05 FDR corrected). Only significant differences in cosine similarities are depicted. The colour gradient depicts the size of the difference where mean prediction accuracies from Alexnet were subtracted from mean prediction accuracies from sentence-computable RCNN. The layer and time complexity of the RCNN increases from the origin of the x-axis.

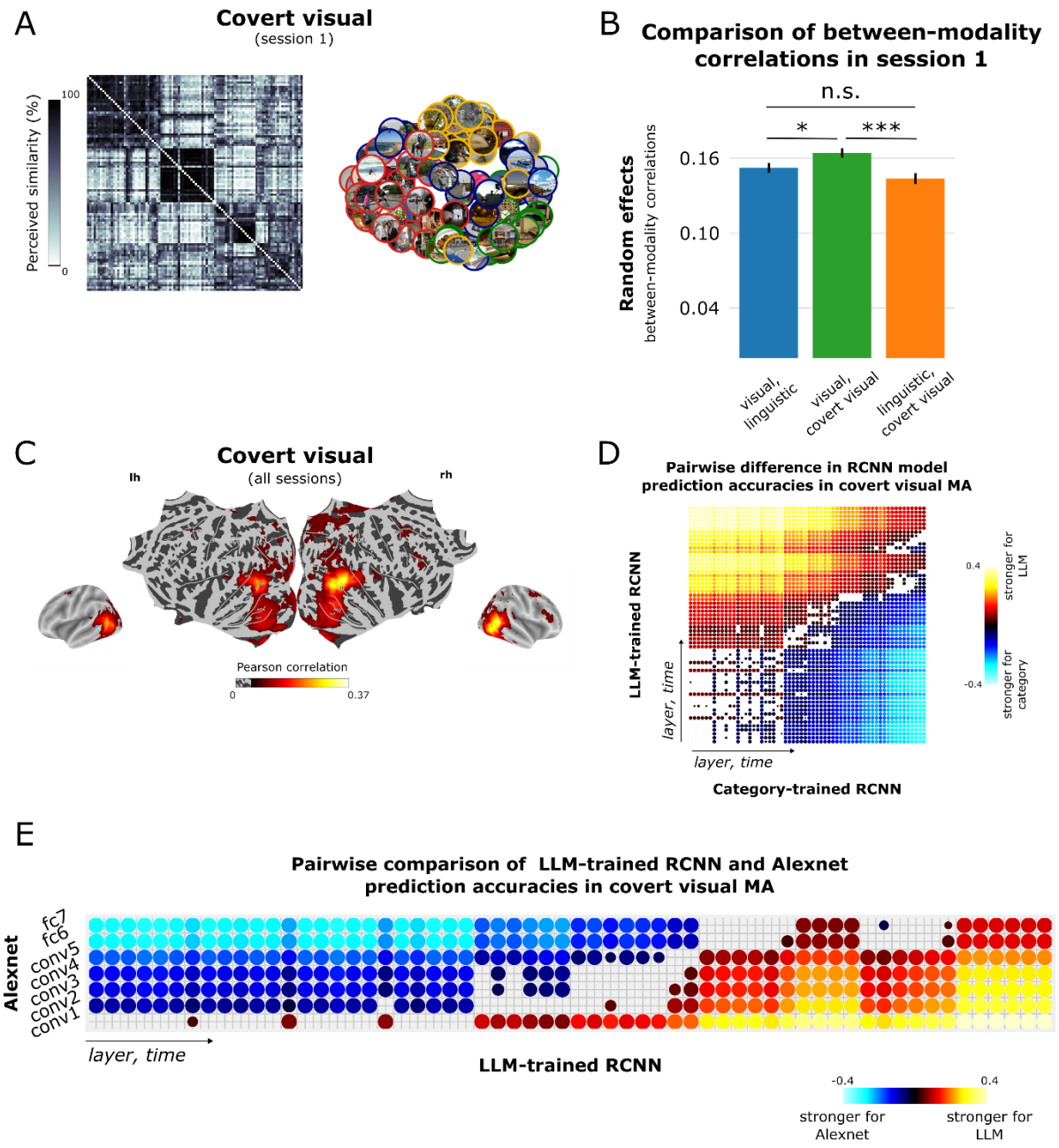


Supplementary Figure 2.9 | Cross-session within-modality comparisons. To investigate the stability of the linguistic and visual representational geometry throughout the experiment, we compared the within-modality correlations across sessions. Based on the assumption that prior exposure to vision narrows down the linguistic representational geometry, we expected to observe a higher stability of the linguistic geometry across sessions and participants compared to the visual representational geometry. First, we correlated modality-specific RDMs from session 1 with modality-

specific RDMs from session 2. We found an average within-modality correlation of 0.199 between the linguistic RDMs (one-sided t-test: $t(22) = 22.272$, $p < 0.0001$) and an average within-modality correlation of 0.177 between the visual RDMs (one-sided t-test: $t(22) = 23.022$, $p < 0.0001$). As expected, the one-tailed independent samples t-test showed that the within-modality cross-session correlations of the linguistic RDMs are significantly greater than the within-modality correlations of the visual RDMs ($t(44) = 1.837$, $p = 0.034$). A similar result was observed when we included within-modality correlations across session 2 (only RDMs that were preceded by linguistic or visual MA were included) and session 3 (all visual and linguistic RDMs regardless of the preceding modality; mean Spearman's ρ linguistic RDMs: 0.208; mean Spearman's ρ visual RDMs: 0.169; independent samples t-test: $t(41) = 5.774$, $p < 0.0001$) and all RDMs altogether (mean Spearman's ρ linguistic RDMs: 0.196; mean Spearman's ρ visual RDMs: 0.174; independent samples t-test: $t(124) = 9.993$, $p < 0.0001$). This observation is likely a direct result of the differences in the cross-modal projection where prior exposure to vision narrows down the representational geometry in language, but prior exposure to language expands the representational geometry in vision. The stars denote the p-values with: *$p < .05$, ** $p < .01$, *** $p < .001$, **** $p < .0001$.



Supplementary Figure 2.10 | Interindividual variability tends to display the opposite effect when controlled for prior exposure. We correlated a participant's RDM from session 1 with other participants' RDMs from the same session and modality to investigate whether individual differences outweigh the within-modality similarity before being exposed to the other modality. Considering that visual geometry in session 1 is better aligned between participants (Figure 2.4D), we expected the within-modality similarity to be comparable to within-participant between-modality similarity, which is what two-tailed paired samples t-test revealed (red and blue bars on the left; $t(11) = 1.314$, $p = 0.108$). We repeated the same procedure for the linguistic modality. Here, we expected to observe the opposite effect because, as shown in Figure 2.4D, (1) linguistic geometry is more likely to capture individual differences before cross-modal projection, and (2) visual geometry tends to employ these idiosyncrasies after exposure to language. Paired samples t-test revealed a trend to significance suggesting that inter-individual variability to some extent does outweigh within-modality similarity in session 1 (red and green bars on the right; $t(10) = 1.463$, $p = 0.087$). The stars denote the p-values with: *p < .05, ** p < .01, *** p < .001, **** p < .0001.



Supplementary Figure 2.11 | Covert visual modality compared to behaviour, brain, and models. We performed additional analyses to determine the nature of the representational geometry in the control MA task that presented the covert visual modality. (A) Visual inspection of the RDM from the session 1 covert MA task reveals similar high-level category clusters as we observed in the visual and linguistic RDMs (Figure 2.4). However, at the centre of the RDM, we can see a dark cluster, which has a detail of a drop in the perceived similarity - similar detail can be observed in the visual RDM but not in the linguistic RDM. The images in the MDS are colour-coded to differentiate between the high-level categories. Closer inspection of the figure reveals a

higher similarity of the covert visual MDS to the linguistic MDS with the animate-human category (red) clustered at the lower left area, animate-animals category (yellow) clustered at the top centre, inanimate-indoor category at the lower right (green), and inanimate-outdoor category (blue) scattered at the inner borders of the clusters of the remaining three categories. (B) Session 1 RDMs were correlated using Spearman's correlation to investigate the representational overlap across the three independent groups of participants. Independent samples t-test revealed that the visual and covert visual modalities were the most strongly correlated geometries. No difference was found in whether the linguistic modality was correlated with the visual or covert visual modality. These findings show that even though the covert visual modality is generally more similar to the visual representational space, it also connects to the linguistic space in a similar manner as the visual modality. It is likely that the covert characteristic introduced in the visual modality had no effect on the way participants configured the space. We further show that the representational geometry of the covert visual modality relates to the visually evoked activity similarly (C). Finally, the representational geometry was found to fit significantly better the representational geometry of sentence embeddings from LLMs compared to structure from category-trained RCNN (D), or AlexNet (E). The stars denote p-values with: *$p < .05$, ** $p < .01$, *** $p < .001$, **** $p < .0001$.

Chapter 3

The semantic distance between a linguistic prime and a natural scene target predicts reaction times in a cross-modal priming experiment

3.1 Abstract

Research into early representation learning indicates that language shapes how we represent the visual world long before we say our first words. A recently raised possibility is that vision and language align in their ability to transform their inputs into a shared similarity space. However, how regularities of such a space support human behaviour remains unknown. 95 participants completed a novel cross-modal priming experiment where a display of a visual target and five distractors was preceded by a linguistic prime. We used representational geometries from behavioural similarity judgements to manipulate the dissimilarity between the prime and the target while keeping dissimilarities between the target and the distractor constant. A linear mixed effects model (LME) revealed a clear cross-modal priming effect where response times (RTs) were faster the closer the linguistic prime and visual target were represented in the similarity space. Strikingly, the strongest effect emerged when RTs were predicted by each participant's (N = 27) own similarity judgements of linguistic primes, but not visual targets. These findings are significant in at least two major respects: firstly, language and vision project their representations onto a shared representational space whose internal geometry is characterised by perceived dissimilarities. Secondly, this space aligns with interindividual differences in linguistic semantics. Finally, to control for the distractors' role in the cross-modal priming effect, we also manipulated the distractors' dissimilarity, in addition to primes' dissimilarity, from the target (N = 51). We found that the cross-modal representation mapping of the representational space is characterised by many of the same properties that were previously identified in exclusively visual experiments. Together, our findings show that vision and language converge onto a shared representational space characterised by perceived dissimilarities opening a discussion as to what extent intricacies of language shape the internal structure of this space.

3.2 Introduction

Research shows that children learn about the visual world through language long before they say their first words. For example, if a spoken word is consistently associated with a set of visual objects, it promotes the commonalities between them, allowing children to represent the objects as members of one category (Althaus, & Mareschal, 2014; Ferry et al., 2013; LaTourrette, & Waxman, 2020; LaTourrette et al., 2023; Waxman, & Markow, 1995). As a result, visual perception is guided by a semantically precise link between a spoken word and its mental representation as early as 7 months of age (Bergelson, & Swingley, 2012; LaTourrette et al., 2023; Tincoff, & Jusczyk, 1999). Behavioural studies even show that language has the capacity to prime recognition of semantically related visual objects by activating a prior representation in a common semantic network (Harder, & Tylén, 2019; Lupyan, & Spivey, 2010; Lupyan, & Swingley, 2012; Meteyard et al., 2007; Meyer et al., 2007; Sperber et al., 1979; Stanfield, & Zwaan, 2001), further demonstrating the interconnected nature of language and vision. This evidence indicates that the visuo-linguistic alignment is critical in early representation learning enabling us to understand and navigate the world around us. What remains less clear, however, is the precise framework that could be used to describe the visuo-linguistic alignment and how its regularities predict human behaviour.

Recent advances in computational methods have enabled researchers to investigate the visuo-linguistic alignment with increasingly complex models. For example, brain responses to viewing natural scene images align with sentence embeddings generated by large language models (LLMs; Doerig et al., 2024). Additionally, several studies show that models that leverage the strengths of both deep neural networks for visual processing and LLMs provide the best means of predicting the visual brain (Wang et al., 2023). What these models have in common

is that their internal representations encode high-dimensional patterns and similarities across visual and textual contexts (Bruni et al., 2011; Li et al., 2023; Lu et al., 2019; Sadeghi et al., 2015). Such visuo-linguistic alignment has its relevance even in behaviour: similarity judgements of sentence captions were found to produce representational geometries that predict many of the same visual regions as identified by LLMs (Chapter 2). It is therefore possible that vision aligns with language in its inherent ability to translate sensory inputs into a unified multidimensional space of perceived similarities.

In the current study, we developed a novel cross-modal priming experiment to investigate whether the cross-modal representation access is predictable by the regularities of this multidimensional similarity space. Prior research provided insights into how visual representational dissimilarities support visual behaviour. It was found that the degree to which our brain perceives something as dissimilar is predictive of performance in behavioural tasks such as visual search (Cohen et al., 2014; Cohen et al., 2015; Cohen et al., 2017; Duncan, & Humphreys, 1989; Moores et al., 2003; Sripati, & Olson, 2010; Yeh et al., 2024), masking (Cohen et al., 2015), and categorisation behaviour (Carlson et al., 2014; Ritchie et al., 2015). For instance, if two visual objects are perceived as more similar, they also compete for a similar neural representation, which then hinders the ability to process the objects in parallel, interfering with visual search efficiency in return (Cohen et al., 2015). This effect decreases with increasing discriminability of the neural representation (Carlson et al., 2014; Ritchie et al., 2015). Providing that language aligns with vision in the common multidimensional space, it was hypothesised that a linguistic prime would induce preparatory activity in the shared representational space (Dils, & Boroditsky, 2010; Gayet, & Peelen, 2022) making the visual target faster to search for. This relationship would then be predictable by regularities of the multidimensional similarity space. Finally, considering evidence from early representation learning

showing that cross-linguistic conventions (Athanasopoulos et al., 2015; Daoutis et al., 2006; Roberson et al., 2000; Roberson et al., 2005; Thierry et al., 2009; Winawer et al., 2007) and individual differences in perceived similarity (Charest et al., 2014) shape the internal representational geometry, we further predicted the cross-modal priming effect using idiosyncratic representational dissimilarities obtained from each participants' similarity judgements data. It was hypothesised that if the multidimensional space is sensitive to interindividual differences in semantics, idiosyncratic representational dissimilarities would provide a better means of predicting the cross-modal priming effect. Characterising how regularities of the multidimensional space relate to human behaviour is instrumental in advancing our understanding of how vision and language are combined to form a complete representation.

3.3 Material and methods

3.3.1 Participants

95 participants (mean 20 ± 3 years; 84 right-handed, one mixed) participated in Experiment 1. The sample size was based on participants' availability in the month during which the experiment was advertised online. A priori power analysis was conducted on data from Experiment 1 to determine the minimum sample size required in Experiment 2. Results indicated a sample size of at least 37 participants to achieve the observed power of 0.75 from Experiment 1 at a significance criterion of $\alpha < 0.001$. A new group of 51 participants (mean 19 ± 1; 46 right-handed) was recruited for Experiment 2 allowing for a dropout rate. Participants in both experiments were recruited via the University of Birmingham SONA recruitment system. Eligibility criteria required participants to have normal or corrected-to-normal vision and have access to a desktop/laptop with at least a 15-inch screen and a keyboard. Participants provided informed consent as part of

a protocol approved by the University of Birmingham Research Ethics Committee and were rewarded course credits for participation.

3.3.2 Materials

The stimulus set was obtained from the Natural Scenes Dataset (NSD, Allen et al., 2022) and consisted of 100 natural scene images and 100 sentence captions describing the images. A greedy algorithm (Allen et al., 2022) selected the stimuli such that they span the full semantic space as defined by embeddings of the images' sentence captions. The sentence captions were collected from five human annotators as part of the Microsoft Common Objects in Context database (MS COCO, Lin et al., 2014). For the current experiment, the most frequent words and phrases across the five annotators were identified and combined to generate one representative sentence caption for every image. The stimulus set was also accompanied by similarity judgements data from the NSD participants (Allen et al., 2022) who completed the multiple arrangements task (MA, Kriegeskorte & Mur, 2012) on the set of 100 images. We selected similarity judgements data from NSD subject 1 (Figure 3.1A) and measured the Euclidean distance between every pair of images to denote their representational dissimilarity. The pairwise dissimilarities were assembled in a representational dissimilarity matrix (RDM), which was then used to sample trial triplets (Figure 3.1B) consisting of a target, a prime, and a distractor while controlling for their dissimilarity within the NSD representational space.

3.3.3 Experiment 1: Cross-modal priming effect

Experiment 1 was designed to investigate whether the NSD representational dissimilarity between the linguistic prime and the visual target predicts response time to the target that was embedded among a set of related visual distractors.

We sampled the trial triplets (Figure 3.1B) consisting of a prime, target, and distractor from the NSD representational space where every image from the stimulus set became a target associated with a prime under three dissimilarity conditions: (1) the prime is the same stimulus as the target where Euclidean distance in the NSD representational space is 0 (identity dissimilarity condition), (2) the prime is halfway between the target and the furthest item from the target where the Euclidean distance is the closest distance to ½ of the maximum distance (medium dissimilarity condition), or (3) the prime is the furthest stimulus from the target where the Euclidean distance is the maximum distance from the target (high dissimilarity condition). The distractor was always sampled as the closest stimulus to the target to induce the priming effect and increase the distractor-target interference (Cohen et al., 2014; Cohen et al., 2015; Moores et al., 2003). A total of 300 triplets were sampled in Experiment 1.

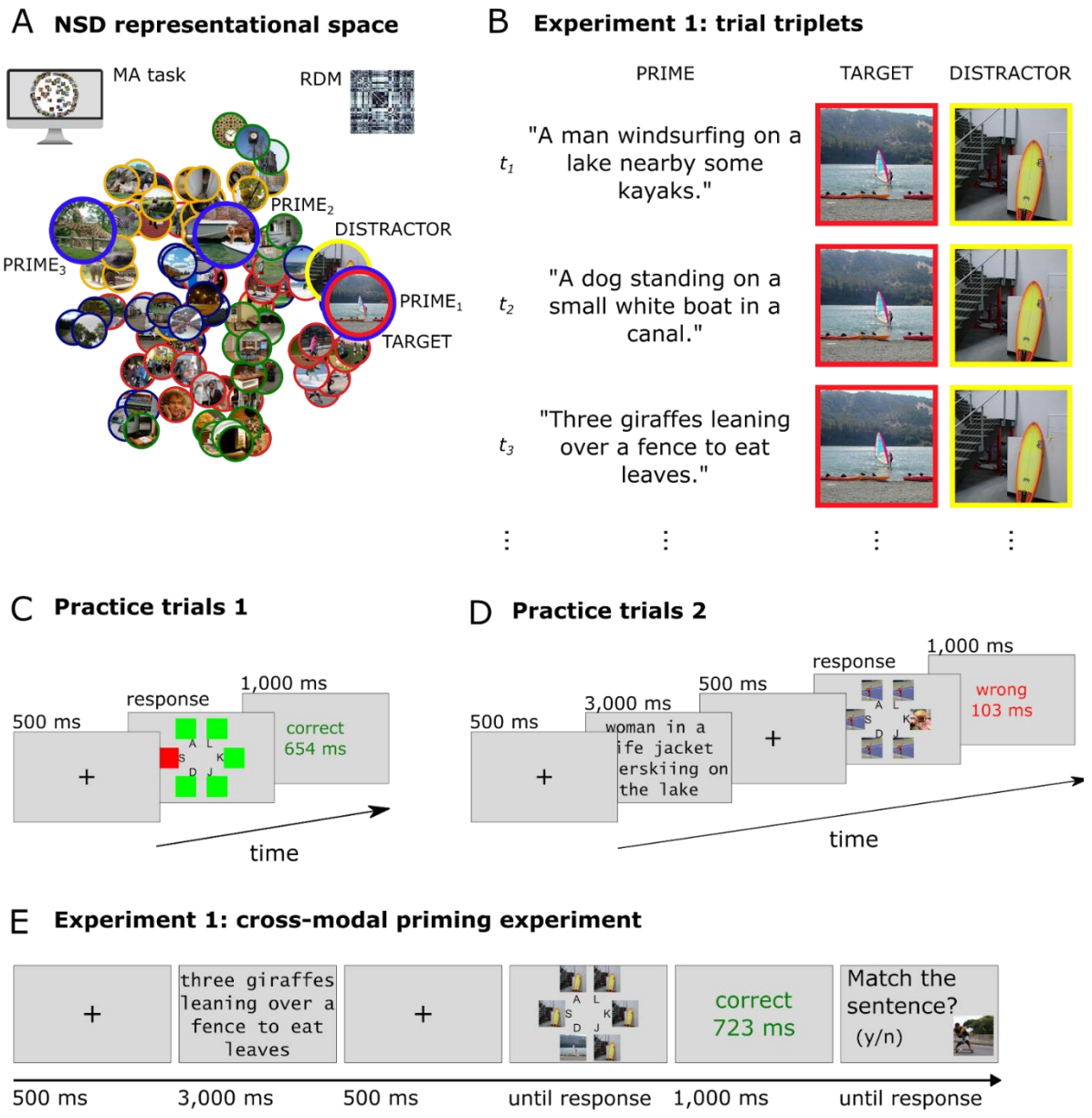


Figure 3.1 | Study design in Experiment 1. (A) We randomly selected one NSD participant who completed the multiple arrangements task on the set of 100 images. Multidimensional scaling was used to visualise the representational space of the stimulus set; the distance between the items represents their dissimilarity and the colour frame represents the category (red - human, yellow - animal, green - indoor, blue - outdoor). RDM from the reference NSD participant was used to sample the trial triplets (B) such that every image became a target (bright red frame) associated with a distractor (bright yellow frame) and a prime (bright blue frame) under three conditions: prime was the identity of the target, prime was the medium dissimilarity between the target and the

furthest item from the target, and prime was the furthest item from the target. The experiment started with a set of practice trials that intended to familiarise participants with the oddball visual search task (C) and with the priming component of the adapted version of the oddball visual search task (D). Trial triplets sampled from the NSD representational space were then embedded in the cross-modal priming experiment (E) which was the adapted version of an oddball visual search task.

3.3.4 Experiment 2: distractors in cross-modal priming experiment

To investigate how the distractors' dissimilarity from the prime moderates the cross-modal priming effect, we manipulated the distractor-target dissimilarity in addition to the prime-target dissimilarity in Experiment 2 (Figure 3.2A). The triplets consisting of a prime, target, and distractor were sampled from the NSD representational space relative to the global average dissimilarity (average Euclidean distance within the NSD RDM of ~0.01305) where every image from the stimulus set became a target associated with a prime and a distractor under four dissimilarity conditions (Figure 3.2B): (1) the prime and the distractor are both close to the target such that their distance from the target is below the global average dissimilarity, (2) the prime is sampled close to the target but the distractor is sampled far from the target such that the distractor's dissimilarity from the target is above the global average dissimilarity, (3) the prime and the distractor are both sampled far from the target such that their distance from the target is above the global average dissimilarity, or (4) the prime is sampled far from the target but the distractor is sampled close to the target such that the distractor's distance from the target is below the global average dissimilarity. A sampling restriction was introduced where the stimulus sampled either as a prime or a distractor was associated with a target only once in any condition. This allowed us to never sample an identical pair of a prime and a distractor in one trial. We also introduced a global restriction where any NSD stimulus was allowed to be

sampled only twice as a prime and as a distractor. This sampling procedure enabled us to sample as many NSD stimuli as possible. A total of 400 triplets were sampled in Experiment 2.

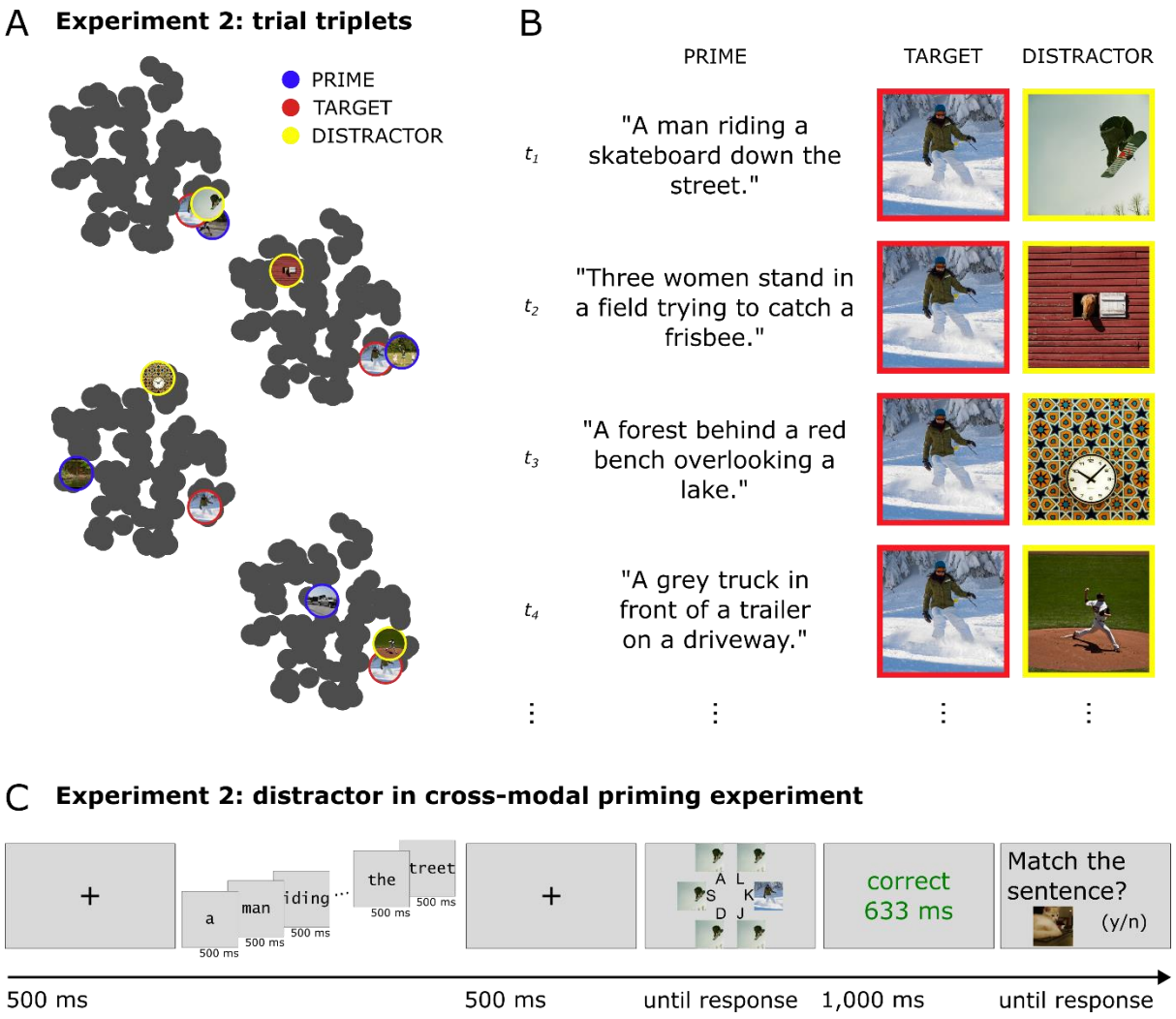


Figure 3.2 | Study design in Experiment 2. (A) The NSD representational space from participant 1 was used to sample trial triplets for Experiment 2 where every image from the NSD set became a target associated with a prime and a distractor under four dissimilarity conditions (B) while we manipulated both the prime's and distractor's dissimilarity from the target. This resulted in 400 trial triplets that were embedded in the cross-modal priming experiment (C).

3.3.5 Procedure

Both experiments were administered online via the web-based platform Meadows (https://meadows-research.com/). Participants were sent an invitation link alongside a few reminders instructing them to only use a laptop or a computer with at least a 15" screen, to maximise their browser window, and to carry out the experiment without any interruptions. Before agreeing to take part, participants were asked to read through the information sheet and to fill in a consent form and a demographics questionnaire. Prior to the main experiment, participants completed two sets of practice trials. The first set of 60 practice trials was intended to familiarise participants with the oddball visual search task (Figure 3.1C). The layout was identical to the main experiment described below, but the primes were omitted and the images were replaced with red and green squares. The colour of the target square and 5 distractor squares randomly alternated. The second set of practice trials consisted of 30 trials and was intended to reinforce fast and accurate responses and attention to the primes (Figure 3.1D). To avoid prior familiarity with the stimuli set in the main experiment, we randomly sampled 20 images and their sentence captions from the MS COCO database (Lin et al., 2014) and created 30 random triplets consisting of a target, a prime, and a distractor. The MS COCO images were cropped and resized to match the NSD stimuli. Each Experiment 1 and Experiment 2 sampled a new set of images and sentence captions for the second set of practice trials.

Triplets consisting of a prime, a target, and a distractor were incorporated in the cross-modal priming experiment, which was an adapted version of the oddball visual search task with an added linguistic prime. The prime was always presented as the sentence caption of the stimulus sampled from the NSD representational space to induce the linguistic representation. Participants were instructed to attentively read the sentences as they appeared on the screen and then to detect an oddball image by pressing the appropriate response key on the keyboard as fast as possible but without sacrificing accuracy. They were also

informed that a decent response time is under 900 ms. Every trial began with a 500 ms fixation followed by a prime displayed against a grey background, in lowercase letters, Lucida Console font and a font size of 2.5% of the screen's available width centred on the screen. Experiment 1 displayed the linguistic primes in a static display for 3 seconds (Figure 3.1E). Experiment 2 presented the linguistic primes in a rapid serial visual presentation format (RSVP); words from the prime were presented one at a time at the speed of 500 ms per word (Figure 3.2C). The benefit of this approach is that it guides the reading task more efficiently. The RSVP approach was also selected for its ability to activate the semantic representation as soon as 250 ms into the reading task (Kutas, & Hillyard, 1980; van Berkum et al., 1999). Next, another fixation of 500 ms appeared and was followed by the visual search display consisting of a target image and one distractor image shown five times. The images were displayed in an imaginary circle with the first image starting at 60° followed by the remaining five images at every 60° angle anti-clockwise. The images were squared 225 x 225 pixels each scaled at 9% of the screen's available width. The images were labelled starting with L for the first item followed by A, S, D, J, and K respectively corresponding to the images' response keys on the keyboard. The visual search display disappeared once a key press was recorded. Participants immediately received feedback on their accuracy and RT. Every trial had a 5% chance of being followed by a catch trial that examined participants' attention to the primes. The catch trial showed an image with a prompt asking whether the image depicts the content of the sentence from the beginning of the trial. The y/n keyboard keys were used for a yes/no response. A break prompt appeared every 60 trials instructing participants to take a short break. Experiment 1 presented 3 repetitions of the triplets yielding 900 trials in total. Experiment 2 presented 2 repetitions of the triplets yielding 800 trials in total. Each experiment was completed by a different participant group and each took approximately 120 minutes to complete.

3.3.6 Estimating the visual and linguistic representational space

A subset of participants from Experiment 1 (*N* = 27) provided similarity judgements data by completing the multiple arrangements task (MA, Chapter 2) on the set of 100 natural scene images and 100 sentence captions describing the images that were used in the experiment. We used the pairwise Euclidean distances from the similarity judgements data to construct two RDMs per participant: one that represented the idiosyncratic representational space of visual similarity judgements and one that represented the space of linguistic similarity judgements. The idiosyncratic RDMs were then used to extract and match the pairwise dissimilarities with corresponding prime-target pairs from the triplets sampled in Experiment 1. This allowed us to predict participants' response times in every trial by their own perceived prime-target dissimilarity either in their visual or linguistic format. To make more general predictions about the visual and linguistic representational space, we also separately averaged the visual and linguistic RDMs across all participants and matched the pairwise dissimilarities from the RDMs with corresponding prime-target pairs in the subset of participants. We then predicted the participants' response times using the averaged representational dissimilarities from each visual and linguistic representational geometry to investigate how well the visual and linguistic similarity structure predicts the semantic priming effect.

3.3.7 Statistical analysis

The data preprocessing was done in Python 3.9.7. A binomial test was performed to identify participants whose accuracy on catch trials was below chance indicating low attention to the linguistic primes. Eleven participants from Experiment 1 and two participants from Experiment 2 were excluded based on the binomial test. Trials with incorrectly identified target images and invalid

response times (i.e., negative response times due to a technical issue) were excluded. Response times were log-transformed to achieve an approximate normal distribution. Log-transformed response times were cleaned from outliers using Tukey's method where any response time smaller than the first quartile minus 1.5 times the interquartile range (IQR) or larger than the third quartile plus 1.5 the IQR was considered an outlier. Trials where the prime-target dissimilarity was zero were also excluded from the analysis. This condition was introduced to induce the priming effect but led to 33% of trials having zero variance in their predictors, which may introduce a bias in the model estimates. The data cleaning process led to the exclusion of 29471, and 1240 data points amounting to 38.7%, and 8% of the available data in Experiment 1, and 2 respectively (see Supplementary Table 3.1 for a full breakdown).

Linear mixed effects model (LME) from lme4 library (Bates et al., 2015) within the R environment was used. Satterthwaite approximation was used to perform inference on LMEs that are reported in the results section. The likelihood ratio test was used to determine whether the increasingly complex fixed effects structure in the LME improves fit to the response times. Both were performed using the anova function from the R stats library (R version 4.3.2, R Core Team, 2023). Akaike Information Criterion (AIC) from the AICcmodavg library (Mazerolle, 2020) was used to (1) assess the goodness of fit of models with maximal random effects structure relative to simplified random effects structure (note that the fixed effect structure remained identical across these models), and (2) compare models with identical structure but different type of fixed effects (i.e., response times predicted either by visual or linguistic prime-target dissimilarity). In Experiment 1, the effect of interest was the interaction between log-transformed response times and prime-target dissimilarity. To enable the fixed effect to vary across participants, each LME was initialised with a maximal random effects structure (Barr et al., 2013) where both prime-target dissimilarity and difference

among participants were treated as random effects. The model specification was as follows:

Eq. 1:

$$log.RT \sim prime.dissimilarity + (prime.dissimilarity \mid participant)$$

In Experiment 2, the analysis was initialised with LME with a single fixed effect predictor for prime-target dissimilarity with maximal random effects structure. However, as suggested by Barr and colleagues (2013) likelihood of convergence issues increases with the complexity of the model. If the model with maximal random effects structure failed to converge, the structure was then simplified in a stepdown fashion. The best model fit of the first analysis in Experiment 2 and analysis in Experiment 1 for averaged visual and linguistic dissimilarities (Figure 3.4) was LME with a fixed slope (respectively AIC -3.01, -1.22, -1.5 relative to LME with maximal random effects structure):

Eq. 2:

$$log.RT \sim prime.dissimilarity + (1 \mid participant)$$

For Experiment 2, we then increased the complexity of the fixed effects structure in a stepwise fashion, which allowed us to examine how the initial cross-modal priming effect changes as we consider additional factors. First, we built up the LME in Eq. 2 by adding the interaction between prime-target and distractor-target dissimilarity with maximal random effects structure. The maximal random effects structure was simplified until the LME converged (AIC -4.52 relative to LME with maximal random effects structure):

Eq. 3:

$$log.RT \sim prime.dissimilarity + prime.dissimilarity : distractor.dissimilarity + (1 \mid participant)$$

We used the likelihood ratio test to determine whether the combination of a single fixed effect and the interaction between fixed effects (Eq. 3) improved the fit to response times data compared to the initial model (Eq. 2). The fixed effects structure was then increased by adding distractor-target dissimilarity as an additional fixed effect. The maximal random effect structure that converged was (AIC -8.37 relative to the maximal random effects structure):

Eq. 4:

$$log.RT \sim prime.dissimilarity + prime.dissimilarity : distractor.dissimilarity + distractor.dissimilarity + (prime.dissimilarity \mid participant)$$

The next part of analyses in Experiment 2 initialised the LME with a single fixed effect for distractor-target dissimilarity and maximal random effects structure for distractor-target dissimilarity and difference among participants. This was done to investigate the effect of distractor-target dissimilarity on response times in isolation from other factors. The maximal structure that converged yielded the following model specification:

Eq. 5:

$$log.rt \sim distractor.dissimilarity + (distractor.dissimilarity \mid participant)$$

The fixed effects structure was then built up by adding the interaction between distractor-target and prime-target dissimilarity as an additional fixed effect. This allowed us to investigate how the effect of distractor-target

dissimilarity on response times changes as we account for additional factors. The maximal model that converged was as follows (AIC -4.24):

Eq. 6:

$$log.RT \sim distractor.dissimilarity + distractor.dissimilarity : prime.dissimilarity + (1 | participant)$$

Finally, the maximal fixed effects structure was achieved by adding fixed effect for prime-target dissimilarity (AIC -11.28):

Eq. 7:

$$log.rt \sim distractor.dissimilarity + distractor.dissimilarity : prime.dissimilarity + prime.dissimilarity + (distractor.dissimilarity : participant)$$

3.4 Results

3.4.1 Behavioural results

The overall performance in the cross-modal priming experiment was good and consistent across both experiments with an average accuracy of 93% (SD = 3.84) in Experiment 1 and 93.1% (SD = 4.55) in Experiment 2. The average accuracy on the catch trials was 88.8% (SD = 9.05) in Experiment 1 and 84% (SD = 13.68) in Experiment 2 indicating that participants were generally attentive to the sentence primes. The untransformed mean response times were 924 (SD = 334) ms and 973 (SD = 219) ms in Experiment 1 and Experiment 2 respectively. The subset of participants from Experiment 1 that also completed the similarity judgements correctly identified the target image in 92.6% (SD = 3.58) trials, scored 91.1% (SD = 9.74) on the catch trials and responded in 875 (SD = 318) ms on average.

3.4.2 Experiment 1: Cross-modal representation access

First, we examine whether there is an effect of the referential prime-target dissimilarity sampled from the NSD representational geometry on response times. It was hypothesised that if the linguistic prime activates a representation in the shared representational space, the search for the visual target should be faster the more semantically related the prime and the target are. This should be predictable by prime-target dissimilarity providing that both vision and language draw on the shared representational space (Chapter 2). The random slope LME (Eq. 1) revealed a significant positive effect of the visual prime-target dissimilarity on the log-transformed response times (estimate = 0.752, SE = 0.191, $F(1, 3336.40) = 15.49$, $p < 0.001$, 95% CI = [0.38, 1.13], Figure 3.3). This observation indicates that lower dissimilarity between the linguistic prime and the visual target leads to faster identification of the target.

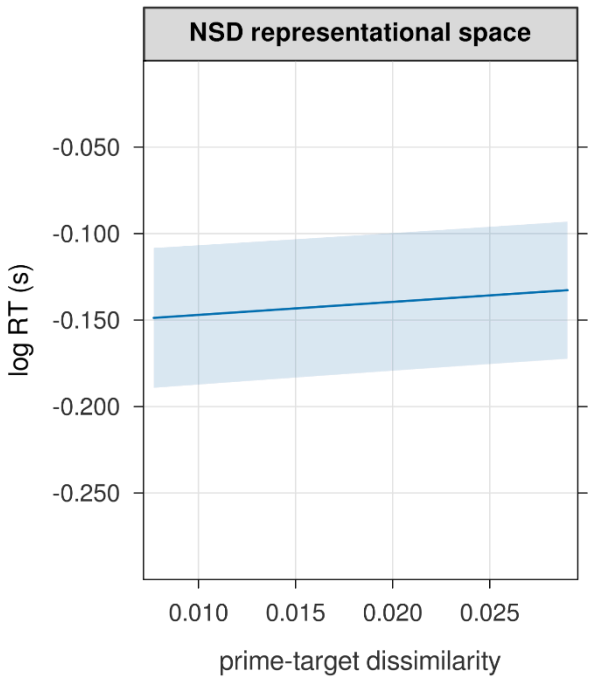


Figure 3.3 | Cross-modal priming effect predicted by NSD representational dissimilarities. Random slope LME was evaluated on log-transformed response times

from Experiment 1 with the prime-target dissimilarity from the NSD representational space as the fixed effect.

3.4.2.1 Linguistic dissimilarity outperforms visual dissimilarity in predicting the cross-modal priming effect

To investigate how well the visual and linguistic dissimilarity of the stimuli predicts the semantic priming effect, we matched the prime-target pairs and the corresponding response times with pairwise dissimilarities obtained from the visual and linguistic similarity judgements task. Again, providing that both modalities activate a shared representational space (Chapter 2), we expect to observe a positive priming effect regardless of whether we account for the visual or linguistic dissimilarity between the primes and targets. Turning first to the dataset with linguistic dissimilarity as the fixed effect, the fixed slope model (Eq. 2) revealed a significant positive relationship between the linguistic prime-target dissimilarity and the log-transformed response times (estimate = 3.297, SE = 1.006, F(,1, 14290.01) = 10.75, p = 0.001, 95% CI = [1.33, 5.27], Figure 3.4A). Next, we turn to the predictions made by visual dissimilarity. The fixed slope LME (Eq. 2) revealed a significant positive relationship between the visual prime-target dissimilarity and the log-transformed response times (estimate = 2.894, SE = 1.382, F(1, 14290.02) = 4.39, p = 0.036, 95% CI = [0.19, 5.60], Figure 3.4B). Although these results show that the semantic priming effect is predictable regardless of whether we account for visual or linguistic dissimilarity, the AIC model selection revealed that the better-fitting model included linguistic dissimilarity as the fixed effect carrying 96% of the cumulative model weight. This suggests that the semantic priming effect is better predicted by statistical regularities of the linguistic similarity space.

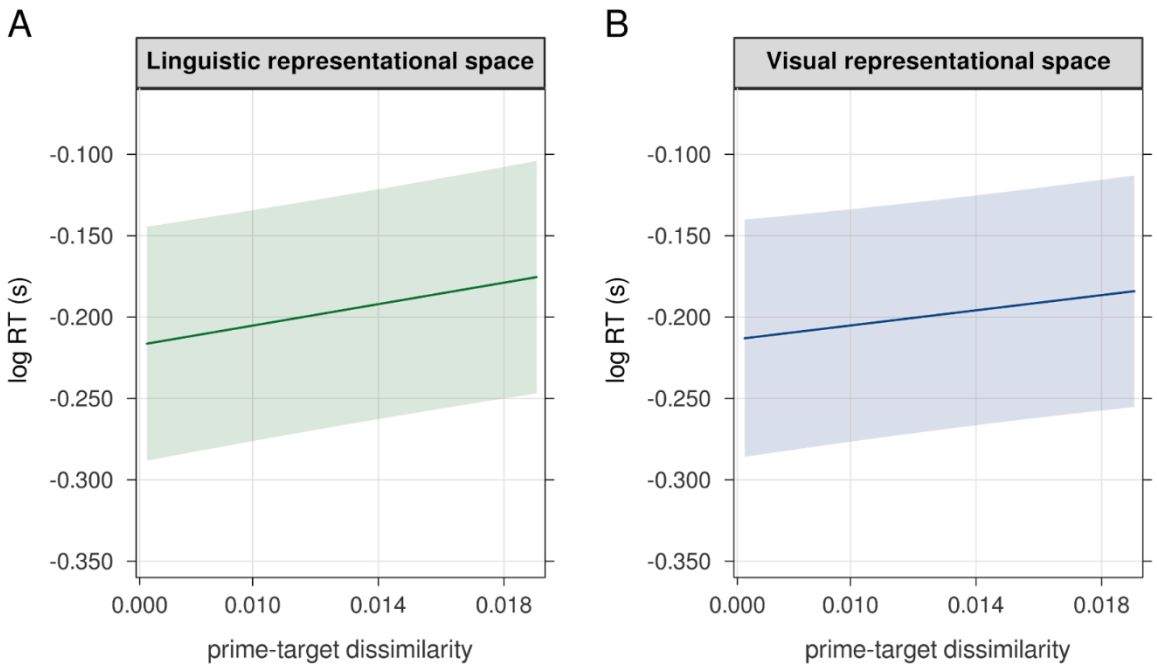


Figure 3.4 | Cross-modal priming effect predicted by averaged linguistic and visual dissimilarities. Fixed slope LME was evaluated on log-transformed response times with (A) linguistic prime-target dissimilarity from the linguistic representational space as the fixed effect, and (B) visual prime-target dissimilarity from the visual representational space as the fixed effect.

3.4.2.2 The idiosyncratic aspect of linguistic dissimilarity predicts the cross-modal priming effect

To explore how individual idiosyncrasies in perceived dissimilarity moderate the cross-modal priming effect, we leveraged the data from the similarity judgements task and constructed participant-specific linguistic and visual RDMs. We then matched the pairwise dissimilarities from each RDM with corresponding prime-target pairs, which enabled us to predict participants' (N = 26) response times by their own idiosyncratic dissimilarity of the stimuli either in their visual or linguistic format. First, we fitted a random slope LME (Eq. 1) on the log-transformed response times using the idiosyncratic linguistic dissimilarity as the fixed effect. This revealed a significant positive relationship between the

idiosyncratic linguistic prime-target dissimilarity and the response times (estimate = 1.240, SE = 0.495, $F(1, 29.75) = 6.29$, $p = 0.018$, 95% CI = [0.27, 2.21], Figure 3.5A). Next, we matched response times and corresponding prime-target pairs with visual pairwise dissimilarities (Eq. 1) from the idiosyncratic visual representational space. However, no evidence was found for any association between idiosyncratic visual dissimilarities and response times (estimate = 0.118, SE = 0.566, $F(1, 23.85) = 0.04$, $p = 0.837$, 95% CI = [-0.99, 1.23], Figure 5B). Taken together, these results indicate that a participant tends to respond faster on trials that present prime-target stimuli pairs that they perceive as more similar in the linguistic similarity judgements.

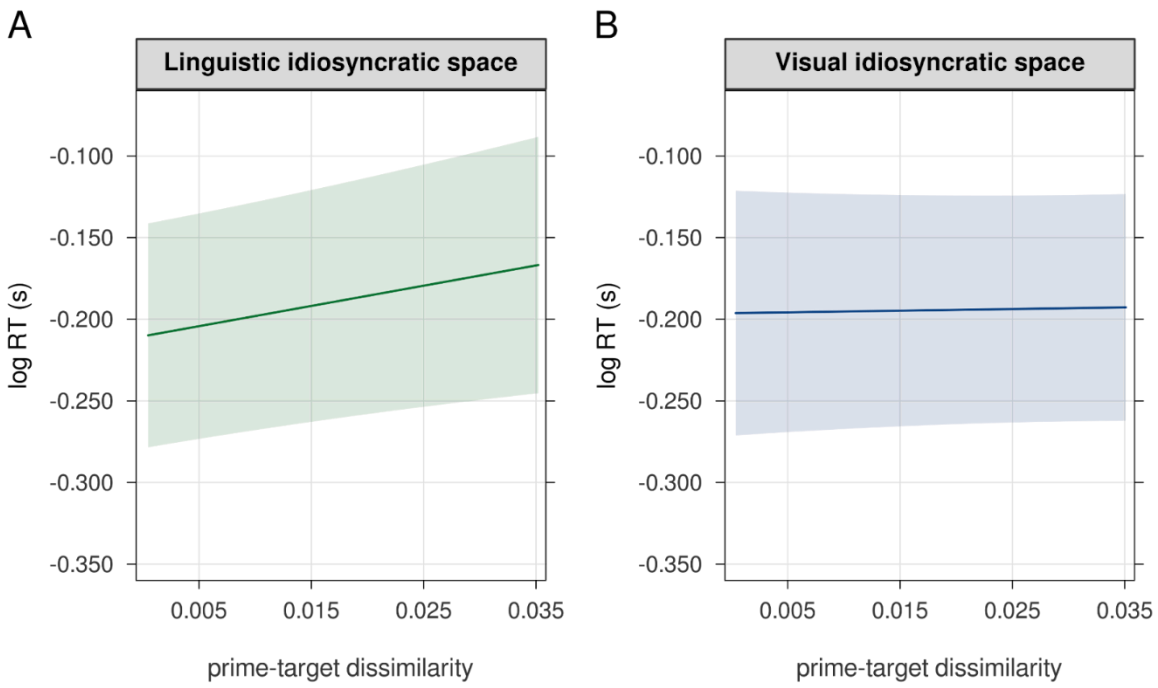


Figure 3.5 | Cross-modal priming effect predicted by idiosyncratic dissimilarities. Random slope LME was evaluated on log-transformed response times with (A) linguistic prime-target dissimilarity from the idiosyncratic representational space as the fixed effect, and (B) visual prime-target dissimilarity from the idiosyncratic representational space as the fixed effect.

Most importantly, the AIC model selection revealed that the model with idiosyncratic linguistic dissimilarities even outperforms the model with averaged linguistic dissimilarities (as introduced in the previous section) carrying 66% of the cumulative model weight. To rule out the possibility that this effect may be related to any intricacies of the current subset of participants, we permuted linguistic RDMs between participants (N = 1000) such that each participant's response times were predicted by linguistic dissimilarities from someone else's similarity space. We then fitted a random slope LME (Eq. 1) on each permuted sample and compared the permuted fixed effect estimates to the observed estimate from the model with idiosyncratic linguistic dissimilarities as predictors. It was hypothesised that if the better fit is driven by the idiosyncratic aspect of linguistic dissimilarities, then the permuted sample would yield a fixed effect estimate as high as the observed estimate in less than 5% of cases. The test revealed that the effect is indeed driven by the idiosyncratic aspect with q = 0.039 (Supplementary Figure 3.1). Taken together, we show that the cross-modal priming effect is best modelled as a function of perceived dissimilarities obtained from the idiosyncratic linguistic similarity judgements.

3.4.3 Experiment 2: Distractor in cross-modal representation access

So far we have demonstrated that linguistic primes promote visual targets the more semantically related they are, indicating that both modalities draw on representations within a shared representational space. It is worth noting that the priming effect remains significant even though the distractor was constantly sampled as the least dissimilar item from the target. Such observation suggests that linguistic primes facilitate visual search in ambiguous search conditions that were previously found to significantly interfere with search behaviour (Cohen et al., 2014; Cohen et al., 2015; Cohen et al., 2017; Moores et al., 2003; Sripati, & Olson, 2010). To explore how the target's dissimilarity from the prime and the

distractor modulates the cross-modal priming effect, we designed Experiment 2 where we manipulated the distractor-target dissimilarity in addition to prime-target dissimilarity. A new group of 51 participants completed Experiment 2.

3.4.3.1 Target's dissimilarity from prime and distractor mediates semantic priming

We fitted a fixed slope LME (Eq. 2) on the log-transformed response times with the prime-target dissimilarity as the fixed effect to investigate whether the priming effect remains even when the distractor-target dissimilarity varies. The model revealed a significant positive effect of the prime-target dissimilarity on response times (estimate = 0.580, SE = 0.272, $F(1, 33956.02) = 4.56$, $p = 0.033$, 95% CI = [0.05, 1.11]). We then added the interaction between the prime-target and distractor-target dissimilarity as an additional fixed effect (Eq. 3) to explore whether the strength of the priming effect depends on this interaction. This significantly improved the model fit ($\chi 2(1) = 86.31$, $p < 0.001$). The model revealed a significant effect of the interaction on response times (Table 3.1; estimate = -172.6, SE = 18.56, $F(1, 33955.01) = 86.42$, $p < 0.001$, 95% CI = [-208.96, -136.19]), and a significant positive effect of the prime-target dissimilarity on response times (estimate = 3.148, SE = 0.387, $F(1, 33955.01) = 66.12$, $p < 0.001$, 95% CI = [2.39, 3.91]). This observation indicates that the strength of the priming effect is dependent on the interaction between the target's dissimilarity from the prime and the distractor. Adding the distractor-target dissimilarity as an additional fixed effect (Eq. 4) did not improve the model fit ($\chi 2(1) = 2.348$, $p = 0.126$).

Table 3.1. The effect of prime-target dissimilarity and its interaction with distractor-target dissimilarity on log-transformed response times.

predictor	estimate	SE	F-value	*p*	95% CI

Prime-target diss.	3.148	0.387	66.12	< 0.001	[-208.96, -136.19]
Prime-target : distractor-target diss.	-172.6	18.56	86.42	< 0.001	[2.39, 3.91]

3.4.3.2 The distractor's dissimilarity from the target interferes with the visual search

Next, we investigate how the distractor moderates the cross-modal priming effect. We fitted a fixed slope LME (Eq. 5) on the log-transformed response times with the distractor-target dissimilarity as the fixed effect. The model revealed a significant negative relationship between the distractor-target dissimilarity and response times (estimate = -2.396, SE = 0.268, $F(1, 33956.01) = 80.24$, $p < 0.001$, 95% CI = [-2.92, -1.87]) indicating that lower distractor-target dissimilarity interferes with the search for the visual target. Adding the interaction between the prime-target and distractor-target dissimilarity as an additional fixed effect (Eq. 6) only marginally improved the model fit ($\chi 2(1) = 3.787$, $p = 0.052$) suggesting that the interference is somewhat independent of this interaction. However, adding both the interaction and the prime-target dissimilarity as additional fixed effects improved the fit. The model converged with a random effect for distractor-target dissimilarity (Eq. 7) and outperformed the other LMEs ($\chi 2(3) = 13.278$, $p = 0.004$). The results revealed a significant effect of interaction between the prime-target and distractor-target dissimilarity (Table 3.2; estimate = -102.85, SE = 49.15, $F(1, 33957.01) = 4.38$, $p = 0.036$, 95% CI = [-199.17, -6.52]), a significant positive effect of the prime-target dissimilarity (estimate = 2.21, SE = 0.725, $F(1, 33957.01) = 9.29$, $p = 0.002$, 95% CI = [0.79, 3.63]), but no significant effect of the distractor-target dissimilarity on response times (estimate = -1.09, SE = 0.715, $F(1, 9094.62) = -1.52$, $p = 0.128$, 95% CI = [-2.49, 0.31]). These results show that the distractors'

interference is mediated by the semantic priming effect where lower prime-target dissimilarity facilitates search in the ambiguous visual condition introduced by the distractor.

Table 3.2. The effect of prime-target and distractor-target dissimilarity and their interaction in response times.

predictor	estimate	SE	F-value	*p*	95% CI
Distractor-target diss.	-1.09	0.715	-1.52	0.128	[-2.49 0.31]
Prime-target : distractor-target diss.	-102.85	49.15	4.38	0.036	[-199.17, -6.52]
Prime-target diss.	2.21	0.725	9.29	0.002	[0.79, 3.63]

3.4.3.3 Semantic priming facilitates search in ambiguous visual conditions

To further demonstrate the facilitatory effect of prime-target dissimilarity, we performed a median split on the distractor-target dissimilarities to isolate trials where the distractor is semantically related to the target (trials below the median distractor-target dissimilarity). This enabled us to replicate results from Experiment 1 proving that linguistic prime facilitates search behaviour in ambiguous visual conditions introduced by the distractor. As expected, the random slope LME (Eq. 1) revealed a significant positive effect of the prime-target dissimilarity on the log-transformed response times in the low distractor-target dissimilarity trails (Figure 3.6A; estimate = 1.036, SE = 0.431, $F(1, 44.88) = 5.79$, $p = 0.020$, 95% CI = [0.19, 1.88]). This corroborates the finding that linguistic primes promote visual targets the more semantically related they are even if the distractors' are closely related to the target. This effect disappeared with increasing distractor-target dissimilarity. We found no significant effect of prime-

target dissimilarity in high distractor-target dissimilarity trials (Figure 3.6B; estimate = 0.266, SE = 0.375, $F(1, 2105.54) = 0.50$, $p = 0.479$, 95% CI = [-0.47, 1.00]). It is possible that the semantic priming effect becomes redundant when the interference between the distractor and target is low, and the low-level perceptual features are sufficient to guide the visual search behaviour.

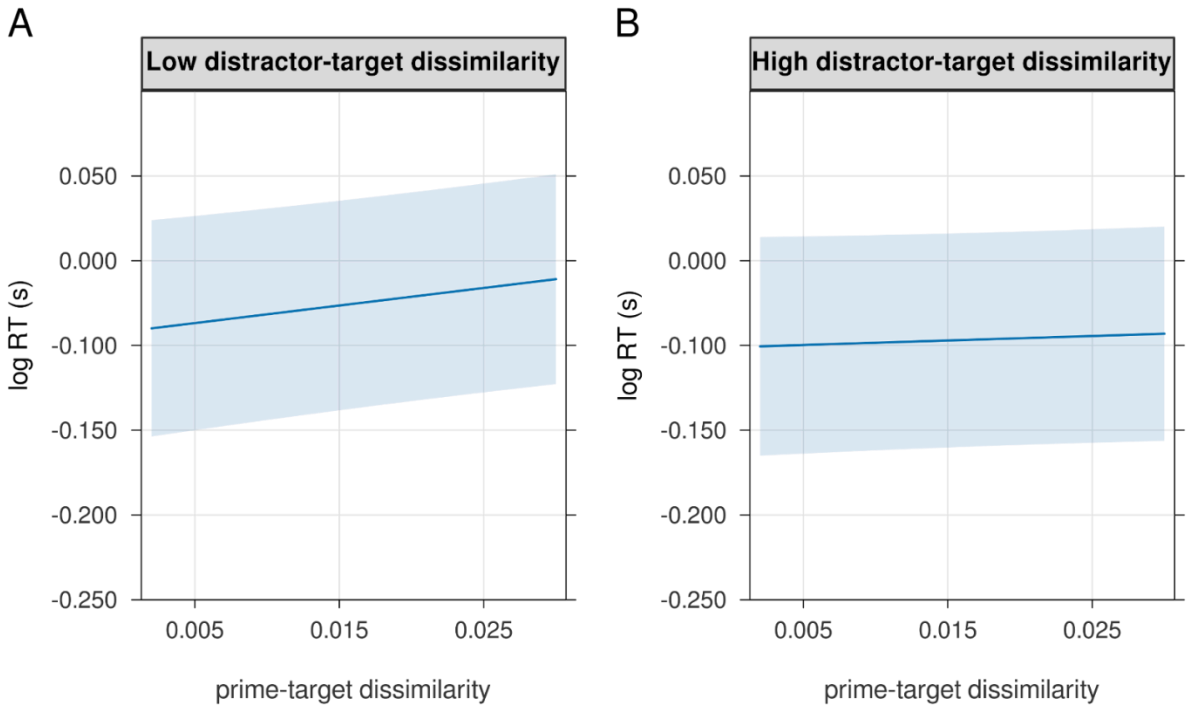


Figure 3.6 | Cross-modal priming effect in low and high distractor-target dissimilarity trials. Random slope LME was evaluated on log-transformed response times with (A) prime-target dissimilarity as the fixed effect where distractor-target dissimilarity is low as defined relative to the median split, and (B) prime-target dissimilarity as the fixed effect where distractor-target dissimilarity is high as defined relative to the median split.

3.5 Discussion

In this study, we were motivated by the hypothesis that vision and language align in their ability to translate their sensory inputs into a unified representational

space of perceived dissimilarities (Chapter 2). We developed a novel cross-modal priming experiment to provide behavioural evidence that the internal representational structure of such a space, traditionally associated with vision (Charest et al., 2014; Cichy et al., 2019; Connolly et al., 2012; Edelman, 1998; Konkle, & Oliva 2012; Kriegeskorte et al., 2008; Mur et al., 2013; Wardle et al., 2016), is readily accessible to language. We found a clear cross-modal priming effect where response times were faster the closer the linguistic prime and the visual target were represented in the similarity space. The most important finding was that the cross-modal priming effect was best predicted using linguistic dissimilarities obtained from each participant's similarity judgements data. These findings are significant in at least two major respects: firstly, linguistic prime activates a prior representation that generalises to a representation obtained in vision. Secondly, this relationship seems to be closely linked to the participant-specific internal representational geometry, which extends previous knowledge of linguistic priming (Harder, & Tylén, 2019; Lupyan, & Spivey, 2010; Lupyan, & Swingley, 2012; Meteyard et al., 2007; Meyer et al., 2007; Sperber et al., 1979) and opens a discussion as to what level interindividual differences in language shape how we represent the visual world.

Additionally, our findings show that the cross-modal representation mapping of the multidimensional space is characterised by many of the same properties that were previously identified in solely visual experiments (Carlson et al., 2014; Cohen et al., 2014; Cohen et al., 2015; Cohen et al., 2017; Ritchie et al., 2015). We replicated prior research that established the effect of visual interference on visual search behaviour (Cohen et al., 2014; Cohen et al., 2015; Cohen et al., 2017; Duncan, & Humphreys, 1989; Moores et al., 2003; Sripati, & Olson, 2010). We showed that the less dissimilar visual distractors were to the target, the longer participants took to identify the target. According to previous research (Duncan, & Humphreys, 1989), such interference happens when multiple

inputs activate a similar representation taking up resources to efficiently guide search behaviour. But what was the prime's role in the visual interference? The results revealed that the cross-modal priming effect was most significant in trials with high distractor-target interference. In these trials, the distractor was sampled to be close to the target within the representational space. It is therefore likely that both the distractor and the target were not only semantically related but mainly activated many overlapping low-level perceptual features in the representational space (i.e., items that share semantic features also often look similar Bracci et al., 2017; Martin et al., 2018) making it difficult to rely solely on low-level perception to guide the visual search. As a result, we speculate that linguistic primes activated semantic properties within the representational space, which then made it easier to identify the visual target in ambiguous visual conditions where visual evidence may be insufficient to guide the search (Yeh, & Peelen, 2022). Consistent with this explanation, the cross-modal priming effect diminished once there was low interference between the distractor and the target. This is in alignment with prior research that shows that if perceptual features are sufficiently distinctive, they take precedence over semantic information in visual search behaviour (Yeh, & Peelen, 2022). These findings are important because they suggest that the representational similarity space flexibly accommodates the cross-modal inputs where both semantic and perceptual properties contribute to form a unified internal representational geometry (Bracci et al., 2017; Martin et al., 2018) for us to observe such an intricate interaction between them.

Further support for this comes from predictions made by idiosyncratic dissimilarities that were extracted from the linguistic and visual similarity judgements. We found that response times were significantly predicted by idiosyncratic prime-target dissimilarities taken from the linguistic similarity judgements, but no effect was found when we used idiosyncratic dissimilarities

from the visual similarity judgements. As the model of one-to-many mapping suggests (Lupyan, & Spivey, 2010; Lupyan, 2012), it is possible that reading a sentence caption evokes multiple meanings and visual interpretations of that sentence that can warp many different visual representations. We speculate that reading the linguistic prime activated an idiosyncratic aspect of the representational space, which then seamlessly generalised to the visual representation (Athanasopoulos et al., 2015; Lupyan, & Spivey, 2010; Winawer et al., 2007). It is, therefore, safe to assume that when participants performed the similarity judgements, they likely relied on similar idiosyncratic representations, which then created a universal similarity structure that was predictive of the visuo-linguistic alignment in real time. This explanation is further exemplified by the fact that the relationship between the response times and idiosyncratic linguistic dissimilarities was irreplicable under the null distribution (Supplementary Figure 3.1). This opens a further discussion (LaTourrette, & Waxman, 2020; LaTourrette et al., 2023; Lupyan, & Spivey, 2010; Lupyan, 2012; Winawer et al., 2007) as to what extent the idisyncratic linguistic semantics shapes the representational similarity space.

Conversely, lower inter-individual variability in the visual similarity judgements (Chapter 2) likely explains why we did not observe a significant effect when using the idiosyncratic visual dissimilarity. Perhaps, any idiosyncratic aspect that the visual dissimilarity captured did not carry enough power to predict response times at the individual level. This explains why we observed significant predictions only for the averaged idiosyncratic visual dissimilarity and the referential visual dissimilarity taken from the NSD representational space. Here, we were predicting participants' response times by the same set of dissimilarities raising the possibility that significant predictions were supported by aggregate group-level relationships. The fact that linguistic predictions were significant at both individual- and group-level further confirms that language taps into the

shared representational space while also accounting for idiosyncratic aspects. It is worth emphasising that the difference in the visual and linguistic predictions cannot be explained by the order of similarity judgement tasks. Previously, we showed that interindividual differences in similarity judgements vary as a function of whether participants completed the linguistic or visual similarity judgements first (Chapter 2). We controlled for this confound by adding the task order as an additional predictor in the LME with idiosyncratic dissimilarities as fixed effects, but no significant effect of task order was found (Supplementary Table 3.2 and Supplementary Table 3.3). This further shows that the difference in the idiosyncratic predictions was related to the semantic prime-target link and the idiosyncratic aspect of this semantic association.

What is the exact nature of the visuo-linguistic alignment? A much-debated question has been regarding the possibility that language relates to visual representations only because it depicts what objects are present regardless of the statistical regularities it captures (Conwell et al., 2023). This would suggest that the linguistic prime elicited an object-specific neural representation that only overlapped with the visual representation based on the objects depicted in the sentence. For instance, reading a sentence such as "A cat lying next to a pair of black boots" would elicit a corresponding representation of a cat and black boots independent of the semantic relationship between them. This, however, cannot explain why we found the link between the prime and the target to facilitate visual search in trials with high distractor-target interference, or why we found the idiosyncratic linguistic dissimilarity, but not idiosyncratic visual dissimilarity, to be a suitable predictor of response times. Our findings somewhat relate to evidence from developmental and cross-linguistic studies that show that labelling can emphasise diagnostic features of visual objects (Althaus, & Mareschal, 2014; Ferry et al., 2013; LaTourrette, & Waxman, 2020; LaTourrette et al., 2023; Waxman, & Markow, 1995), which are then represented as more similar (respective as

members of one category). The linguistic prime likely drew on a store of personally meaningful representations that, by definition, were more general and flexibly warped the representation of the visual target. Such a relationship was then best predicted by the salience of the linguistic idiosyncratic dissimilarity that was defined by more abstract semantic relations than just by what objects are present, further demonstrating how individual experience may shape the shared semantic space.

In summary, the success in predicting the cross-modal priming effect using behavioural similarity judgements suggests that perceived dissimilarities present an adequate framework for characterising the internal representation of the shared representational space. Our findings provide behavioural relevance to a visuo-linguistic framework where representations in vision and language are projected onto a unified similarity space – a framework previously found to benefit computational tasks (Andrews et al., 2009; Kiela, & Bottou, 2014; Silberer, & Lapata, 2014; Vong et al., 2024; Wang et al., 2023). Future research should further investigate how vision and language project onto the shared space in real-time and to what degree the intricacies of one's language shape its internal geometry.

3.6 References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., Hutchinson, J. B., Naselaris, T., & Kay, K. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience, 25*(1), 116-126.

Althaus, N., & Mareschal, D. (2014). Labels direct infants' attention to commonalities during novel category learning. *PLoS One, 9*(7).

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review, 11*(3), 463-498.

Athanasopoulos, P., Bylund, E., Montero-Melis, G., Damjanovic, L., Schartner, A., Kibbe, A., Riches, N., & Thierry, G. (2015). Two languages, two minds: flexible cognitive processing driven by language of operation. *Psychological Science, 26*(4), 518-526.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language, 68*(3), 255-278.

Bates, D., Machler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1-48.

Bergelson, E., & Swingley, D. (2011). At 6-9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences of the United States of America, 109*(9), 3253-3258.

Bracci, S., Ritchie, J. B., de Beeck, H. O. (2017). On the partnership between neural representations of object categories and visual features in the ventral visual pathway. *Neuropsychologia, 105*, 153-164.

Bruni, E., Tran, N. K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research, 49*(1), 1-47.

Carlson, T. A., Ritchie, J. B., Kriegeskorte, N., Durvasula, S., & Ma, J. (2014). Reaction time for object categorization is predicted by representational distance. *Journal of Cognitive Neuroscience, 26*(1), 132-142.

Charest, I., Kievit, R. A., Schmitz, T. W., Deca, D., & Kriegeskorte, N. (2014). Unique semantic space in the brain of each beholder predicts perceived similarity. *Proceedings of the National Academy of Sciences of the United States of America, 111*(40), 14565-14570.

Cichy, R. M., Kriegeskorte, N., Jozwik, K. M., van den Bosch, J. J., & Charest, I. (2019). The spatiotemporal neural dynamics underlying perceived similarity for real-world objects. *Neuroimage, 194*, 12-24.

Cohen, M. A., Alvarez, G. A., Nakayama, K., & Konkle, T. (2017). Visual search for object categories is predicted by the representational architecture of high-level visual cortex. *Journal of Neurophysiology, 117*(1), 388-402.

Cohen, M. A., Konkle, T., Rhee, J. Y., Nakayama, K., & Alvarez, G. A. (2014). Processing multiple visual objects is limited by overlap in neural channels. *Proceedings of the National Academy of Sciences of the United States of America, 111*(24), 8955-8960.

Cohen, M. A., Nakayama, K., Konkle, T., & Stantic, M. (2015). Visual awareness is limited by the representational architecture of the visual system. Journal of *Cognitive Neuroscience, 27*(11), 2240-2252.

Connolly, A. C., Guntupalli, J. S., Gors, J., Hanke, M., Halchenko, Y. O., Wu, Y. C., Abdi, H., Haxby, J. V. (2012). The representation of biological classes in the human brain. *The Journal of Neuroscience, 32*(8), 2608-2618.

Conwell, C., Prince, J. S., Alvarez, G., & Konkle, T. (2023). Language models of visual cortex: where do they work? And why do they work so well where they do? *Journal of Vision.*

Daoutis, C. A., Franklin, A., Riddett, A., Clifford, A., & Davies, I. R. (2006). Categorical effects in children's colour search: a cross-linguistic comparison. *The British Journal of Developmental Psychology, 24*(2), 373-400.

Dils, A. T., & Boroditsky, L. (2010). Visual motion aftereffect from understanding motion language. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(37), 16396-16400.

Doerig, A., Kietzman, T. C., Allen, E., Wu, Y., Naselaris, T., Kay, K., & Charest, I. (2024). Visual representations in the human brain are aligned with large language models. *arXiv*.

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review, 96*(3), 433-458.

Edelman, S. (1998). Representation is representation of similarities. *The Behavioural and Brain Sciences, 21*(4), 449-467.

Fedorenko, E., Scott, T. L., Brunner, P., Coon, W. G., Pritchett, B., Schalk, G., & Kanwisher, N. (2016). Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences of the United States of America, 113*(41), 6256-6262.

Ferry, A. L., Hespos, S. J., & Waxman, S. R. (2013). Nonhuman primate vocalizations support categorization in very young human infants. *Proceedings of the National Academy of Sciences of the United States of America, 110*(38), 15231-15235.

Gayet, S., & Peelen, M. V. (2022). Preparatory attention incorporates contextual expectations. *Current Biology, 32*(3), 687-692.

Harder, M., & Tylen, K. (2019). What is up and down in embodied language processing? An experimental study on semantic priming of visual perception. *Cognitive Semiotics, 12*(2).

Kiela, D., & Bottou, L. (2014). Learning image embeddings using convolutional neural networks for improved multi-modal semantics. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP*), 36-45.

Kriegeskorte, N., & Mur, M. (2012). Inverse MDS: inferring dissimilarity structure from multiple item arrangements*. Frontiers in Psychology, 3*.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P. A. (2008). Matching categorical representations in inferior temporal cortex of man and monkey. *Neuron, 60*(6), 1126-1141.

Konkle, T., & Oliva, A. (2012). A real-world size organization of object responses in occipitotemporal cortex. *Neuron, 74*(6), 1114-1124.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science, 207*(4427), 203-205.

LaTourrette, A. S., & Waxman, S. R. (2020). Naming guides how 12-month-old infants encode and remember objects. *Proceedings of the National Academy of Sciences of the United States of America, 117*(35), 21230-21234.

LaTourrette, A., Chan, D. M., & Waxman, S. R. (2023). A principled link between object naming and representation is available to infants by seven months of age. *Scientific Reports, 13*(14328).

Li, J., Kementchedjhieva, Y., Fierro, C., & Sogaard, A. (2023). Do vision and language models share concepts? A vector space alignment study. *arXiv*.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., & Zitnick, C. L. (2014). Microsoft COCO: common objects in context. *Computer Vision - ECCV 2014*, 740-755.

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining task-agnostic visuolinguistic representations for vision-and-language tasks. *arXiv*.

Lupyan, G. (2012). Linguistically modulated perception and cognition: the label-feedback hypothesis. *Frontiers in Psychology, 3,* 54.

Lupyan, G., & Spivey, M. J. (2010). Making the invisible visible: verbal but not visual cues enhance visual detection. *PLoS One, 5*(7).

Lupyan, G., & Swingley, D. (2012). Self-directed speech affects visual search performance. *Quarterly Journal of Experimental Psychology, 65*(6), 1068-1085.

Martin, C. B., Douglas, D., Newsome, R. N., Man, L. L., & Barense, M. D. (2018). Integrative and distinctive coding of visual and conceptual object features in the ventral visual stream. *eLife, 7.*

Mazerolle, M. J. (2023, March 20). Model selection and multimodel inference using the AICcmodavg package. https://mirrors.cloud.tencent.com/CRAN/web/packages/AICcmodavg/vignettes/AICcmodavg.pdf

Meteyard, L., Bahrami, B., & Vigliocco, G. (2007). Motion detection and motion verbs: language affects low-level visual perception. *Psychological Science, 18*(11), 1007-1013.

Meyer, A. S., Belke, E., Telling, A. L., & Humphreys, G. W. (2007). Early activation of object names in visual search. *Psychonomic Bulletin & Review, 14*(4). 710-716.

Moores, E., Laiti, L., & Chelazzi, L. (2003). Associative knowledge controls deployment of visual selective attention. *Nature Neuroscience, 6*(2), 182-189.

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A., & Kriegeskorte, N. (2013). Human object-similarity judgements reflect and transcend the primate-IT object representation. *Frontiers in Psychology, 4*.

R core Team (2023). R: A language and environment for statistical computing. Vienna, Austria. https://www.R-project.org/.

Ritchie, J. B., Tovar, D. A., & Carlson, T. A., (2015). Emerging object representations in the visual system predict reaction times for categorization. *PLoS Computational Biology, 11*(6).

Roberson, D., Davidoff, J., Davies, I. R., & Shapiro, L. R. (2005). Color categories: evidence for the cultural relativity hypothesis. *Cognitive Psychology, 50*(4), 378-411.

Roberson, D., Davies, I., & Davidoff, J. (2000). Color categories are not universal: replications and new evidence from a stonge-age culture. *Journal of Experimental Psychology, 129*(3), 369-398.

Sadeghi, Z., McClelland, J. L., & Hoffman, P. (2015). You shall know an object by the company it keeps: an investigation of semantic representations derived from object co-occurrence in visual scenes. *Neuropsychologia, 76*, 52-61.

Silberer, C., & Lapata, M. (2014). Learning grounded meaning representations with autoencoders*. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 1*, 721-732.

Sperber, R. D., McCauley, C., Ragain, R. D., & Weil, C. M. (1979). Semantic priming effects on picture and word processing. *Memory & Cognition, 7*(5), 339-345.

Sripati, A. P., & Olson, C. R. (2010). Global image dissimilarity in macaque inferotemporal cortex predicts human visual search efficiency. *The Journal of Neuroscience, 30*(4), 1258-1269.

Stanfield, R. A., & Zwaan, R. A. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science, 12*(2), 153-156.

Thierry, G., Athanasopoulos, P., Wiggett, A., Dering, B., & Kuipers, J. R. (2009). Unconscious effects of language-specific terminology on preattentive color perception. *Proceedings of the National Academy of Sciences of the United States of America, 106*(11), 4567-4570.

Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science, 10*(2), 172-175.

Van Berkum, J. J., Hagoort, P., & Brown, C. M. (1999). Semantic integration in sentences and discourse: evidence from the N400. Journal of Cognitive Neuroscience, 11(6), 657-671.

Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. M. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science, 383*(6682), 504-511.

Wang, A. Y., Kay, K., Naselaris, T., Tarr, M. J., & Wehbe, L. (2023). Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence, 5,* 1415-1426.

Wardle, S. G., Kriegeskorte, N., Grootswagers, T., Khaligh-Razavi, S. M., & Carlson, T. A. (2016). Perceptual similarity of visual patterns predicts dynamic neural activation patterns measured with MEG. *NeuroImage, 132*, 59-70.

Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: evidence from 12- to 13-month-old infants. *Cognitive Psychology, 23*(3), 257-302.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences of the United States of America, 104*(19), 7780-7785.

Woolnough, O., Donos, C., Murphy, E., Rollo, P. S., Roccaforte, Z. J., Dehaene, S., & Tandon, N. (2023). Spatiotemporally distributed frontotemporal networks for sentence reading. *Proceedings of the National Academy of Sciences of the United States of America, 120*(17).

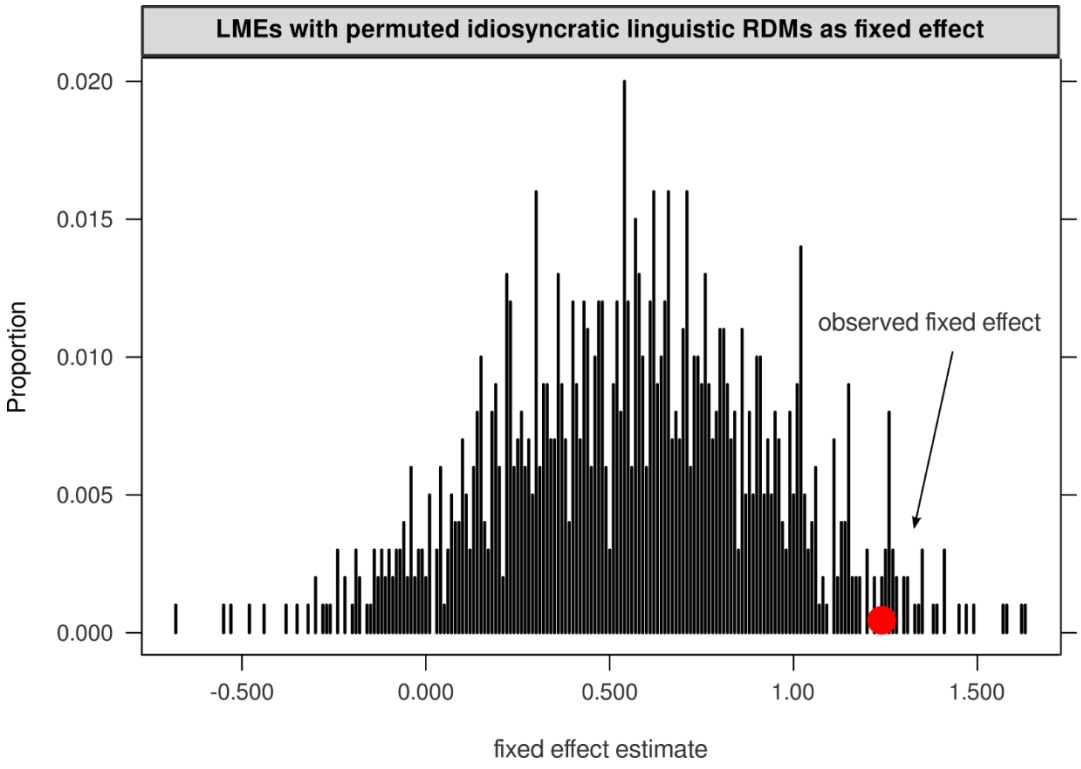
Yeh, L. C., & Peelen, M. V. (2022). The time course of categorical and perceptual similarity effects in visual search. *Journal of Experimental Psychology. Human Perception and Performance, 48*(10), 1069-1082.

Yeh, L. C., Throat, S., & Peelen, M. V. (2024). Predicting cued and oddball visual search performance from neural representational similarity. *bioRxiv*.

3.7 Supplementary material

Supplementary Table 3.1. Overview of removed data in Experiment 1, MA subset of participants from Experiment 1, and Experiment 2.

	participants	incorrect			invalid			outliers			identity condition		
		total	*mean*	*SD*	*total*	*mean*	*SD*	*total*	*mean*	*SD*	*total*	*mean*	*SD*
Exp. 1	11	5316	63	25	1	-	-	1025	12	20	23129	275	17
Exp. 1	1	1745	67	33	1	-	-	383	15	34	7074	272	15
Exp. 2	5	2531	55	36	2	0.4	0.2	414	9	18	-	-	-



Supplementary Figure 3.1 | LME predictions using permuted idiosyncratic linguistic RDMs. Idiosyncratic linguistic RDMs were permuted between participants such that each participants' response times were predicted by someone else's representational

geometry. The figure shows distribution of fixed effect estimates that were obtained from random slope LMEs where permuted dissimilarities were used as predictors. The distribution shows that less than 5% of LMEs produced a fixed effect estimate as high as the observed estimate.

Supplementary Table 3.2. The effect of prime-target dissimilarity obtained from idiosyncratic visual space and the task order effect on response times.

predictor	estimate	SE	F-value	*p*	95% CI
Prime-target diss.	0.113	0.554	0.040	0.841	[-0.97, 1.20]
Task order	0.093	0.069	1.820	0.189	[-0.04, 0.23]

Supplementary Table 3.3. The effect of prime-target dissimilarity obtained from idiosyncratic linguistic space and the task order effect on response times.

predictor	estimate	SE	F-value	*p*	95% CI
Prime-target diss.	1.237	0.486	6.470	0.016	[0.28, 2.19]
Task order	0.098	0.068	2.110	0.158	[-0.03, 0.23]

Chapter 4

Representational dynamics of vision and language in perception and imagery

4.1 Abstract

Emerging evidence suggests that the latent semantic space obtained from models jointly trained on visual and linguistic data improves brain response prediction in both vision and language. It remains unclear, however, how vision and language share meaningful representations in the brain. Here, we collected ~90 hours of electroencephalography (EEG) data from seven participants while they performed perception and imagery tasks on 100 natural scene images and 100 sentence captions describing the images. We used representational similarity analysis (RSA) to investigate if, and if so how, representations in vision and language project onto a representational space derived from behavioural similarity judgements. We found that the multidimensional space significantly correlated with neural geometry in visual perception as early as 242 ms after the image onset. A comparison to feature-based models revealed that this overlap envelopes features of colour, distance, animacy, and place. We also observed some alignment between the similarity space and representational dynamics in sentence reading and sentence imagery conditions, however, these comparisons did not survive after cluster-based correction. Further assessment of the noise ceiling pointed to the high measurement noise to be the most likely reason why only a limited alignment was found. Finally, we used linear discriminant analysis (LDA) to validate the categorical distinction identified in the representational similarity space. The most reliable categorical representations were identified in visual perception, including successful cross-decoding for categorical representation of people and outdoor scenes in vision and language. Together, these results showed some, albeit limited, representational overlap in visuo-linguistic perception and imagery. We discuss the implications of our findings in light of emerging attempts to develop a unified visuo-linguistic framework and offer recommendations for future research.

4.2 Introduction

Mapping language to corresponding mental representations, including visual images, is critical for language comprehension and learning (Bergen et al., 2007; Savarimuthu, & Ponniah, 2024; Speed et al., 2024). But even when visual imagery is not explicitly necessary to perform a linguistic task (Savarimuthu, & Ponniah, 2024), language spontaneously forms a prior image (Barsalou, 1999; Dils, & Boroditsky, 2010; Liu et al., 2024; Stanfield, & Zwaan, 2001; Zwaan et al., 2002) that facilitates behavioural performance in visual tasks (Chapter 3). Recent research has revealed that representational geometry obtained in similarity judgements of sentences predicts many of the same regions that are engaged during visual perception (Chapter 2). Most importantly, this effect remained even when participants engaged with sentence descriptions of corresponding images that they had never seen before. A much-debated possibility is that language draws on visual representations in a shared multidimensional space that encodes the statistical regularities of the world (Bonner, & Epstein, 2021; Doerig et al., 2024; Stanfield, & Zwaan, 2001). The space appears to be multidimensional in that it encodes the rich world knowledge in its visual (Golomb, & Kanwisher, 2012; Gomez et al., 2018; Kaiser, & Cichy, 2018) along with semantic aspects (Draschkow, & Vo, 2017; Yeh, & Peelen, 2022) and generalises across modalities (Tang et al., 2023; Wang et al., 2023). While it is now well established that neural activity patterns in vision reflect the representational geometry of such a space (Charest et al., 2014; Cichy et al., 2019; Mur et al., 2013; Wardle et al. 2016), its role in language comprehension remains less clear.

Mental imagery offers a plausible way to experimentally investigate how we represent knowledge about the world in vision and language. Most research has focused on visual imagery and the extent to which it overlaps with corresponding perceptual representations in the visual cortex. Early research established that

visual images maintain many perceptual properties that enable the observer to inspect the mental image just as if they were to inspect a real-world scene using their sight (Lewis et al., 2011; Kosslyn, 1975; Kosslyn et al., 1978). Neuroimaging evidence associated this with overlapping representations activated in the visual cortex during both visual perception and imagery (Albers et al., 2013; Dijkstra et al., 2017; Dijkstra et al., 2019; Ganis, & Schendan, 2008; Harrison, & Tong, 2009; Horikawa, & Kamitani, 2017; Klein et al., 2004; Lee et al., 2012). To date, only a handful of studies have used spoken word (Mitchell, & Cusack, 2016; Xie et al., 2020) or scene descriptions (Stecher, & Kaiser, 2023) to induce mental imagery, all of which support the idea that the mental and perceptual representations converge at multiple levels of abstraction including visual, semantic, and emotional properties. Several authors have recently even reported that, if adequately combined, vision and language carry complementary representations that can improve performance in many computational tasks (Collell, & Moens, 2016; Collell, & Moens, 2018; Lahner et al., 2024; Tang et al., 2023; Wang et al., 2023). This converging evidence underscores the importance of developing a unified framework for visuo-linguistic integration, which could potentially lead to advancements in models of human cognition. A challenge, however, remains in how we can describe the representational overlap in real time.

We recorded electroencephalography (EEG) data while participants performed perception and imagery tasks on 100 natural scene images and 100 sentence captions describing the images. The stimulus set was obtained from the Natural Scenes Dataset (NSD, Allen et al., 2022) and covered the full semantic space, based on embeddings of the images' captions. This and the fact that the stimuli were natural scenes allowed us to obtain an extensive sampling of the neural representational geometry in its visual and linguistic format. We then used representational similarity analysis (RSA, Kriegeskorte et al., 2008) to investigate if, and if so how, neural geometry in vision and language aligns with the internal

geometry of the multidimensional space. We modelled such a space by re-weighting similarity judgements of the visual and linguistic stimuli (Chapter 2), which allowed us to understand the time course of how the brain evaluates semantic associations and similarities in vision and language. We then used feature-based models to reveal the precise nature of the neural geometry and explore whether neural geometries across the modalities retain modality-specific characteristics. Finally, we trained a linear discriminant classifier to validate the categorical structure of the semantic space and to better understand at what point representations from multimodal systems converge and translate into behaviour.

4.3 Results

Seven participants attended eight EEG sessions in which they perceived and imagined 100 natural scene images and 100 sentence captions describing the images. Participants were counterbalanced across the task modalities such that four participants started the experiment with the visual modality in their first four EEG sessions and continued with the linguistic modality in the remaining EEG sessions. Three participants completed the experiment in reversed order. Each trial started with a perception phase where participants either viewed an image in the visual modality (Figure 4.1A) or read a sentence in the linguistic modality (Figure 4.1B). This was followed by an imagery phase in which participants imagined the content of the stimulus and maintained the mental image for 3 seconds. Participants then rated the vividness of the imagery on a scale between 1 (no image at all) and 5 (as vivid as normal vision). Wilcoxon signed-rank test did not reveal a significant difference in vividness ratings between visual (mean = 3.831, SD = 0.441) and linguistic imagery (mean = 3.832, SD = 0.611; W = 10, z = 0, p = 1, one participant was excluded from this analysis due to the completion of only one modality). Each trial ended with an attention check where participants

decided whether the displayed stimulus matched the one shown in the perception phase. Participants performed well on the attention checks with an average accuracy of 84% (SD = 9.97%) and 88.71% (SD = 6.28) in visual and linguistic modality respectively. We found that the average vividness rating was higher in trials where participants successfully matched the stimuli (W = 22, z = -2.153, p = 0.031). This is a reassuring finding because it implies that the quality of the imagery was determined by participants' attentiveness to stimuli displayed in the perception phase.

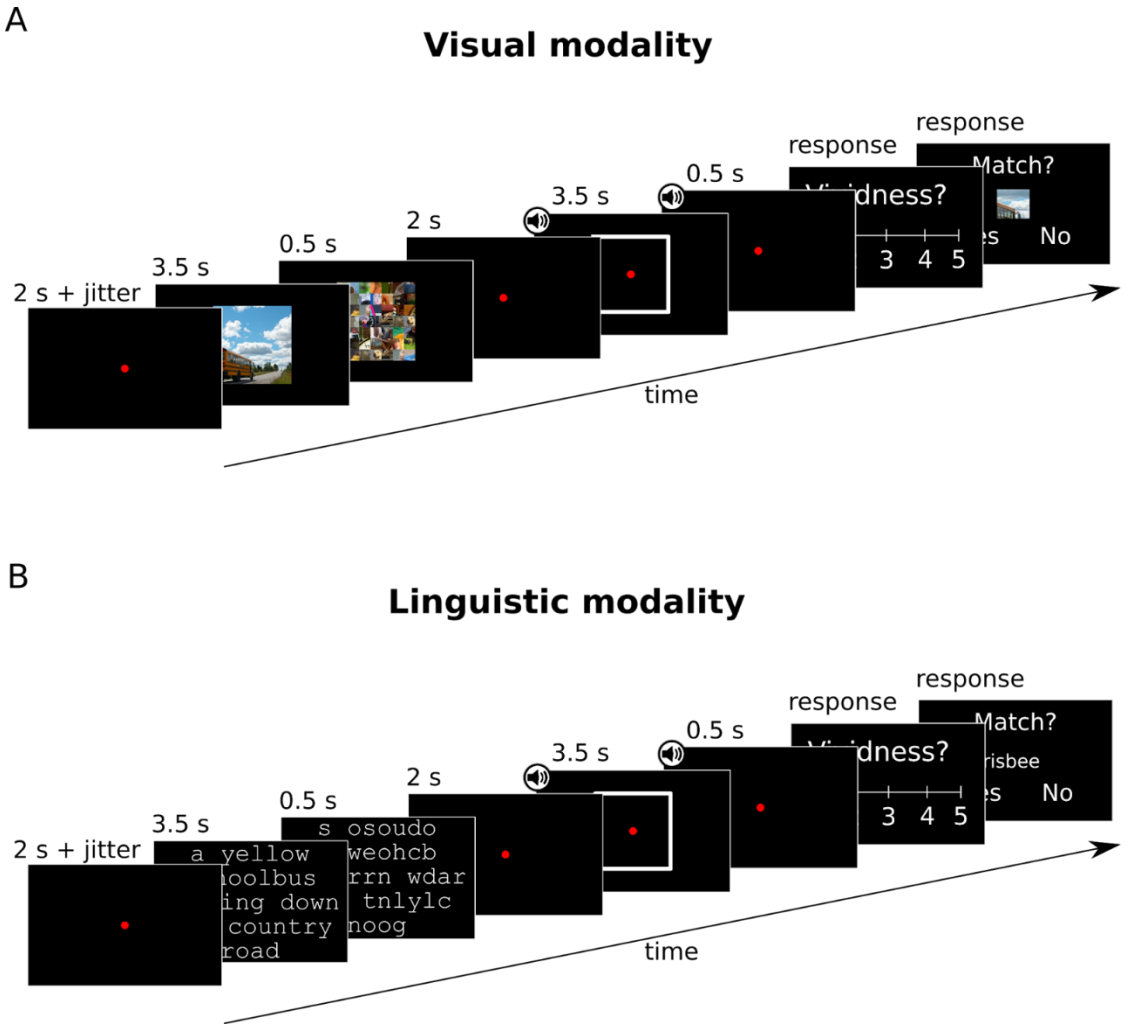


Figure 4.1 | EEG experimental design. Participants were counterbalanced across eight experimental sessions in which they completed two imagery tasks while we recorded their EEG. (A) Four participants started with the visual modality in their first four EEG

sessions. They were shown one image at a time from the set of 100 images taken from the Natural Scenes Dataset (NSD, Allen et al., 2022). This was followed by a visual mask of scrambled images. Participants were then asked to imagine the image on the sound cue and hold the imagery for 3 seconds. Then, they rated the vividness of the imagery and decided whether the tile displayed on the screen matched the image from the perception phase. (B) Three participants started with the linguistic modality in which they viewed sentence captions of the images taken from the NSD. The procedure was identical to the one in the visual modality, but the visual mask consisted of scrambled sentences and the attention check at the end of the trial presented a word. In both modalities, participants completed 3-4 runs per EEG session each presenting the full set of 100 stimuli. Participants switched the task modality once they attended the first four sessions.

4.3.1 Representational dynamics in multimodal perception and imagery predicted by multidimensional similarity space

Similarity judgements in vision and language were found to project their representations onto a multidimensional similarity space that aligned with visual brain responses (Chapter 2). How does the multidimensional similarity space align with a time-varying representation in multimodal perception and imagery? To represent the multidimensional space, we first created a weighted similarity model by re-weighting geometries from behavioural similarity judgements (Chapter 2) that were performed on the same set of images and sentences used in the present Chapter (Figure 4.2A). We then compared the weighted similarity model to time-resolved EEG RDMs to investigate whether the representational geometry of the model correlates with neural representational dynamics. We found that the neural dynamics of visual perception aligned with the weighted similarity model as early as 242 ms after the image appeared on the screen (peak t value $t(5) = 8.919$, $p < 0.05$ cluster-corrected). The cosine similarities peaked 578 ms post-stimulus followed by a sharp decline around 800 ms. As shown in Figure 4.2B, another significant representational overlap appeared between 871 and

1352 ms after the stimulus onset (peak t value $t(5) = 8.024$, $p < 0.05$ cluster-corrected) with peak cosine similarity at 1070 ms. These findings suggest that neural dynamics of visual perception aligns with the multidimensional similarity space. However, no overlap was found between the weighted similarity model and representational dynamics in visual imagery (Figure 4.2C).

Of particular interest here is how neural representations in language comprehension align with the multidimensional similarity space. First, we compared the weighted similarity model to neural dynamics in the sentence reading condition. As can be seen from Figure 4.2D, the time-resolved comparisons revealed cosine similarities significant at the uncorrected $p < 0.05$ level appearing early around the sentence reading onset. This was followed by a sparsely populated cluster of uncorrected cosine similarities each less than 89 ms apart lasting between 328 and 512 ms (9 time points, $p < 0.05$ uncorrected). Then, cosine similarities peaked 1414 ms after the sentence onset, which was adjacent to another sparsely populated cluster of cosine similarities significant at the uncorrected level lasting between 1348 and 1508 after the sentence onset (30 time points, $p < 0.05$ uncorrected). In contrast to findings in visual imagery, we also observed a significant alignment between the weighted similarity model and neural dynamics of linguistic imagery lasting from 891 ms to 1027 ms after the imagery onset (17 time points, $p < 0.05$ uncorrected; Figure 4.2E). Together, these findings indicate that there is some link between the multidimensional similarity space and the neural dynamics of language comprehension, but a note of caution is due here as neither of these comparisons remained after we performed the cluster-based permutation test for p-value correction.

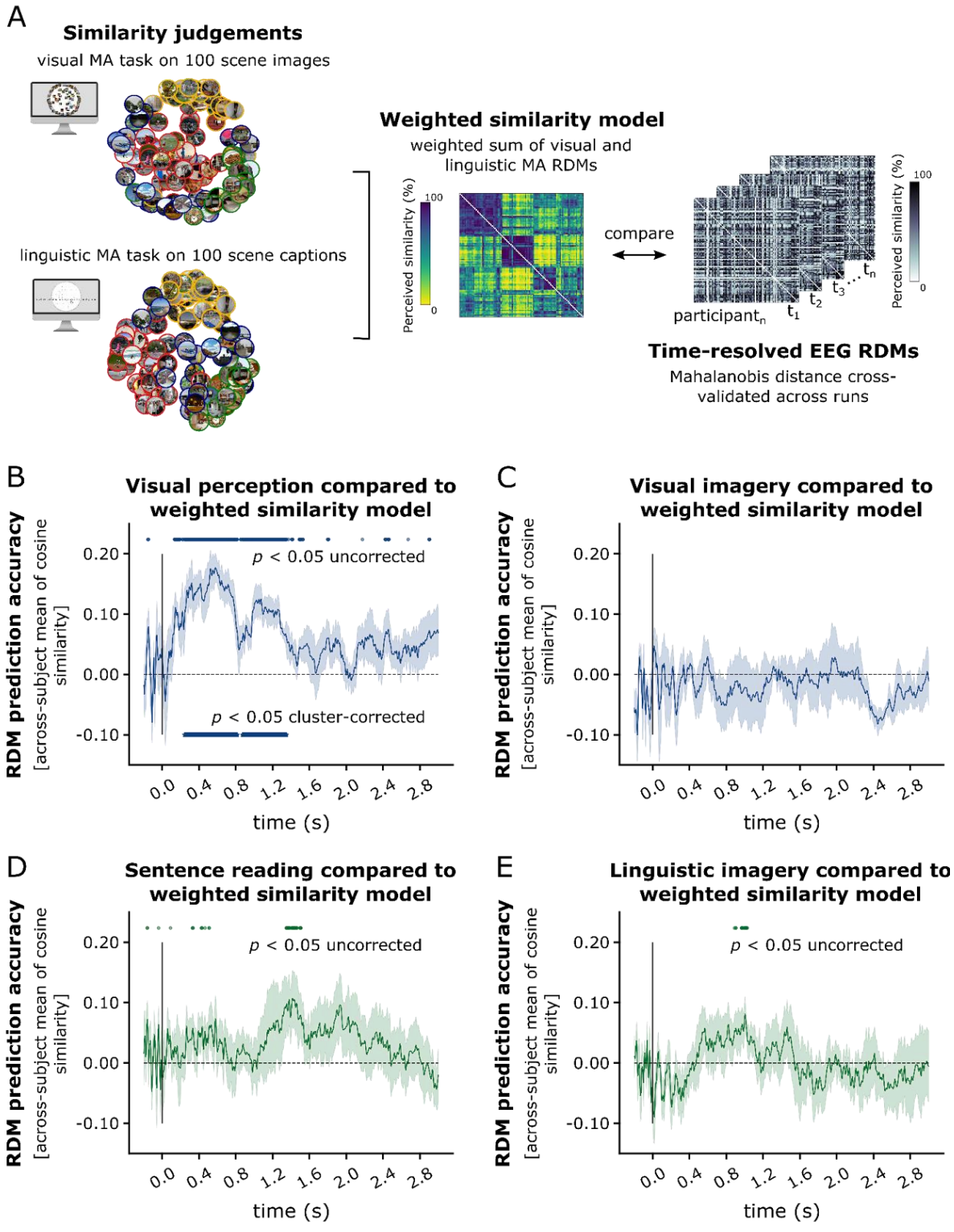


Figure 4.2 | Neural representational dynamics compared to the weighted similarity model of behavioural similarity judgements. (A) A weighted similarity model was computed as a weighted sum of RDMs obtained from visual (top) and linguistic (bottom) similarity judgements. Similarity judgements are projected onto a 2D plane using

multidimensional scaling (MDS) where the distance between the items represents their perceived dissimilarity and the colour represents the category membership (red – people, yellow – animals, green – indoor scenes, blue – outdoor scenes). The weighted similarity model was compared to time-resolved EEG RDMs using cosine similarity. The bottom panel shows averaged cosine similarities across participants in visual perception (B), visual imagery (C), sentence reading (D), and linguistic imagery (E). The shaded area depicts the standard error of the mean (SEM). The points at the top of each figure depict cosine similarities significant at the $p < 0.05$ level (one-sided t-test). The stars at the bottom depict cluster-corrected cosine similarities ($p < 0.05$, cluster-based permutation one-sided t-test).

4.3.2 Mid- and high-level features: unfolding dimensions of neural representational geometry

Previous research identified several mid- and high-level features that contribute to the representational geometry in the brain and behaviour (Charest et al., 2014; Cichy et al., 2019; Contier et al., 2024; Hebart et al., 2020). To quantify their weighting relative to the weighted similarity model, we built four models based on feature-based annotations of colour, distance, animacy, and place of the visual stimulus set (Figure 4.3A). We then fitted the models to a spatiotemporal version of EEG RDMs, which was based on vectorised neural activity from time clusters that were significantly predicted by the weighted similarity model (note that visual imagery was not included in this analysis due to non-significant predictions, see Figure 4.2C). It was hypothesised that if the neural geometry envelopes multidimensional features in order to construct a complete semantic representation (Anderson et al., 2019; Collell, & Moens, 2016; Tang et al., 2021), then the weighted similarity model would outperform the fixed feature-based models.

4.3.2.1 Neural representational geometry of visual perception represents multidimensional features

First, we examine the model fit in the early cluster 242 to 816 ms in visual perception (Figure 4.2B). Here, all four fixed feature-based models yielded significant predictions of the neural representational geometry with an average cosine similarity of 0.155 (SEM = 0.032), 0.138 (SEM = 0.028), 0.138 (SEM = 0.025), and 0.125 (SEM = 0.025) for models of colour, distance, animacy, and place respectively (Figure 4.3C, one-tailed t-test, $p < 0.05$ Bonferroni-corrected for five models). The model that re-weighted the geometry of behavioural similarity judgements produced the highest fit with cosine similarity averaging at 0.178 (SEM = 0.032), but the difference in model performance was not significant (two-tailed test, $q > 0.05$ FDR corrected). As depicted in Figure 4.3C, only models of colour, animacy, and weighted similarity performed significantly above the lower bound of the noise ceiling (two-tailed t-test, $p < 0.05$ Bonferroni-corrected for five models). This indicates that these models explain a substantial portion of the variability in the neural geometry cued by a natural scene image as would be expected by an unknown true model.

Next, we fitted the models to the late cluster in visual perception that lasted between 871 and 1352 ms after the image onset (Figure 4.2B). The neural representational geometry was significantly predicted by feature-based models of colour and place with an average cosine similarity of 0.102 (SEM = 0.030) and 0.088 (SEM = 0.026) respectively, but no alignment with models of animacy or distance was found. Similarly to the early cluster, the weighted similarity model explained the neural geometry best with cosine similarity averaging at 0.115 (SEM = 0.033, one-tailed t-test, $p < 0.05$ Bonferroni-corrected for five models). However, there was no statistical evidence that the weighted model outperformed any of the fixed feature-based models (two-tailed test, $q < 0.05$ FDR corrected), or that

any of the models significantly differed from the lower bound of the noise ceiling (Figure 4.3D, two-tailed t-test, all $p > 0.05$ Bonferroni-corrected for five models).

4.3.2.2 No evidence for multidimensional features being represented in the neural representational geometry of language comprehension

The next set of analyses aimed to identify feature representations in neural geometry of language comprehension. Note that the weighted similarity model produced significant predictions of the time-varying representation only at the uncorrected level of $p < 0.05$. For this reason, the spatiotemporal version of EEG RDMs was based on neural activity from clusters that were created around time points significant at the uncorrected level and less than 100 ms apart. This singled out two clusters in sentence reading with an early cluster lasting between 328 and 512 ms after the sentence onset (Figure 4.2D, 9 time points), and a late cluster lasting between 1348 and 1508 ms after the sentence onset (30 time points). There was no evidence that any of the models explained the neural geometry in the early cluster of sentence reading (Figure 4.3E, one-tailed t-test, all $p > 0.05$ Bonferroni-corrected for five models). We observed average cosine similarities of 0.040 (SEM = 0.037), 0.039 (SEM = 0.032), 0.019 (SEM = 0.025), and 0.018 (SEM = 0.026) for the feature-based models of distance, colour, place, and animacy respectively. The weighted similarity model produced seemingly the highest average correlation (0.052, SEM = 0.092), but pairwise model comparisons found no difference in model performance (two-tailed test, $q > 0.05$ FDR corrected).

We found that the models fitted the later neural geometry in sentence reading slightly better with the highest cosine similarities averaging at 0.101 (SEM = 0.057), 0.096 (SEM = 0.041), and 0.089 (SEM = 0.043) for the weighted model and feature-based models of distance and colour respectively. The feature-based models of animacy and place fitted the neural geometry less with average cosine

similarity of 0.056 (SEM = 0.033) and 0.049 (SEM = 0.033). However, neither of these models performed significantly above zero (Figure 4.3F, one-tailed t-test, all $p > 0.05$ Bonferroni-corrected for five models) or outperformed other models (two-tailed test, $q > 0.05$ FDR corrected). Finally, we fitted the models to a cluster lasting between 891 and 1027 ms in the linguistic imagery condition (Figure 4.2E, 17 time points). Note that the time-resolved predictions were also significant only at the uncorrected level of $p < 0.05$. As was the case with sentence reading, there was no evidence for any of the models to produce a statistically significant fit with the neural geometry (Figure 4.3G, one-tailed t-test, all $p > 0.05$ Bonferroni-corrected for five models). No cosine similarity greater than 0.063 (SEM = 0.034) produced by the weighted similarity model was observed. Although neither of these model comparisons in the linguistic modality performed significantly above zero, they all reached the lower bound of the noise ceiling (Figure 4.3E - Figure 4.3G). This points to the surprisingly low estimates of the noise ceiling where the expected minimum correlation was close to zero across all linguistic conditions and even the late cluster in visual perception (Figure 4.3D). This observation implies that there is a substantial level of measurement noise in the current data, which may have affected the true neural representational structure.

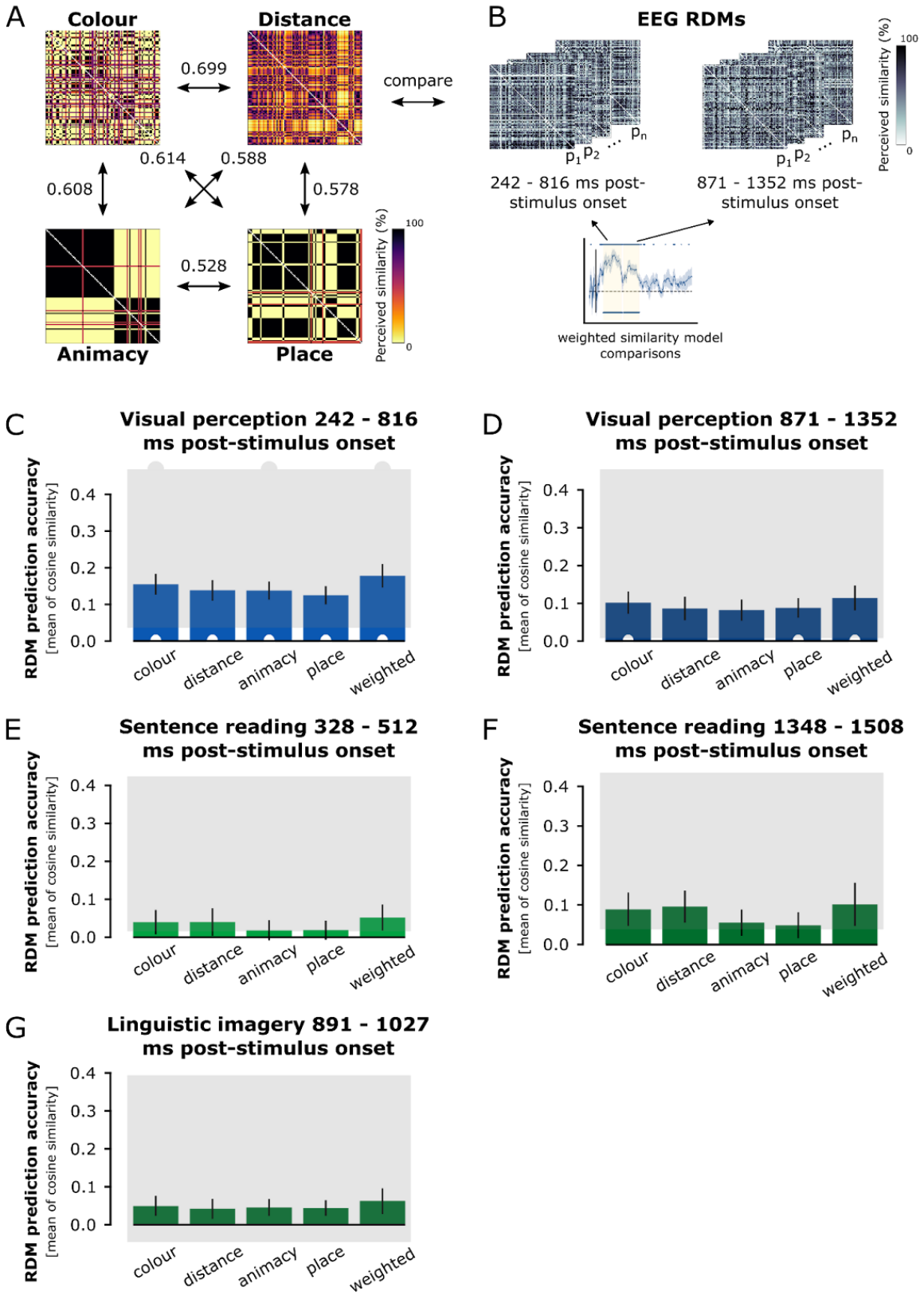


Figure 4.3 | Representation of mid- and high-level features in neural geometry of multimodal perception and imagery. (A) An independent group of participants

provided annotations of the visual stimulus set from which four feature-based models of colour, distance, animacy, and place were constructed. These models were then compared to a spatiotemporal version of EEG RDMs (B), which represented neural activity from clusters significantly predicted by the weighted similarity model. The bar graphs at the bottom show a comparison of feature-based models to the early (C), and late cluster (D) of spatiotemporal EEG RDMs from the visual perception condition. The bars show cosine similarity averaged across bootstrapped samples of participants and stimulus conditions. Error bars indicate estimates of standard errors and white dew drops on the horizontal axis indicate which models perform significantly above zero. The grey areas depict the lower and upper bounds of the noise ceiling with grey dew drops marking models that are significantly different from the lower bound of the noise ceiling. Both early (E) and late (F) clusters in the reading conditions were also compared to the weighted similarity model to visualise their performance relative to feature-based models. (G) depicts model comparisons to neural geometry in linguistic imagery.

4.3.3 Categorical representations: Within-modality decoding across time

Multidimensional scaling (MDS) of the similarity judgements revealed four main categorical clusters of animals, people, indoor, and outdoor scenes (Figure 4.2A). These categories are known to be tied to the architecture of visual representational geometry (Epstein, & Kanwisher, 1998; Haxby et al., 2001; Ishai et al., 1999; Kanwisher et al., 1997) and significant decoding performance in neural dynamics of vision (Carlson et al., 2013; Kaiser et al., 2016a; Kaneshiro et al., 2015), but the granularity of categorical representations in language remain less clear. It was hypothesised that if both vision and language project onto the shared multidimensional space (Chapter 2, Chapter 3), then it would be possible to use the categorical distinctions identified in the MDS (Figure 4.4A) to discriminate between multimodal representations in perception and imagery. For this, we selected 10 exemplars from each of the categorical clusters (Figure 4.4B, see Methods for detail on how the stimuli were selected) and trained a linear

discriminant (LDA) classifier to perform a 4-class classification on the pre-processed epochs. The classifier distinguished between the categories in visual perception 238 ms into the natural scene onset (peak t value $t(5) = 8.708$, $p < 0.05$ cluster-corrected) with maximum 30.6% classification accuracy (1/4 = 25% chance level). However, there was no evidence that any of the remaining conditions carried meaningful categorical representations (Supplementary Figure 4.1). We then resolved to a 2-class classification to better account for the granularity of categorical representations and related features.

4.3.3.1 Visual perception and imagery

In visual perception, the binary LDA classifier reliably discriminated between categorical representations of *animals* versus *people* (Figure 4.4C), *animals* versus *outdoor scenes* (Figure 4.4D), *people* versus *outdoor scenes* (Figure 4.4F), and *people* versus *indoor scenes* (Figure 4.4G). These results indicate that the neural dynamics in vision contains sufficient information to allow decoding of many animate versus scene representations. Categorical representations of animals versus people, animals versus outdoor scenes, and people versus outdoor scenes were discriminable as early as 238 ms after the image appeared on the screen. This early onset was followed by a sharp increase in above-chance classification scores reaching a peak of 0.600 at 389 ms on average. The highest classification score occurred when discriminating between *people* versus *outdoor scenes* (Figure 4.4F) with classification scores reaching a maximum 320 ms after the image onset (max AUC = 0.628, peak t value $t(5) = 107.454$, $p < 0.05$ cluster-corrected). This peak was then followed by a sharp decline with the first cluster of above-chance classification scores ending by 600 ms. Another sustained cluster of above-chance decoding occurred from 871 to 953 ms and 1457 to 2606 ms. These findings validate categorical distinctions found in the MDS where representations of *people* and *outdoor scenes* are represented at the opposite

ends of the multidimensional space (Figure 4.4A). Opposedly, a late cluster of above-chance classification scores between 2355 and 2477 ms was produced when discriminating between *people* versus *indoor scenes* (Figure 4.4G; max AUC = 0.583, peak t value $t(5) = 40.317$, $p < 0.05$ cluster-corrected). This cluster was preceded by above-chance classification scores significant only at the uncorrected level of $p < 0.05$ starting 226 ms after the image onset. Similar findings were found when discriminating between representations of *animals* versus *indoor scenes*. As shown in Figure 4.4E, there was a sharp increase in above-chance classification scores starting around 461 ms and lasting until 1395, but neither of these scores survived cluster-based correction. There was no evidence that the classifier reliably discriminated between representations of outdoor versus indoor scenes either, likely due to them sharing many overlapping features.

Moving onto categorical representations in visual imagery, we observed a sparse distribution of above-chance decoding when discriminating between *animals* versus *people* (Figure 4.4C) and *animals* versus *outdoor scenes* (Figure 4.4D). The former yielded above-chance decoding 496 ms after the imagery onset with a peak classification score of 0.580 at 563 ms. The latter revealed sporadic above-chance scores between 27 and 1438 ms after the imagery onset (all $p < 0.05$ uncorrected). However, neither of these classification scores survived cluster-level correction. There was no evidence for a reliably discriminable representation of animals versus indoor scenes, people versus outdoor/indoor scenes, and outdoor versus indoor scenes.

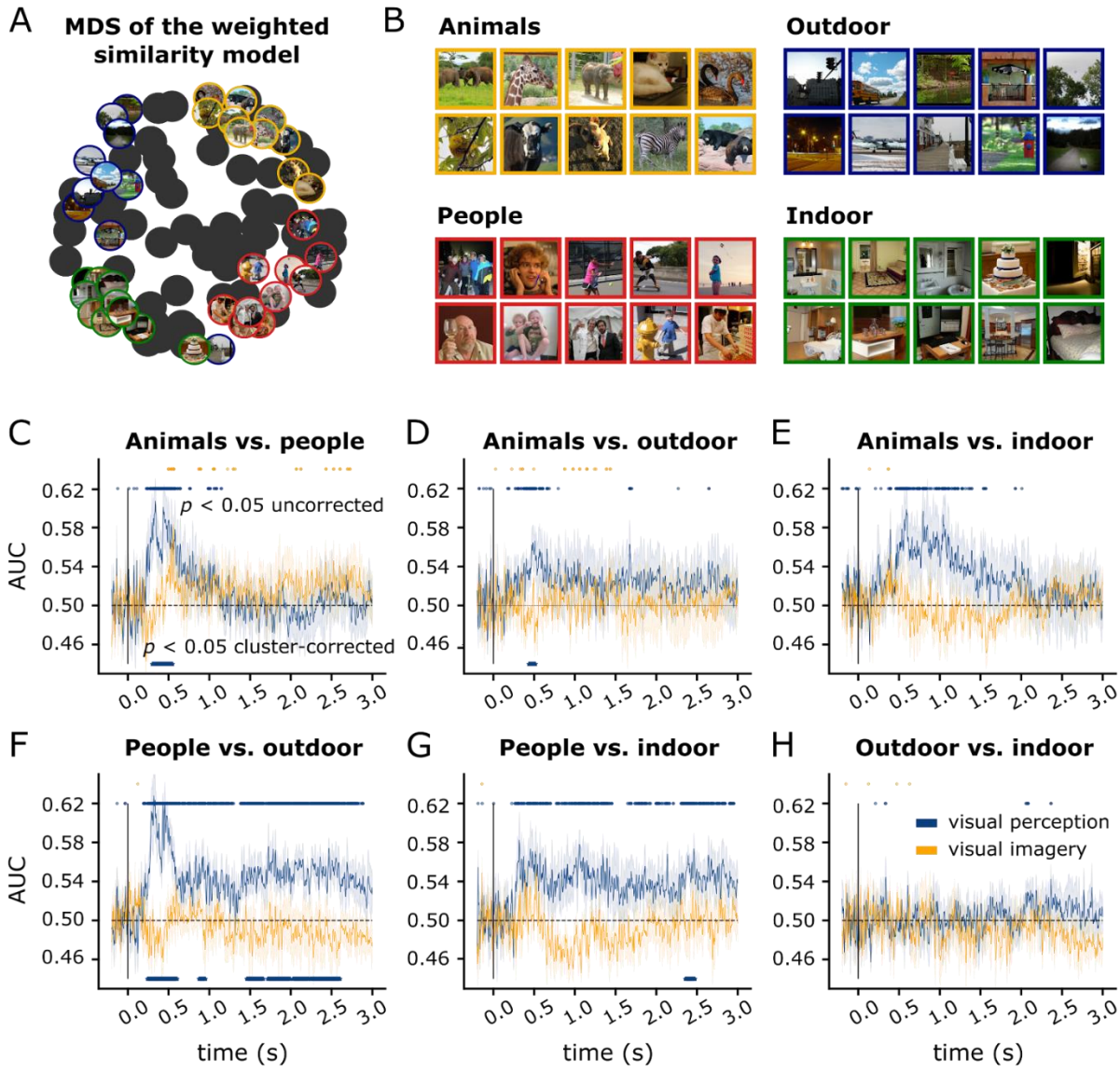


Figure 4.4 | Categorical representations in visual perception and imagery. (A) Multidimensional scaling of the weighted similarity model revealed four main categorical clusters. (B) We selected a balanced set of 10 exemplars per category. These were selected such that each scene is representative of the category (see Methods for detail). The bottom panel depicts results from a binary LDA performed on the pre-processed EEG epochs in visual modality discriminating between animals and people (C), animals and outdoor scenes (D), animals and indoor scenes (E), people and outdoor scenes (F), people and indoor scenes (G), outdoor and indoor scenes (H). For each participant, classification scores were averaged across cross-validation folds. The shaded area depicts the SEM across participants. The significance of classification scores across participants was determined by a one-sided t-test against chance (1/2 = 0.5 chance level, $p < 0.05$ cluster-

corrected). Stars in solid colour at the bottom of each graph mark significant clusters. Light-coloured points at the top mark significant classification scores at the uncorrected $p < 0.05$ level.

4.3.3.2 Sentence reading and linguistic imagery

Next, we investigated whether we can identify categorical representations in neural activity induced in sentence reading and linguistic imagery. In sentence reading, the classifier produced isolated time points that exhibited above-chance classification scores only at the uncorrected $p < 0.05$ level. Overall, there was no evidence for reliable discrimination of any of the categorical representations. Interestingly, the classifier generally performed better in discriminating categorical representations in linguistic imagery. We observed sparsely populated clusters of above-chance classification scores at the uncorrected $p < 0.05$ level across all categorical representations. The most sustained clusters were identified in categorical representations of *animals* versus *outdoor scenes* (max AUC = 0.558), *animals* versus *indoor scenes* (max AUC = 0.561), *people* versus *outdoor scenes* (max AUC = 0.554), and *outdoor* versus *indoor scenes* (max AUC = 0.551). These above-chance scores were largely produced early in the epoch starting as early as 98 ms before the imagery onset and lasted until 1755 ms on average. No classification score higher than 0.558 was observed. Neither of these scores, however, remained after cluster-based correction.

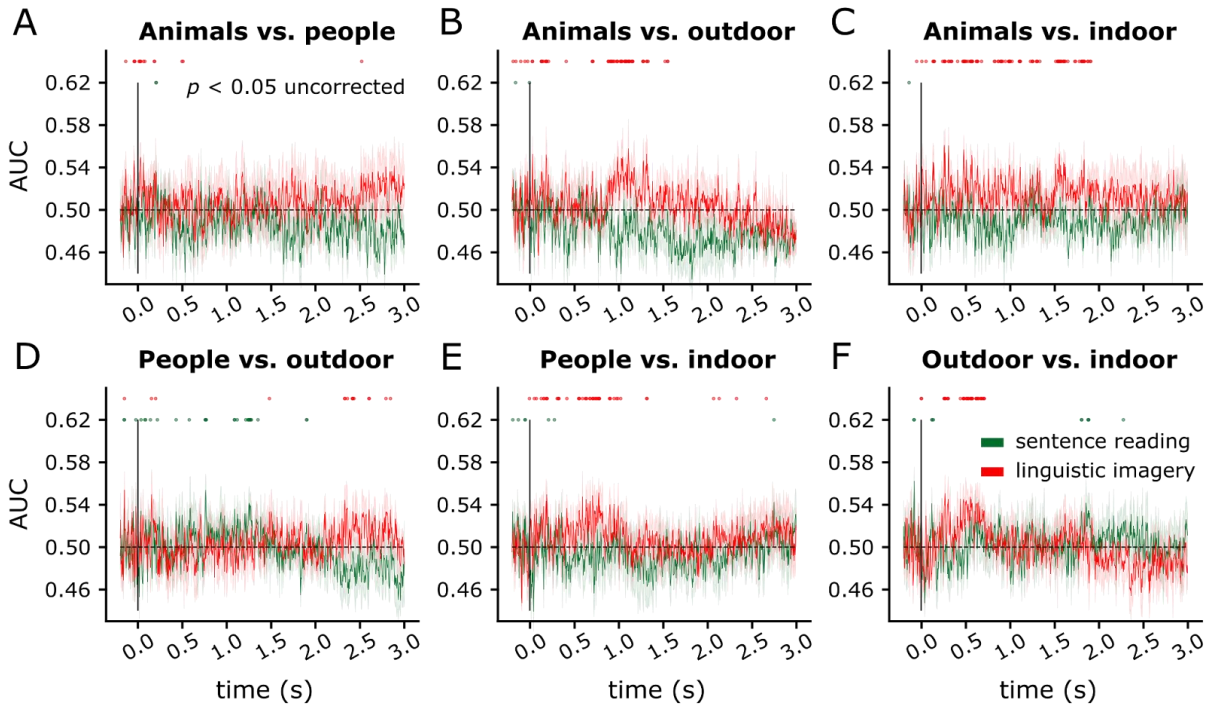


Figure 4.5 | Categorical representations in sentence reading and linguistic imagery. The figure depicts averaged classification scores across participants from a binary LDA classifier that discriminated between animals and people (A), animals and outdoor scenes (B), animals and indoor scenes (C), people and outdoor scenes (D), people and indoor scenes (E), outdoor and indoor scenes (F). The shaded area depicts the SEM across participants. The significance of classification scores was determined by a one-sided t-test against chance (1/2 = 0.5 chance level, $p < 0.05$ cluster-corrected). Light-coloured points at the top mark significant classification scores at the uncorrected $p < 0.05$ level.

4.3.4 Categorical representations: generalisation across time and modalities

Research shows that language can modulate behavioural performance in visual tasks suggesting that there is a point at which representations from multimodal systems converge and translate into behaviour (Chapter 3; Stanfield, & Zwaan, 2001; Zwaan et al., 2002). Here, we perform cross-decoding across time and modalities to characterise such a representational overlap in real time. Since the most reliable decoding performance was found in visual perception, we first

explored whether these features generalise and potentially help decoding in the remaining experimental conditions. For this, we trained a binary LDA classifier on all six pairwise categorical representations on epochs in visual perception and tested the classifier on epochs in visual imagery, sentence reading, and linguistic imagery. The temporal generalisation revealed that the early categorical representation of *people* and *outdoor scenes* in visual perception was similar to the same categorical representation in linguistic imagery (Figure 4.6A; max AUC = 0.575, peak t value $t(5) = 47.116$, $p < 0.05$ cluster-corrected). This overlap was evident from 1000 ms and lasted until the end of the linguistic imagery epoch. There was no significant representational overlap between the other categorical representations trained on visual perception and tested on any of the remaining experimental conditions (Supplementary Figure 4.2A-C, all $p > 0.05$ cluster-corrected).

We repeated the same procedure to generalise categorical representations from the sentence reading condition. We found that the representation of *people* and *outdoor scenes* generalised to both perception and imagery from the visual modality. We observed a significant overlap between the late-stage sentence reading and almost the entire epoch in visual perception (Figure 4.6B; max AUC = 0.565, peak t value $t(5) = 35.656$, $p < 0.05$ cluster-corrected). The late linguistic representation of *people* and *outdoor scenes* also overlapped with early representation in visual imagery until approximately 2500 ms into the imagery onset (Figure 4.6C; max AUC = 0.567, peak t value $t(5) = 28.147$, $p < 0.05$ cluster-corrected). These findings suggest that there is a degree of overlap between vision and language particularly noticeable in the representation of *people* and *outdoor scenes*, likely related to the internal geometry of those categories (Figure 4.4A). However, there was no evidence that any of the remaining categorical representations trained on sentence reading generalised to any other modalities (Supplementary Figure 4.2D-F, all $p > 0.05$ cluster-corrected). We also did not

observe any significant periods of overlap when we trained the classifier on the imagery conditions (Supplementary Figure 4.3, all $p > 0.05$ cluster-corrected).

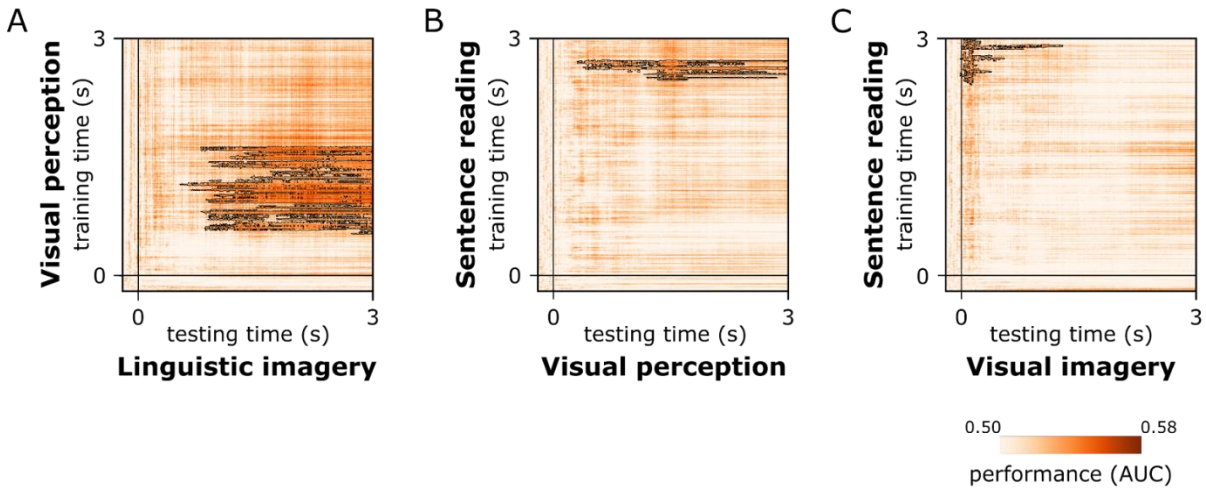


Figure 4.6 | Generalisation of categorical representation for people and outdoor scenes across time and modalities. A binary LDA classifier was trained and tested to explore how all six pairwise categorical representations generalise across time and modalities. We found significant overlap when training on visual perception and testing on linguistic imagery the representation of *people* and *outdoor scenes* (A). A significant overlap was also found when training on sentence reading and testing on both visual perception (B) and imagery (C) for the same categories. No periods of overlap were found for any of the remaining conditions (see Supplementary Figure 4.2 and Supplementary Figure 4.3). The performance was thresholded at chance level (1/2 = 0.5) and tested for significance using a cluster-based permutation test at the $p < 0.05$ level.

4.4 Discussion

Here, we collected EEG data to investigate how the representational overlap in vision and language unfolds in real time. We showed that the multidimensional space derived from re-weighting behavioural similarity judgements of vision and language reliably predicted representational dynamics of visual perception. These results are consistent with previous research that showed alignment between

perceived dissimilarities and both neural representational structure (Charest et al., 2014; Magri, & Konkle, 2019; Mur et al., 2013) and dynamics (Cichy et al., 2019; Wardle et al., 2016) of vision. However, the onset of significant predictions appeared far later than those observed by Cichy and colleagues (2019) or Wardle and colleagues (2016) who found that the geometry of perceived dissimilarities and human vision aligned by 100 ms after the image onset. This discrepancy may be because both studies used simplified stimuli such as visual gratings and isolated objects. The later onset in our study may reflect the time required for the human brain to engage in higher-order visual processing including semantic interpretation (Greene, & Oliva, 2009; Kaiser et al., 2016b) and contextual analysis (Bonner, & Epstein, 2021; Võ et al., 2019; Stansbury et al., 2013), which may have not been as prominent in the aforementioned research. This interpretation is consistent with Proverbio and Riva (2009) who showed that semantic congruence when viewing natural scene images is reflected in the recognition potential that emerges between 200 - 500 ms post-stimulus.

We also observed some alignment between the multidimensional similarity space and neural dynamics in sentence reading with the first and second peaks appearing 328 and 1414 ms into the reading onset. The early peak aligns with the N400 potential typically linked to early semantic integration (Hagoort et al., 2004; Kutas, & Hillyard, 1980; van Berkum et al., 1999). The later peak likely reflects contextual update (van Petten, & Luka, 2012) where participants adjust the mental representations based on their understanding of the sentence's meaning. These findings are in agreement with prior research suggesting that language processing involves a distributed network of semantic representations (Anderson et al., 2019; Huth et al., 2016; Popham et al., 2021; Tang et al., 2021). We expand on this by showing that the brain evaluates semantic similarity and associations across different sentence processing stages. However, a note of caution is due here because the predictions failed to survive cluster-based correction. This may

be attributed to inconsistency in reading speed where research shows that covert reading and speech production are subject to substantial temporal variations (Martin et al., 2016; Rayner et al., 2006). Because the sentence was presented in a static display, the inconsistency in reading speed perhaps caused a misaligned geometry across participants. This then led the weighted similarity model to underestimate the predictions in real-time. Future research should employ a rapid serial visual presentation (RSVP) design to better time lock the neural response to the sentence enabling further investigation of how different sentence constituents contribute to the neural geometry.

We then performed model comparisons to investigate what information is represented in the neural geometry. We observed significant performance for all models in the early cluster of visual perception, and models of colour, place, and weighted similarity in the late cluster. These models performed above the lower bound of the noise ceiling, indicating that the models capture meaningful patterns in the data. However, contrary to our expectation that more complex models would outperform those based on isolated features (Anderson et al., 2019; Doerig et al., 2024; Tang et al., 2021; Wang et al., 2023), we found that the fixed feature-based models performed just as well as the weighted similarity model. This could be attributed to the high correlation between the models (Figure 4.3A), possibly reflecting the natural co-occurrence and hierarchy between the features. An example of this may be that objects that are similar in appearance also share many semantic features, often making such features not entirely separable (Bracci, & Op de Beeck, 2016; Bracci et al., 2017; Martin et al., 2018). It seems therefore possible that the visual brain is tuned to integrate high-dimensional features into a unified representational geometry. However, with a small sample size, an alternative explanation may be that the analysis did not have enough power to reveal the true difference in model performance.

This is supported by further assessment of the noise ceiling where we found that the expected minimum correlation was close to zero, which implies low alignment of neural geometries across participants. Moreover, the expected maximum correlation fell far below the perfect correlation, indicating that even the best achievable model would perform relatively low on the current data (Nili et al., 2014). It is, therefore, possible that the model comparisons were somewhat obscured by high measurement noise (Nili et al., 2014; Walther et al., 2016), which was most prominent in the linguistic conditions as depicted by the wide noise ceiling range (Figure 4.3). It is important to note that we took several precautions to minimise multivariate noise in the data, which included cross-validating the EEG RDMs across runs (Diedrichsen et al., 2020; Walther et al., 2016), pooling across time points to increase the signal-to-noise ratio (Naselaris et al., 2015) and using bootstrapping method across stimuli and participants to avoid overfitting (Schütt et al., 2023). On top of that, the wide noise ceiling range appeared even in the early cluster of visual perception where it would be expected to observe better alignment between participants since the neural response was driven by an external stimulus. These observations point to the sample size to be the most likely cause of the high measurement noise, which could be mitigated in the future.

We found no evidence that visual imagery aligned with the multidimensional similarity space, nor that its neural dynamics represented any meaningful category information, which contradicts previous research that successfully identified semantic representations in imagery (Dijkstra et al., 2018; Mitchell, & Cusack, 2016; Stecher, & Kaiser, 2023). However, this discrepancy is not an isolated case, as several studies have also reported limited performance in imagery decoding (Kosmyna et al., 2018; Mitchell, & Cusack, 2016; Lee et al., 2022; Shatek et al., 2019). Several factors may have affected our ability to decode imagery representations, including the complexity of the natural scene stimuli.

Dijkstra (2024) recently proposed a link between the quality of mental imagery and high sensitivity to size and orientation in the visual cortex. Providing that imagery activates representations in the visual cortex (Albers et al., 2013; Dijkstra et al., 2017; Dijkstra et al., 2019; Ganis, & Schendan, 2008; Harrison, & Tong, 2009; Horikawa, & Kamitani, 2017; Klein et al., 2004; Lee et al., 2012), it is possible that the mental image was highly dispersed across participants since they were not instructed about what exactly they should imagine. This is even more relevant since the stimulus set consisted of natural scenes that depicted multiple objects of interest, which enabled participants to imagine various elements of the scene. To address these challenges, future research should consider diversifying the stimuli by including more exemplars from within related categories – this approach could strengthen the neural representation of corresponding features and improve decoding performance.

It may also be that the mental image was propagated in a step-down fashion where general representation was re-instantiated first (Dijkstra et al., 2020) followed by mental exploration of isolated objects (Lewis et al., 2011; Kosslyn, 1975; Kosslyn et al., 1978; Shepard, & Metzler, 1971). It is therefore possible that the fragmented representation – likely dispersed across the imagery epoch, lacked sufficient power to align with the complex multidimensional similarity space or carry a stable categorical representation. This inconsistency may be exemplified by substantial variability in the imagery onset (Dijkstra et al., 2018) and personal imagery style (Bergmann et al., 2016; Bigelow et al., 2023; Cui et al., 2007). One way to mitigate the differential time-course would be to employ time-frequency over time-resolved analysis, which has proven efficient in identifying imagery representations in the alpha frequency band (Stecher, & Kaiser, 2023; Xie et al., 2020).

Surprisingly, linguistic imagery displayed more densely populated clusters of significant categorical decoding compared to visual imagery, and even a late alignment with the weighted similarity model. Although the results were only significant at the uncorrected level, language may induce stronger mental imagery for several reasons. First, reading involves dual coding (Paivio, 1971; Paivio, 1986) where a sentence is processed at the syntactic level at first (Friederici et al., 2000; Friederici et al., 2003; Pallier et al., 2011), followed by linking its semantic meaning to an appropriate mental image (Barsalou, 1999; Dils, & Boroditsky, 2010; Liu et al., 2024; Stanfield, & Zwaan, 2001; Zwaan et al., 2002). Such higher demands in cognitive effort may have strengthened the imagery as opposed to visual imagery, which is preceded by passive image viewing. We speculate that the methodological choice to include a visual mask after the perception phase may have even reinforced this effect. The mask may have disrupted the immediate visual representation, forcing participants to rely on recall of the natural scene image. The visual imagery was then more likely to be imprecise because of non-commitment, which is a pervasive feature in visual imagery (Bigelow et al., 2023). This contrasts with the linguistic condition, where sentences are typically structured to point to the main subject of interest in the scene (e.g., a yellow school bus driving down the country road"), which may have provided a more exact instruction of what to imagine.

This study set out to explore how the visuo-linguistic representational overlap unfolds in real-time. The results showed that visual perception reliably aligns with the multidimensional space obtained in behavioural similarity judgements (Chapter 2), and captures sufficient information for category representations to be identified. We also observed some potential in using this multidimensional space to explain neural geometry in language. These findings support the concept of distributed semantic coding (Federmeier, & Kutas, 1999; Haxby et al., 2001; Huth et al., 2016; Ishai et al., 1999; Kiani et al., 2007; Newman

et al., 2010; Tang et al., 2021; Wang et al., 2023), where both vision and language appear to rely on their inherent ability to encode the similarity of stimuli. However, the generalisability of these results is subject to limitations due to high measurement noise and failure to observe significant results after applying correction methods. With the current data, we also cannot conclusively determine the extent to which neural representations in vision and language converge onto a shared representational space. Although we observed some significant periods in which categorical representations generalised across the modalities, this was limited only to scenes depicting people and outdoor scenes. This may be related to the internal geometry of the multidimensional space along with diagnostic features of each of these categories, however, the lack of reliable generalisation between the remaining conditions invites a further discussion as to what extent vision and language truly align (Conwell et al., 2023) and what is the precise nature of such an overlap. As seen in prior research (Kosmyna et al., 2018; Lee et al., 2022; Mitchell, & Cusack, 2016; Shatek et al., 2019), mental imagery may still be a challenging phenomenon to study in real-time. Further research should therefore be carried out to mitigate some of the methodological concerns to capture more precise representations in vision and language. This includes diversifying the dataset by introducing different stimuli, but similar in content to obtain more fine-grained sampling of the neural geometry, along with increasing the sample size. The study would also benefit from an RSVP design in the reading condition to better time-lock events related to various language processing stages (Friederici et al., 2000; Friederici et al., 2003; Pallier et al., 2011) and how these generalise to visual feature coding. Finally, future research should also consider prioritising time frequency over time-resolved analysis to further investigate how multimodal systems converge in specific frequency bands. Once these issues are addressed, a natural progression of this work would be to perform a sensor-based analysis to further explore how the visuo-linguistic overlap is distributed across the brain.

4.5 Materials and methods

4.5.1 Participants

Seven participants recruited from the University of Birmingham (mean 20±2 years, five female, one non-binary, one left-handed, one mixed) participated in the experiment in return for financial compensation of £8/h. Eligibility criteria required participants to have normal or corrected-to-normal vision and not suffer from psychiatric or neurological disorders. Participants provided informed consent as part of a protocol approved by the University of Birmingham Research Ethics Committee.

4.5.2 Stimuli

The stimulus set was obtained from the Natural Scenes Dataset (NSD, Allen et al., 2022) and consisted of 100 natural scene images and 100 sentence captions describing the images. A greedy algorithm (Allen et al., 2022) selected the set of 100 images such that they maximally span the high-level semantic space, which was defined by embeddings of the images' sentence captions. The benefit of this approach was that the stimulus set depicted a wide range of real-world scenarios allowing us to obtain a comprehensive sampling of the neural representational geometry in the EEG experiment. The sentence captions were collected from five human annotators as part of the Microsoft Common Objects in Context database (MS COCO, Lin et al., 2014). For the purpose of our experiment, we identified the most frequent words and phrases across the five annotators and combined these to generate one representative sentence caption for every image. The stimulus set was used in two experiments in which participants passively viewed a stimulus and then performed mental imagery of that stimulus either in its visual or linguistic format while we recorded EEG.

The experiment was programmed in PsychoPy3 (v2022.1.2, Peirce et al., 2019) library in Python. Stimuli were displayed against a black background on an Asus PC screen with 1920 x 1080 resolution and 60 Hz refresh rate. Participants were seated approximately 50 cm from the screen. Images in the original size of 425 x 425 pixels were displayed at 10 degrees of visual angle and sentences were displayed at 1 degree of visual angle in white Lucida Console font. Every stimulus presentation in the perception phase was followed by the display of a static mask. This was done to reduce variability in the activation of mental imagery across participants (Dijkstra et al., 2018; Xie et al., 2020) by (1) removing the immediate representation of the stimulus from working memory (Fahrenfort et al., 2007) and (2) time-locking feedback processing of the stimulus at the beginning of the imagery phase (Linde-Domingo et al., 2019). The post-perception masks in the visual modality were created by randomly selecting 289 25 x 25 pixel quadrants from any of the 100 images. These were then assembled into a new image in the size of the NSD stimuli (425 x 425 pixels). A new visual mask was presented in every trial. The post-perception linguistic mask preserved the format of the sentence displayed in the perception phase (i.e., the number of letters in every word and their order) but letters were randomly shuffled. A red fixation point centred on the screen was displayed at 0.2 degrees of visual angle in the imagery phase, and between trials. All responses to behavioural tasks were made using a computer mouse with the right hand.

4.5.3 Procedure

Participants attended eight EEG sessions in total. In half of the sessions, we administered the visual modality of the EEG experiment and in the other half, we administered the linguistic modality of the EEG experiment. The order of the modalities was counterbalanced across participants such that four participants started with the visual modality in their first four EEG sessions followed by the

linguistic modality in the remaining four EEG sessions. Three participants completed the experiment in reversed order enabling us to capture the linguistic representations without prior exposure to the visual stimuli. On average, participants completed 3-4 runs per session each presenting the full stimulus set and lasting approximately 25 minutes. Each session was scheduled on different days and lasted approximately 3.5 hours. We acquired a total of 84 repetitions per image from the visual stimulus set and 94 repetitions per sentence from the linguistic stimulus set amounting to a total of ~90 hours of EEG recordings.

Before starting the EEG experiment in the first session, we administered the Vividness of Visual Imagery Questionnaire (VVIQ, Marks, 1973) to assess participants' ability to form mental images. Participants rated the vividness of 16 imagined scenarios on a scale between 1 (no image at all) and 5 (as vivid as normal vision) once with eyes open and once with eyes closed. The average VVIQ score for eyes open was 58.5 (SD = 12.1) and for eyes closed was 69 (SD = 10.6). Neither of the participants needed to be excluded due to low VVIQ scores. The EEG design is depicted in Figure 4.1A and Figure 4.1B. Trials started with a presentation of a red fixation dot for 2 seconds (+ random jitter between 0 - 0.250 s) followed by the perception phase displaying a stimulus for 3.5 seconds. The stimuli were presented in random order in every run. The perception phase was followed by a static post-perception mask for 0.5 seconds, and another fixation dot for 2 seconds. The beginning of the imagery window was indicated by a 1000 Hz sound cue lasting 100 ms and the display of a rectangle with a white frame centred at the screen. Participants were instructed to imagine the stimulus on the first sound cue and maintain the imagery until the second sound cue indicating the end of the 3-second imagery phase. Then, a 0.5-second fixation appeared followed by a vividness rating where participants were asked to rate the vividness of the imagery using a slider on a scale between 1 (no image at all) and 5 (as vivid as normal vision). Every trial was followed by an attention check to ensure

participants were attentive to stimuli presented in the perception phase. Participants were shown either a 25 x 25 pixel quadrant selected randomly from an image in the visual modality or a randomly selected word in the linguistic modality. Participants were asked to decide whether the displayed stimulus matched the stimulus from the perception phase. The stimuli matched 50% of the trials.

4.5.4 EEG recording and preprocessing

EEG data were recorded at 2048 Hz using a 128-channel BioSemi EEG system. Additional electrodes were placed at the left and right mastoids, and the superior/inferior orbital rim to monitor eye movements. The impedances were kept below 10 kΩ. EEG preprocessing was performed using the MNE-Python package (Gramfort et al., 2013). The EEG signal was filtered at a bandpass filter of 0.05 - 40 Hz and re-referenced to an average reference. We opted to average re-referencing due to the lower quality of the mastoids' signal in some participants. The EEG signal was downsampled to 256 Hz to reduce the computational load. Epochs were segmented between -0.2 seconds and 3 seconds relative to stimulus/imagery onset. Baseline correction was applied 0.2 seconds pre-stimulus/imagery. Four types of epochs were created including visual perception, sentence reading, visual imagery, and linguistic imagery.

4.5.5 Feature annotations of NSD stimuli

Feature-based annotations of the 100 natural scene images were collected from a naive sample of two participants recruited by word-of-mouth. An online questionnaire was administered online via the web-based platform Meadows (https://meadows-research.com/) and took approximately 30 minutes to complete. Participants viewed one image at a time and were cued to annotate the

image according to colour, distance, animacy, and place. The first forced-choice question asked participants to decide whether red, green, or blue was the most prominent colour in the scene. Second, participants used a slider to indicate how far away the main object in the scene was. For instance, the slider was placed all the way to the left if the scene depicted an object photographed at a close range (e.g., a close-up of a cat), but all the way to the right if the scene was open and/or the dominant object was depicted in the distance (e.g., a mountain on the horizon). Third, participants indicated whether the scene was animate and lastly, participants decided whether the scene was situated indoors or outdoors in two forced-choice questions. Participants were instructed to make their decisions to the best of their judgement and to consider the most prominent object in the scene.

4.5.6 Representational similarity analysis

4.5.6.1 Construction of time-resolved EEG data RDMs

We used the cross-validated Mahalanobis distance from the rsatoolbox library (version 0.1.3.dev44, van den Bosch et al., in prep, https://github.com/rsagroup/rsatoolbox/tree/main) in the Python environment to compute EEG representational dissimilarity matrices (RDMs). The benefit of the cross-validated Mahalanobis distance is that it corrects multivariate noise by validating the difference in dissimilarity estimates across measurements (Walther et al., 2016). We loaded a participant's EEG epochs across all runs for a given experimental condition. To compute time-resolved EEG RDMs, neural measurements across the epochs were extracted one time point at a time leaving one measurement per EEG channel. The measurements were then split by run into a train and test set where all but one run were assigned to the train set. We then averaged the measurements by stimulus condition (i.e., NSD stimulus type) across the train runs and used the Mahalanobis distance to compute dissimilarity

estimates between measurement vectors for each stimulus condition in the train and test sets. The dissimilarity estimates were then averaged across all possible cross-validation folds and stored in one EEG RDM for a given participant and experimental procedure comprising one dissimilarity array per time point.

4.5.6.2 Weighted similarity model compared to neural representational dynamics

We created a weighted similarity model based on similarity judgements data collected in Chapter 2. Briefly, two independent groups of participants completed the multiple arrangements (MA, Kriegeskorte, & Mur, 2012) task on the visual and linguistic stimuli used in the EEG experiment. The dissimilarities between each pair of stimuli in the MA task were computed using the Euclidean distance and assembled in one RDM per participant. The participant RDMs were then averaged for each task modality, which resulted in two behavioural RDMs each representing a similarity structure either from the visual or linguistic similarity judgements. The weighted similarity model was computed as a weighted sum of the visual and linguistic RDMs. This approach was chosen because the relative weighting of features in each of the behavioural RDMs was not known a priori (Schütt et al., 2023), and was likely mixed in the neural measurements. The weighted similarity model was then compared to time-resolved EEG RDMs for each participant using cosine similarity. Similarity estimates were then tested against zero using a one-sided cluster-based permutation test from the MNE-Python library (Gramfort et al., 2013).

4.5.6.3 Construction of spatiotemporal EEG data RDMs

Clusters of significant similarity estimates between the weighted similarity model and time-resolved EEG RDMs were used to create a spatiotemporal version of EEG RDMs (Figure 4.3B). If similarity estimates were significant only at the uncorrected $p < 0.05$, we identified consecutive time steps where significant

predictions were no more than 100 ms apart. We then extracted neural measurements from the pre-specified time windows, vectorised the measurements across the time and channels, and computed new EEG RDMs using the cross-validation procedure as described above. We chose this approach to increase signal-to-noise ratio of the EEG RDMs (Diedrichsen et al., 2020; Naselaris et al., 2015), and quantify relative weighting of mid- and high-level features that were disregarded by the weighted similarity model.

4.5.6.4 Construction of feature-based models

To determine the relative weighting of mid- and high-level features, we constructed four feature-based fixed models based on the annotations of the visual stimuli. The annotations from the forced-choice questions (colour, animacy, place) were converted into a numerical format using one-hot encoding. For example, if a participant indicated red as the dominant colour in the scene, the stimulus was assigned a label for the red channel, and 0 for the blue and the green channel. The encoded labels were then averaged across the annotators and pairwise Euclidean distances were calculated between every pair of stimuli. The distance values from the slider were linearly scaled between the minimal (0) and maximal (1) values before the pairwise Euclidean distances were calculated. This procedure resulted in four RDMs each representing features of colour, distance, animacy, or place (Figure 4.3A). We then used these RDMs to create four feature-based fixed models using the rsatoolbox.

4.5.6.5 Weighted and fixed models evaluated against spatiotemporal neural representation

The weighted similarity and fixed feature-based models were evaluated on the spatiotemporal EEG RDMs to investigate the relative weighting of mid- and high-level features in the neural representational geometry. We resolved to 2-fold

participant and 5-fold stimulus cross-validation across 1000 bootstrap resamples using the eval_dual_bootstrap function from the rsatoolbox (version 0.1.3.dev44, van den Bosch et al., in prep). A major advantage of this method is that (1) the model similarity estimates are not biassed by overfitting to either subject or conditions (Schütt et al., 2023), and (2) it helps to generalise from the sampled participants and stimuli to the population of participants and stimuli (Nili et al., 2014). The test subset of fixed feature-based models was evaluated on a test subset of EEG RDMs. Before the weighted similarity model was evaluated on the test set of EEG RDMs, it was fitted to a training subset of EEG RDMs where each MA RDM was assigned flexible theta parameters. The weighted similarity model was then predicted using these weights and its subset was evaluated on the test subset of EEG RDMs using cosine similarity. At the same time, the upper and lower bounds of the noise ceiling were calculated to determine the expected model estimates given the noise in the EEG RDMs. To calculate the upper bound of the noise ceiling, the fit of the true model's RDM is overestimated by correlating the test subset of EEG RDMs to the full pool of test EEG RDMs. The lower bound is underestimated by correlating the test subset of EEG RDMs with the training subset of EEG RDMs. The estimates were averaged across all possible folds resulting in one estimate for upper and lower bound per bootstrap. A one-sided test was used to determine the significance of the estimates against zero (a two-tailed test was used for comparisons to the lower bound of the noise ceiling) while correcting for multiple comparisons for five models at the level of $q < 0.05$ Bonferroni-corrected. Pairwise comparisons were performed using a two-sided test ($p < 0.05$ FDR corrected).

4.5.7 Linear discriminant analysis

Linear discriminant (LDA) classifier from the MNE-Python library (version, Gramfort et al., 2013) was trained to discriminate between categorical

representations in neural activity patterns. We defined four categories consisting of people, animals, indoor, and outdoor scenes based on the categorical clusters visualised in the multidimensional scaling (MDS) of the similarity judgements (Figure 4.2A). 10 stimuli from each categorical cluster were selected and associated with each category. This was done to ensure a balanced sample across categories, and that the stimuli are representative of the category (Figure 4.4B). For example, in both animate categories, we selected stimuli in which the animate object was depicted at the centre of the image and was also described in the sentence caption of that image as the main subject. Similarly, indoor scenes were selected such that the image clearly depicted/sentence caption clearly described an interior without referring to any animate objects. Outdoor scenes were open-plan scenes depicting/describing natural artefacts.

4.5.7.1 Within-modality decoding across time

Categorical representations were discriminated in a binary LDA classification across time on the pre-processed EEG epochs within a participant. EEG data features were standardised by removing the mean from each data point and scaling to unit variance. For each participant, the LDA classifier (scikit-learn version, Pedregosa et al., 2011) was trained and tested for each time point using leave-one-run-out cross-validation from the MNE-Python library (Gramfort et al., 2013). Classification scores were then averaged across the folds and tested across participants against chance using a one-sided cluster-based permutation test from the MNE-Python library at the level of $p < 0.05$.

4.5.7.2 Generalisation across time and modalities

To investigate shared representations in real-time, we performed cross-decoding of categorical representations identified in the MDS. The LDA classifier was trained for each time point on epochs from a given experimental condition

and tested for each time point on epochs from left-out experimental conditions. The significance of classification scores was determined using the spatial one-sided cluster-based permutation test from the MNE-Python library at the level of $p < 0.05$.

4.6 References

Albers, A. M., Kok, P., Toni, I., Dijkerman, H. C., & de Lange, F. P. (2013). Shared representations for working memory and mental imagery in early visual cortex. *Current Biology, 23*(15), 1427-1431.

Allen, J. E., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., Hutchinson, J. B., Naselaris, T., & Kay, K. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience, 25*, 116-126.

Anderson, A. J., Binder, J. R., Fernandino, L., Humphries, C. J., Conant, L. L., Raizada, R. D. S., Lin, F., & Lalor, E. C. (2019). An integrated neural decoder of linguistic and experiential meaning. *The Journal of Neuroscience, 39*(45), 8969-8987.

Barsalou, L. W. (1999). Perceptual symbol systems. *The Behavioural and Brain Sciences, 22*(4), 577-609.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59*, 617-645.

Bergen, B. K., Lindsay, S., Matlock, T., & Narayanan, S. (2007). Spatial and linguistic aspects of visual imagery in sentence comprehension. *Cognitive Science, 31*(5), 733-764.

Bergmann, J., Genc, E., Kohler, A., Singer, W., & Pearson, J. (2016). Smaller primary cortex is associated with stronger, but less precise mental imagery. *Cerebral Cortex, 26*(9), 3838-3850.

Bigelow, E., J., McCoy, J. P., & Ullman, T. D. (2023). Non-commitment in mental imagery. *Cognition, 238*.

Bonner, M. F., & Epstein, R. A. (2021). Object representations in the human brain reflect the co-occurrence statistics of vision and language. *Nature Communications, 12*(1).

Bracci, S., & Op de Beeck, H. (2016). Dissociations and associations between shape and category representations in the two visual pathways. *The Journal of Neuroscience, 36*(2), 432-444.

Bracci, S., Ritchie, J. B., & Op de Beeck, H. (2017). On the partnership between neural representations of object categories and visual features in the ventral visual pathway. *Neuropsychologia, 105*, 153-164.

Carlson, T., Tovar, D. A., Alink, A., & Kriegeskorte, N. (2013). Representational dynamics of object vision: the first 1000 ms. *Journal of Vision, 13*(10), 1-1.

Charest, I., Kievit, R. A., Schmitz, T. W., Deca, D., & Kriegeskorte, N. (2014). Unique semantic space in the brain of each beholder predicts perceived similarity. *Proceedings of the National Academy of Sciences of the United States of America, 111*(40), 14565-14570.

Cichy, R. M., Kriegeskorte, N., Jozwik, K. M., van den Bosch, J. J. F., & Charest, I. (2019). The spatiotemporal neural dynamics underlying perceived similarity for real-world objects. *NeuroImage, 194*, 12-24.

Collell, G., & Moens, M. F. (2016). Is an image worth more than a thousand words? On the fine-grain semantic differences between visual and linguistic representations. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2807-2817.

Collell, G., Moens, M. F. (2018). Do neural network cross-modal mappings really bridge modalities? *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 462-468.

Contier, O., Baker, C. I., & Hebart, M. N. (2024). Distributed representations of behaviour-derived object dimensions in the human visual system. *Nature Human Behaviour,* 1-15.

Conwell, C., Prince, J. S., Alvarez, G., & Konkle, T. (2023). Language models of visual cortex: where do they work? And why do they work so well where they do? *Journal of Vision.*

Cui, X., Jeter, C. G., Yang, D., Montague, P. R., Eagleman, D. M. (2007). Vividness of mental imagery: individual variability can be measured objectively. *Vision Research, 47*(4), 474-478.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron, 73*(3), 415-434.

Diedrichsen, J., Berlot, E., Mur, M., Schutt, H. H., Shahbazi, M., & Kriegeskorte, N. (2020). Comparing representational geometries using whitened unbiased-distance-matrix similarity. *arXiv*.

Dijkstra, N. (2024). Uncovering the role of the early visual cortex in visual mental imagery. *Vision Research, 8*(2).

Dijkstra, N., Ambrogioni, L., Vidaurre, D., & van Gerven, M. (2020). Neural dynamics of perceptual inference and its reversal during imagery. *ELife, 9.*

Dijkstra, N., Bosch, S. E., & van Gerven, M. A. (2017). Vividness of visual imagery depends on the neural overlap with perception in visual areas. *The Journal of Neuroscience, 37*(5), 1367-1373.

Dijkstra, N., Bosch, S. E., van Gerven, M. A. J. (2019). Shared neural mechanisms of visual perception and imagery. *Trends in Cognitive Sciences, 23*(5), 423-434.

Dijkstra, N., Mostert, P., Lange, F. P., Bosch, S., & van Gerven, M. A. (2018). Differential temporal dynamics during visual imagery and perception. *ELife, 7.*

Dils, A. T., & Boroditsky, L. (2010). Visual motion aftereffect from understanding motion language. *Proceedings of the National Academy of Sciences of the United States of America, 107*(37), 16396-16400.

Doerig, A., Kietzmann, T. C., Allen, E., Wu, Y., Naselaris, T., Kay, K., & Charest, I. (2024). Visual representations in the human brain are aligned with large language models. *bioRxiv*.

Draschkow, D., & Vo, M. L. (2017). Scene grammar shapes the way we interact with objects, strengthens memories, and speeds search. *Scientific Reports, 7*(1), 16471.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature, 392*(6676), 598-601.

Fahrenfort, J. J., Scholte, H. S., & Lamme, V. A. (2007). Masking disrupts reentrant processing in human visual cortex. *Journal of Cognitive Neuroscience, 19*(9), 1488-1497.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: long-term memory structure and sentence processing. *Journal of Memory and Language, 41*(4), 469-495.

Friederici, A. D., Opitz, B., & von Cramon, D. Y. (2000). Segregating semantic and syntactic aspects of processing in the human brain: an fMRI investigation of different word types. *Cerebral Cortex, 10*(7), 698-705.

Friederici, A. D., Ruschemeyer, S. A., Hahne, A., & Fiebach, C. J. (2003). The role of left inferior frontal and superior temporal cortex in sentence comprehension: localizing syntactic and semantic processes. *Cerebral Cortex, 13*(2), 170-177.

Ganis, G., & Schendan, H. E. (2008). Visual mental imagery and perception produce opposite adaptation effects on early brain potentials. *Neuroimage, 42*(4), 1714-1727.

Golomb, J. D., & Kanwisher, N. (2012). Higher level visual cortex represents retinotopic, not spatiotopic object location. *Cerebral Cortex, 22*(12), 2794-2810.

Gomez, J., Natu, V., Jeska, B., Barnett, M., & Grill-Spector, K. (2018). Development differentially sculpts receptive fields across early and high-level human visual cortex. *Nature Communications, 9*(1), 788.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hamalainen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience, 7.*

Greene, M. R., & Oliva, A. (2009). The briefest of glances: the time course of natural scene understanding: the time course of natural scene understanding. *Psychological Science, 20*(4), 464-472.

Güçlü U., & van Gerven, M. A. (2015). Semantic vector space models predict neural responses to complex visual stimuli, *arXiv*.

Hagoort, P., Hald, L., Bastiaansen, M., Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science, 304*(5669), 438-441.

Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature, 458*(7238), 632-635.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293*(5539), 2425-2430.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour, 4*(11), 1173-1185.

Horikawa, T., & Kamitani, Y. (2017). Hierarchical neural representation of dreamed objects revealed by brain decoding with deep neural network features. *Frontiers in Computational Neuroscience, 11*(4).

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature, 532*(7600), 453-458.

Ishai, A., Ungerleider, L. G., Martin, A., Schouten, J. L., & Haxby J. V. (1999). Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Sciences of the United States of America, 96*(16), 9379-9384.

Kaiser, D., Azzalini, D. C., & Peelen, M. V. (2016a). Shape-independent object category responses revealed by MEG and fMRI decoding. *Journal of Neurophysiology, 115*(4), 2246-2250.

Kaiser, D., & Cichy, R. M. (2018). Typical visual-field locations enhance processing in object-selective channels of human occipital cortex. *Journal of Neurophysiology, 120*(2), 848-853.

Kaiser, D., Oosterhof, N. N., & Peelen, M. V. (2016b). The Neural Dynamics of Attentional Selection in Natural Scenes. *The Journal of Neuroscience, 36*(41), 10522-10528.

Kaiser, D., Quek, G. L., Cichy, R. M., & Peelen, M. V. (2019). Object vision in a structured world. *Trends in Cognitive Sciences, 23*(8), 672-685.

Kaneshiro, B., Perreau-Guimaraes, M., Kim, H. S., Norcia, A. M., & Suppes, P. (2015). A representational similarity analysis of the dynamics of object processing using single-trial EEG classification. *PLoS One, 10*(8).

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience, 17*(11), 4302-4311.

Kiani, R., Esteky, H., Mirpour, K., & Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology, 97*(6), 4296-4309.

Klein, I., Dubois, J., Mangin, J. F., Kherif, F., Flandin, G., Poline, J. B., Denis, M., Kosslyn, S. M., & Le Bihan, D. (2004). Retinotopic organization of visual mental images as revealed by functional magnetic resonance imaging. *Brain Research, Cognitive Brain Research, 22*(1), 26-31.

Kosmyna, N., Lindgren, J. T., & Lecuyer, A. (2018). Attending to visual stimuli versus performing visual imagery as a control strategy for EEG-based brain-computer interfaces. *Scientific Reports, 8*(1).

Kosslyn, S. M. (1975). Information representation in visual images. *Cognitive Psychology, 7*(3), 341-370.

Kosslyn, S. M., Ball, T. M., & Reiser, B. J. (1978). Visual images preserve metric spatial information: evidence from studies of image scanning. Journal of Experimental Psychology. *Human Perception and Performance, 4*(1), 47-60.

Kriegeskorte, N., & Mur, M. (2012). Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology, 3,* 245.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2*.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science, 207*(4427), 203-205.

Lahner, B., Dqivedi, K., Iamshchinina, P., Graumann, M., Lascelles, A., Roig, G., Gifford, A. T., Pan, B., Jin, S., Murty, N. A. R., Kay, K., Oliva, A., & Cichy, R. (2024). Modeling short visual events through the BOLD moments video fMRI dataset and metadata. *Nature Communications, 15*(1).

Lee, S., Jang, S., & Jun, S. C. (2022). Exploring the ability to classify visual perception and visual imagery EEG data: toward an intuitive BCI system. *Electronics, 11*(17).

Lee, S. H., Kravitz, D. J., & Baker, C. I. (2012). Disentangling visual imagery and perception of real-world objects. *Neuroimage, 59*(4), 4064-4073.

Lewis, K. J., Borst, G., & Kosslyn, S. M. (2011). Integrating visual mental images and visual perceptions: new evidence for depictive representations. *Psychological Research, 75*(4), 259-271.

Lin, T. Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Lawrence Zitnick, C., & Dollar, P. (2014). Microsoft COCO: common objects in context. *arXiv*.

Linde-Domingo, J., Treder, M. S., Kerren, C., & Wimber, M. (2019). Evidence that neural information flow is reversed between object perception and object reconstruction from memory. *Nature Communications, 10*(1).

Liu, D., Wang, L., & Han, Y. (2024). Mental simulation of colour properties during language comprehension: influence of context and comprehension stages. *Cognitive Processing*, 1-14.

Magri, C., & Konkle, T. (2019). Comparing facets of behavioral object representation: implicit perceptual similarity matches brains and models. *2019 Conference on Cognitive Computational Neuroscience*.

Marks, D. F. (1973). Visual imagery differences in the recall of pictures. *British Journal of Psychology, 64(*1), 17-24.

Martin, S., Brunner, P., Iturrate, I., Millan, J. del R., Schalk, G., Knight, R. T., & Pasley, B. N. (2016). Word pair classification during imagined speech using direct brain recordings. *Scientific Reports, 6*(1).

Martin, C. N., Douglas, D., Newsome, R. N., Man, L. L., & Barense, M. D. (2018). Integrative and distinctive coding of visual and conceptual object features in the ventral visual stream. *Elife, 7.*

Mitchell, D. J., & Cusack, R. (2016). Semantic and emotional content of imagined representations in human occipitotemporal cortex. *Scientific Reports, 6.*

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A., & Kriegeskorte, N. (2013). Human object-similarity judgements reflect and transcend the primate-IT object representation. *Frontiers in Psychology, 4.*

Naselaris, T., Olman, C. A., Stansbury, D. E., Ugurbil, K., & Gallant, J. L. (2015). A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. *Neuroimage, 105*, 215-228.

Newman, S. D., Ikuta, T., & Burns, Jr T. (2010). The effect of semantic relatedness on syntactic analysis: an fMRI study. *Brain and Language, 113*(2), 51-58.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS Computational Biology, 10*(4).

Paivio, A. (1971). *Imagery and verbal processes*. New York: Holt, Rinehart, and Winston.

Paivio, A. (1986). *Mental representations*: *A dual-coding approach.* New York: Oxford University Press.

Pallier, C., Devauchelle, A. D., & Dehaene, S. (2011). Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences of the United States of America, 108*(6), 2522-2527.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., & Duchesnay, M. P. E. (2011). Scikit-learn: machine learning in Python. *The Journal of Machine Learning Research, 12,* 2825-2830.

Pei, X., Leuthardt, E. C., Gaona, C. M., Brunner, P., Wolpaw, J. R., & Schalk, G. (2011). Spatiotemporal dynamics of electrocorticographic high gamma activity during overt and covert word repetition. *Neuroimage, 54*(4), 2960-2971.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Hochenberger, R., Sogo, H., Kastman, E., & Lindelow, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods, 51*(1), 195-203.

Popham, S. F., Huth, A. G., Bilenko, N. Y., Deniz, F., Gao, J. S., Nunez-Elizalde, A. O., & Gallant, J. L. (2021). Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature Neuroscience, 24*(11), 1628-1636.

Proverbio, A. M., & Riva, F. (2009). RP and N400 ERP components reflect semantic violations in visual processing of human actions. *Neuroscience Letters, 459*(3), 142-146.

Quilty-Dunn, J., Porot, N., & Mandelbaum, E. (2022). The best game in town: the reemergence of the language-of-thought hypothesis across the cognitive sciences. *The Behavioral and Brain Sciences, 46*.

Rayner, K., Reichle, E. D., Stroud, M. J., Williams, C. C., & Pollatsek, A. (2006). The effect of word frequency, word predictability, and font difficulty on the eye movements of young and older readers. *Psychology and Aging, 21*(3), 448-465.

Savarimuthu, A., & Ponniah, R. J. (2024). The mind's eye is not a metaphor: visuospatial working memory and mental imageries for learning and language comprehension. *Psychological Studies, 69*(2), 158-168.

Schutt, H. H., Kipnis, A. D., Diedrichsen, J., & Kriegeskorte, N. (2023). Statistical inference on representational geometries. *ELife, 12.*

Seydell-Greenwald, A., Wang, X., Newport, E. L., Bi, Y., & Striem-Amit, E. (2023). Spoken language processing activates the primary visual cortex. *PLoS One, 18*(8).

Shatek, S. M., Grootswagers, T., Robinson, A. K., & Carlson, T. A. (2019). Decoding images in the mind's eye: the temporal dynamics of visual imagery. *Vision, 3*(4).

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, 171*(3972), 701-703.

Speed, L. J., Eekhof, L. S., & Mak, M. (2024). The role of visual imagery in story reading: evidence from aphantasia. *Consciousness and Cognition, 118*(103645).

Stanfield, R. A., & Zwaan, R. A. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science, 12*(2), 153-156.

Stansbury, D. E., Naselaris, T., & Gallant, J. L. (2013). Natural scene statistics account for the representation of scene categories in human visual cortex. *Neuron, 79*(5), 1025-1034.

Stecher, R., & Kaiser, D. (2023). Imaginary scenes are represented in cortical alpha activity. *bioRxiv*.

Tang, J., LeBel, A., & Huth, A. G. (2021). Cortical representations of concrete and abstract concepts in language combine visual and linguistic representations. *bioRxiv*.

Tang, J., Du, M., Vo, V. A., Lal, V., & Huth, A. G. (2023). Brain encoding models based on multimodal transformers can transfer across language and vision. *arXiv*.

Van Berkum, J. J., Hagoort, P., & Brown, C. M. (1999). Semantic integration in sentences and discourse: evidence from the N400. *Journal of Cognitive Neuroscience, 11*(6), 657-671.

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: benefits, costs, and ERP components. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology, 83*(2), 176-190.

Võ, M. L., Boettscher, S. R., & Draschkow, D. (2019). Reading scenes: how scene grammar guides attention and aids preception in real-world environments. *Current Opinion in Psychology, 29*, 205-210.

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *Neuroimage, 137*, 188-200.

Wang, A. Y., Kay, K., Naselaris, T., Tarr, M. J., & Wehbe, L. (2023). Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence, 5,* 1415-1426.

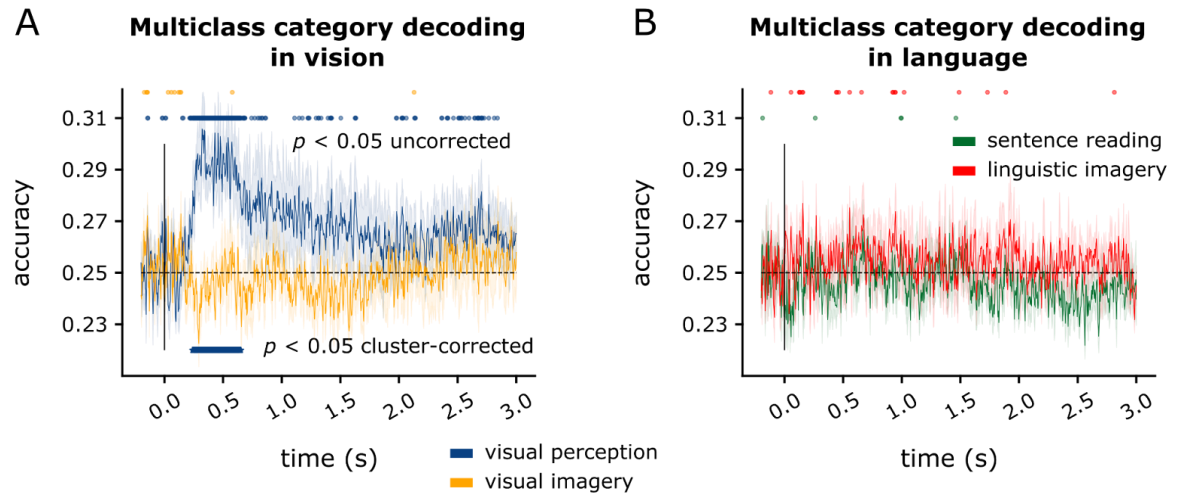
Wardle, S. G., Kriegeskorte, N., Grootswaggers, T., Khalig-Razavi, S. M., & Carlson, T. A. (2016). Perceptual similarity of visual patterns predicts dynamic neural activation patterns measured with MEG. *Neuroimage, 132*, 50-70.

Xie, S., Kaiser, D., & Cichy, R. M. (2020). Visual imagery and perception share neural representations in the alpha frequency band. *Current Biology, 30*(15).

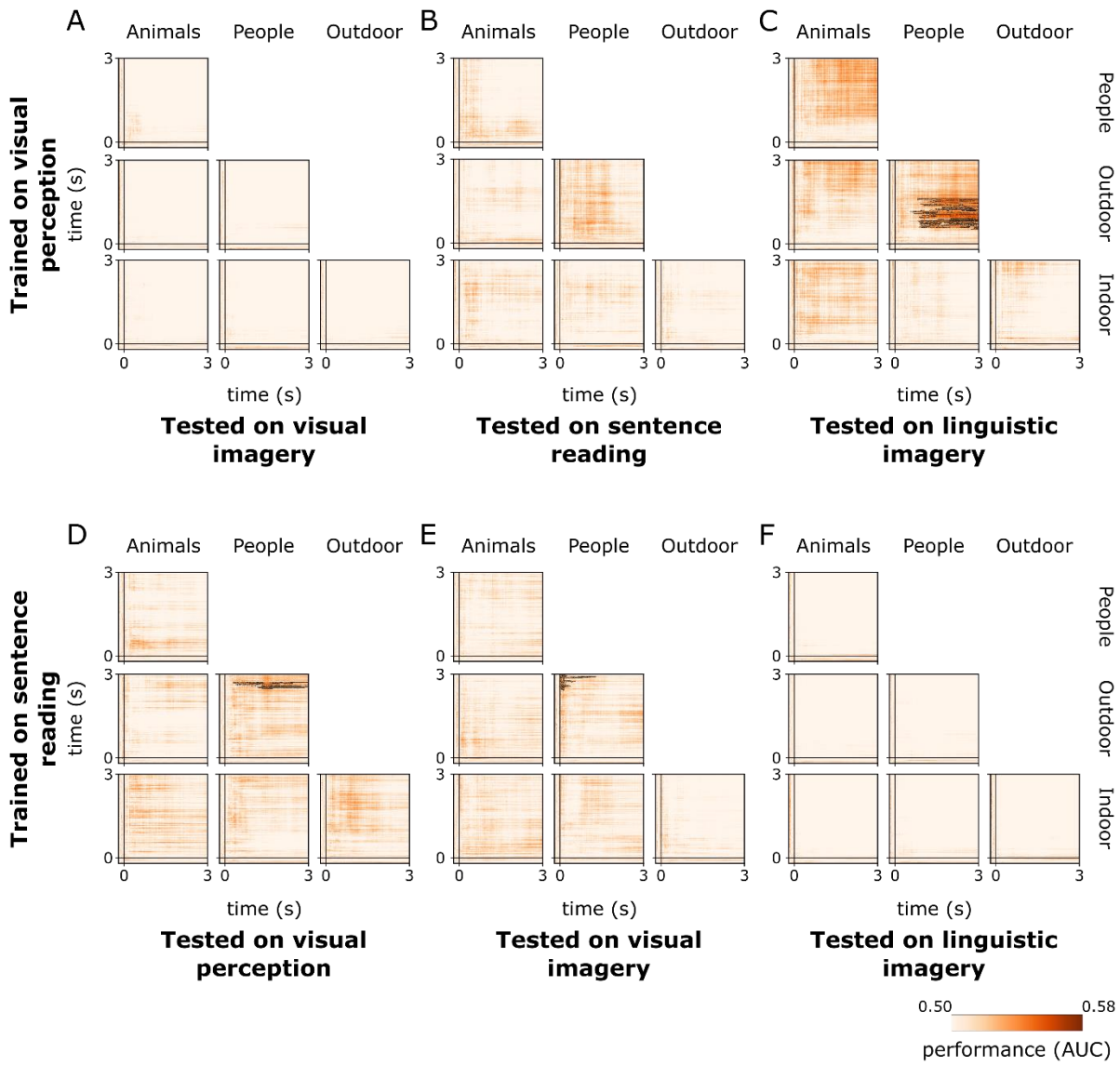
Yeh, L. C., & Peelen, M. V. (2022). The time course of categorical and perceptual similarity effects in visual search. *Journal of Experimental Psychology. Human Perception and Performance, 48*(10), 1069-1082.

Zwaan, R. A., Stanfield, R. A., & Yaxley, R. H. (2002). Language comprehenders mentally represent the shapes of objects. *Psychological Science, 13*(2), 168-171.

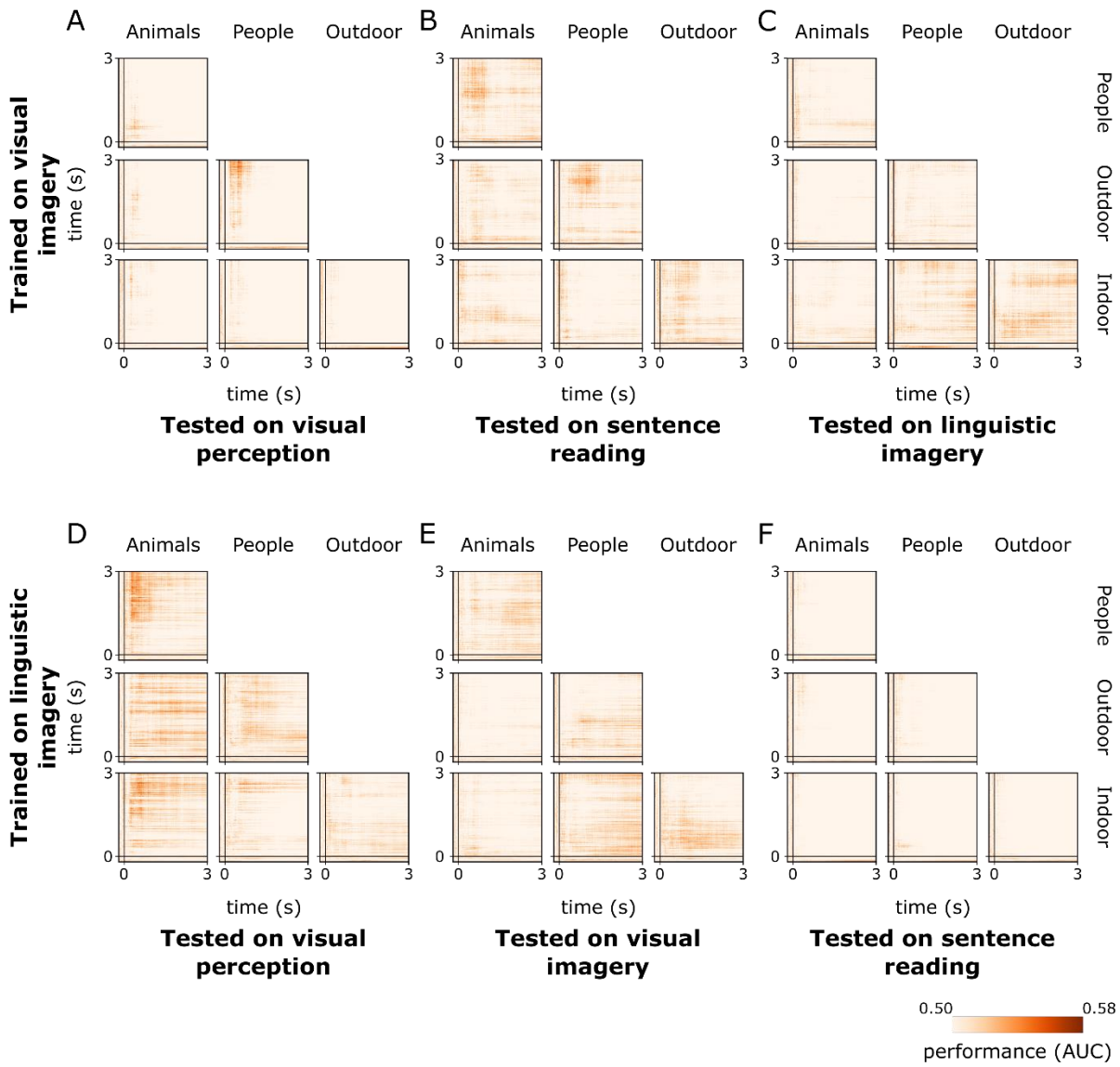
4.7 Supplementary material



Supplementary Figure 4.1 | Multiclass decoding in multimodal perception and imagery. Motivated by the clear categorical distinction observed in the similarity judgements, we decided to solidify the representational geometry by performing a 4-class classification. We selected 10 exemplars from each category cluster (animals, people, indoor, outdoor) and trained a 4-class LDA classifier within participants to discriminate between the categorical representations in the visual (A) and linguistic (B) modalities. The line plot shows averaged classification scores across participants. The shaded area depicts the SEM across participants. The significance of classification scores across participants was determined by a one-sided t-test against chance (1⁄4 = 25% chance level, p < 0.05 cluster-corrected). Stars in solid colour at the bottom of each graph mark significant clusters. Light-colour points at the top mark significant classification scores at the uncorrected p < 0.05 level.



Supplementary Figure 4.2 | Generalisation of categorical representations from visual perception and sentence reading. A binary LDA classifier was trained on all six pairwise categorical representations in visual perception (A-C) and sentence reading (D-F) to investigate how the representation generalises across time and modalities. The performance was thresholded at the chance level (1/2 = 0.5) and tested for significance using a cluster-based permutation test at the $p < 0.05$ level.



Supplementary Figure 4.3 | Generalisation of categorical representations across time and modalities. A binary classifier was trained on all six pairwise categorical representations in visual imagery and linguistic imagery to investigate how the representation generalises across time and modalities. The performance was thresholded at the chance level (1/2 = 0.5) and tested for significance using a cluster-based permutation test at the $p < 0.05$ level.

Chapter 5

General discussion

Imagine an elephant playing with a pink feather duster. It is unlikely that you have seen a similar scene before, yet with just a few words, you can effortlessly draw a vivid image of an elephant and a pink feather duster. You can also come to the conclusion that such a scene is more similar to a cat chasing a mouse rather than a scene where a car stops at a traffic light. How does language do that? Through carefully designed behavioural experiments, this thesis set out to investigate how representations simulated in language overlap with representations in vision and what is its relevance to behaviour. To illustrate, in Chapter 2, we collected similarity judgements of natural scene images and sentence captions describing these images. We recovered the visual and linguistic representational geometries to answer questions such as: does the similarity structure of language correspond to the similarity structure of the directly perceived visual scenes? How do the representational geometries align with brain response patterns to viewing the corresponding natural scene images? How stable are the geometries across individuals and sessions? We found that the similarity judgements were not only similar in the behavioural context but also that they both predicted a strikingly similar network of brain response patterns along the mid- and high-level visual regions.

In Chapter 3, we presented a novel cross-modal priming experiment to investigate whether the regularities of the similarity space predict overt behaviour. We found a clear cross-modal priming effect where response times were faster the closer the linguistic prime and the visual target were represented in the similarity space. Next, upon discovering that the linguistic similarity judgements displayed higher levels of variability between participants (Chapter 2), we leveraged participants' own representational geometries to predict their own response times. This showed that the idiosyncratic linguistic dissimilarities were the strongest predictors of the cross-modal priming effect, revealing the unique role of language in emphasising personally meaningful representations. Finally, in

Chapter 4, we used electroencephalography to explore the neural underpinnings of the visuo-linguistic overlap in a mental imagery task performed on the visual and linguistic natural scene stimuli. Representational similarity analysis revealed some, albeit limited, representational overlap of the visual and linguistic neural patterns with the behavioural similarity judgements. An assessment of the noise ceiling, however, pointed to a high measurement noise likely affecting the reliability of the findings. Together, this thesis is the first comprehensive investigation of how representations in vision and language converge onto a shared representational space, and how its similarity structure guides overt behaviour. In the following sections, we draw on these findings to describe what characteristics give rise to the visuo-linguistic alignment and offer recommendations for future research.

5.1 Representations in vision and language converge in their ability to encode regularities of the visual world

5.1.1 Representations simulated in language resemble representations of perceived scenes.

On the question of representational overlap, our findings demonstrate that the internally simulated representation in language does indeed resemble representations of the directly perceived objects. Evidence for this comes from Chapter 3 where we found a clear cross-modal priming effect when a linguistic prime preceded the search for a visual target. Prior research showed that when participants form a preparatory object representation, visual search for the corresponding object is faster because the preparatory activation triggers a neural representation shared with the target object (Gayet, & Peelen, 2022). Therefore, for cross-modal priming to occur, language must have triggered an overlapping representation with vision, inducing the preparatory activity. This

finding supports the literature that suggests that for us to obtain an accurate mental representation, language must draw on more experiential domains, such as vision (Dils, & Boroditsky, 2010; Harder, & Tylén, 2019; Lupyan, & Spivey, 2010; Lupyan, & Swingley, 2012; Meteyard et al., 2007; Meyer et al., 2007; Sperber et al., 1979; Stanfield, & Zwaan, 2001). Similarly, without an internal representation resembling that of directly perceived visual scenes, participants would not have been able to configure the visual and linguistic similarity judgements as accurately as they did. Most importantly, we observed this effect even though participants performed the linguistic similarity judgements never seeing the visual scenes before. What mechanisms enable language to generate an internal representation that is isomorphic to the one induced by vision?

5.1.2 Multimodal systems are optimised to encode behaviourally relevant regularities of the visual world

The most likely explanation is that both vision and language converge in their inherent ability to recognise and encode regularities of the visual world, further supporting the ideas initially presented by Shepard and Chipman (1970). In Chapter 2, we discovered a strong correspondence between similarity judgements in vision and language, showing that internal representations in language preserve the relational structure of directly perceived scenes (Shepard, & Chipman, 1970). As a result of Chapter 3, we now know that the regularities of the same representational space can even predict overt behaviour, with response times being faster the closer the linguistic prime and visual target are represented in the representational similarity space. Based on the proposal of Kriegeskorte and colleagues (2008a), it is possible that representations elicited by both the prime and target project as points onto a shared representational space, where the distance between them represents their dissimilarity. From this perspective, response times – serving as an indirect measure of the representational

dissimilarity – decrease linearly with decreasing distance between the prime's and target's representation in the space. These findings provide a new understanding of how the representational similarity space, previously exclusively associated with vision (Carlson et al., 2014; Cohen et al., 2014; Cohen et al., 2015; Cohen et al., 2017; Kriegeskorte et al., 2008b; Mur et al., 2013; Yeh et al., 2024), supports cross-modal integration in vision and language.

Further analyses in Chapter 2 revealed that the representational geometries derived from visual and linguistic similarity judgements both aligned more closely with the similarity structure of large language model (LLM, Doerig et al., 2024) sentence embeddings than the image-trained AlexNet (Krizhevsky et al., 2012). Similarly, many authors previously reported that LLMs are great predictors of visually- and language-evoked activity in the human brain (Doerig et al., 2024; Lahner et al., 2024). The success of LLMs likely lies in their ability to capture rich contextual information, which parallels how the human brain extracts semantic information about its sensory inputs (Federmeier, & Kutas, 2001; Haxby et al., 2001; Heilbron et al., 2022; Ishai et al., 1999; Kiani et al., 2007; Kaiser, & Cichy, 2018; Kriegeskorte et al., 2008b; Schrimpf et al., 2021). What's significant about our findings is that LLMs also seem to mirror some aspects of human cognitive processes, including how people categorise concepts and relationships based on semantic similarity. This is even more significant considering that the present LLM embeddings were derived from a model that was trained on substantially less data than its counterparts (Doerig et al., 2024). These findings are important in two aspects: first, the findings show that multimodal systems are optimised to encode behaviourally relevant representations about the visual world. Second, they underscore the potential of the LLM-trained recurrent convolutional neural network (RCNN) in modelling and understanding human cognition.

5.2 Stability of the representational similarity space as a function of task demands

5.2.1 Representation in language is grounded in visual experience

Any further answer to how the shared representational space supports behaviour requires a deeper understanding of its stability in response to task demands. To illustrate, the results on cross-modal projection in Chapter 2 revealed that the stability of the representational geometry varies depending on the cognitive load in vision and language. First, we discovered that if participants completed the visual similarity judgements first, then this experience was more likely to narrow down the number of potential candidates in the linguistic judgements. This manifested as a decreased interindividual variability in the configuration of the linguistic similarity judgements. What's interesting about this observation is that if participants completed the linguistic similarity judgements first, the configuration was more variable between individuals, likely as a result of personally meaningful simulations of the scenes. Consistent with the concept of non-committance in visual imagery (Bigelow et al., 2023), a possibility is that the internal representation in language was less accurate and only reflected the general concepts of the sentence. For example, the sentence "an elephant playing with a pink feather duster" would elicit an imperfect representation of several features that are diagnostic enough of an elephant (e.g., four legs, big ears, grey, etc.). Such flexibility would then enable the visual representation to be easily imprinted into the similarity space of the linguistic similarity judgements in the following session. This shows that while language does simulate an accurate representation without a direct reference to a currently perceived object (otherwise we would have not seen the overlap between the similarity judgements, or the cross-modal priming effect), embodied experience seems to be integral to a deeper, more concrete understanding (Barsalou, 2008; Kemmerer,

2023; Tang et al., 2021; Zwaan, 2014). This study is the first to provide a concrete example of embodiment in human behaviour, and offer relevance to artificial models of human cognition that naturally benefit from joint visuo-linguistic training (Andrews et al., 2009; Lazaridou et al., 2015; Kiela, & Bottou, 2014; Silberer, & Lapata, 2014; Vong et al., 2024; Wang et al., 2023).

5.2.2 Language as a unique descriptor of representations in personally meaningful contexts

The most interesting result that came from the cross-modal projections, however, was the transferability of idiosyncrasies of the linguistic similarity judgements. We found that if participants completed the linguistic similarity judgements first, then the space of potential representational candidates expanded, in other words, there could be more than one visual image linked with a single description. This manifested as an increase in interindividual variability in the follow-up visual similarity judgements (note that these were initially closely aligned between participants). It is worth emphasising that these findings cannot be attributed entirely to memory because otherwise, we would have found both modalities to have a similar effect on one another. Additionally, participants were not aware that the stimuli covaried between the modalities, and were scheduled to complete the judgements at least a week apart, minimising the likelihood of a memory effect. An important implication of this observation therefore is that language has a unique ability to re-activate idiosyncratic representations, perhaps related to previous experiences with the visual world and personally meaningful semantic relations. The fact that the simulation of personally meaningful representations even shapes the perceived similarity of visual representations in real-time (which was demonstrated as the increase in inter-individual variability in the visual judgements) is a characteristic previously only described in early

representation learning (Bergelson, & Swingley; 2012; LaTourrette, & Waxman, 2020; LaTourrette et al., 2023; Tincoff, & Jusczyk, 1999).

Notably, we observed a similar language-to-vision mapping in the cross-modal priming experiment in Chapter 3. We discovered that idiosyncratic dissimilarities derived from participants' own linguistic similarity judgements outperformed visual dissimilarities in predicting participants' response times. A possible explanation for such an imbalance in cross-modal mappings comes from Tversky (1977) who noted that there are many asymmetries in similarity often arising between items that differ in familiarity, prototypicality, or salience. For example, a representation of a wolf might be more likely to elicit the more familiar member of the category – a dog, rather than the other way around. We now know from Chapter 2 that language simulates representations based on personally meaningful concepts. It is therefore possible that the linguistic prime elicited the idiosyncratic aspect whose relationship to the visual target was best predicted by the salience of idiosyncratic linguistic dissimilarities. This also relates to what is frequently described in artificial models of human cognition showing differences in language-to-vision mapping (Lupyan, 2012), with language typically being easier to map onto visual representations. This is because a single sentence can correspond to many potential scene images. Consider the sentence from the beginning of the discussion: "An elephant playing with a pink feather duster". This sentence could match an image of an elephant standing in a grassy field; however, it could equally match an image where the elephant stands in a desert. A similar relationship is, however, much more difficult to observe in vision-to-language mapping because fewer sentences could match this image. It is therefore possible that the visual similarity judgements were much more constrained, as indicated by the lower interindividual differences observed in Chapter 2. This constraint may have resulted in visual similarity judgements being less accurate predictors of the prime-target relationship in Chapter 3.

5.2.3 Dimensions of the representational similarity space

We found that the stability of the representational similarity space varies not only as a function of individual differences but also in relation to task demands. Chapter 3 showed that the linguistic prime facilitated the search for a visual target when the target was placed among visually related distractors. The cross-modal priming effect diminished with increasing dissimilarity of the target to the simultaneously presented distractors. This is a phenomenon frequently studied in visual experiments where it is well established that the semantic link between a prime and a target takes precedence if visual features are insufficient to guide the search behaviour (Yeh, & Peelen, 2022). These findings speak about the dimensions of the representational similarity space. As suggested by several studies (Contier et al., 2024; Hebart et al., 2020; Kaiser, & Cichy, 2018), representational similarity spaces are selective to a multitude of visual features and semantic categories. Along with the finding that the complex structure of LLM-trained RCNNs (Doerig et al., 2024) is a well-suited predictor of the representational similarity space (Chapter 2), the shared space for vision and language likely encodes multidimensional features. These findings provide a deeper insight into how the human brain may aggregate multimodal representations into a unified similarity space. We show that the features of this space are multidimensional, flexible, and shift depending on the task demands, further demonstrating the complementary, albeit unique, features that vision and language encode (Collell, & Moens, 2016). An outstanding question is in regards to where in the brain the shared representational similarity space is embedded.

5.3 Is the shared representational space grounded in the visual cortex?

Chapter 2 presented key evidence suggesting that the representational geometry from linguistic similarity judgements is embedded in the modality-specific system

for visual perception. We demonstrated that the judgements derived from vision and language both predicted a strikingly similar network of visually evoked response patterns along the mid- and high-level visual regions. These findings are consistent with prior research that demonstrated a widely distributed network evoked by language, which overlapped with regions typically associated with vision (Huth et al., 2016; Popham et al., 2021; Tang et al., 2021; Tang et al., 2023). Most importantly, these predictions cannot be attributed to task order effects because the results remained even when we predicted the neural response patterns using linguistic similarity judgements that were made by participants who had never seen the actual images before. What this suggests is that participants simulated internal representations that were stable and prototypical enough for them to align with neural responses to viewing the corresponding visual scenes. Along with the observation that language has a unique role in drawing personally meaningful representations (Chapter 2, Chapter 3), these findings may potentially have an important impact on how we understand semantic representations in human vision.

5.4 Suggestions for future work

5.4.1 What is the degree to which the neural activity patterns in language overlap with activity patterns in vision?

The question of how language shapes the representational space still requires further research, particularly into the degree to which language is grounded in the modality-specific systems for visual perception. Chapter 4 was designed to address this question by exploring the neural patterns in response to visual and linguistic imagery. The findings were, however, limited by the relatively small sample, which led to high levels of measurement noise and inconsistent results. Notwithstanding these limitations, Chapter 4 was the first to associate

behavioural representational geometry with language-related neural activity patterns. We found some overlap between the neural patterns and the similarity judgements, which may indirectly imply that there is a degree of isomorphism between the visually evoked fMRI, multimodal EEG, and behavioural representational geometries (Kriegeskorte, & Kievit, 2013). Whilst these results did not survive correction for multiple comparisons, they offer a new insight into how language-related discourse evolves over time, suggesting a role of fine-grained similarity coding (Federmeier, & Kutas, 1999; Federmeier, & Kutas, 2001; Heilbron et al., 2022; Kutas, & Federmeier, 2011). A natural progression of this work is to expand the sample size and identify the neural source of the representational overlap with the similarity judgements by using searchlight methods in the sensor space. Future research should also adopt the rapid serial presentation (RSVP) paradigm to more precisely time-lock internal events of the developing discourse in the sentence reading condition. It is also necessary to identify the neural source of the overlap between the modalities by generalising the linear classifiers in the sensor space (Haufe et al., 2014). Once these issues are addressed, further questions regarding the impact of idiosyncratic linguistic semantics on the grounded model of cognition should be addressed.

5.4.2 Impact of linguistic semantics on grounded models of cognition

In 2023, David Kemmerer asserted that if language is grounded, then its connection to the representational space in the modality-specific systems must shape how we perceive the world. Prior literature provided supporting evidence for Kemmerer's theory. It was found that cross-linguistic differences in grammar (Athanasopoulos et al., 2015) and semantic categorisation (Winawer et al., 2007) affect performance in semantic, but also relatively low-level perceptual tasks. In fact, Charest and colleagues (2014) showed that the semantic space in the visual brain predicts individuals' own judgements of similarity. The observation that

participants in Chapter 2 exhibited higher interindividual variability in the linguistic similarity judgements suggests that language is a particularly useful tool in eliciting the idiosyncratic component of the representational semantic space. Building on this finding, subsequent research should aim to further explore the feasibility of using language to describe the individually unique configuration of the internal representational space. We also found that the idiosyncratic aspect derived from linguistic judgements was then projected onto the visual representational space in session 2 (Chapter 2) and that it was the strongest predictor of the relationship between the primed linguistic and observed visual target representation (Chapter 3). A logical next step would be to replicate these findings to confirm that the cross-modal projection is not due to memory, but rather a direct result of the imbalance in the language-to-vision versus vision-to-language mapping. These lines of inquiry may potentially have a profound impact on our understanding of how the stability of the grounded representational space changes in response to differential semantic, respective linguistic, typology.

5.5 Concluding remarks

This thesis is the first comprehensive investigation of how representations in vision and language project onto a shared representational space. First, we showed that both vision and language converge in their inherent ability to encode the regularities of the visual world and that both behavioural geometries resemble the internal representational structure of visually evoked brain responses (Chapter 2). We then showed that the representational geometry of such a space is a well-suited predictor of the cross-modal priming effect, indicative of its relevance in overt behaviour (Chapter 3). Upon the finding that linguistic similarity judgements exhibit a higher inter-individual variability (Chapter 2), we leveraged the idiosyncratic linguistic dissimilarities to make predictions about the cross-modal mapping and stability of the shared representational space. This

revealed that language has a unique role in drawing on the idiosyncratic aspect of the representational space, underscoring the necessity for grounded models of cognition to recognise the role of linguistic semantics. Despite the relatively limited sample in Chapter 4, our findings suggest that there is some isomorphism between the similarity structures of the simulated representations in language (Chapter 4) and visually evoked responses to viewing natural scene images (Chapter 2). In summary, this thesis is the first step toward establishing a unified semantic space that encodes behaviourally relevant representations about multimodal sensory experience.

5.6 References

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. Psychological Review, 11(3), 463-498.

Athanasopoulos, P., Bylund, E., Montero-Melis, G., Damjanovic, L., Schartner, A., Kibbe, A., Riches, N., & Thierry, G. (2015). Two languages, two minds: flexible cognitive processing driven by language of operation. Psychological Science, 26(4), 518-526.

Barsalou, L. W. (2008). Grounded cognition. Annual Review of Psychology, 59, 617-645.

Bergelson, E., & Swingley, D. (2011). At 6-9 months, human infants know the meanings of many common nouns. Proceedings of the National Academy of Sciences of the United States of America, 109(9), 3253-3258.

Bigelow, E., J., McCoy, J. P., & Ullman, T. D. (2023). Non-commitment in mental imagery. Cognition, 238.

Carlson, T. A., Ritchie, J. B., Kriegeskorte, N., Durvasula, S., & Ma, J. (2014). Reaction time for object categorization is predicted by representational distance. Journal of Cognitive Neuroscience, 26(1), 132-142.

Charest, I., Kievit, R. A., Schmitz, T. W., Deca, D., & Kriegeskorte, N. (2014). Unique semantic space in the brain of each beholder predicts perceived similarity. Proceedings of the National Academy of Sciences of the United States of America, 111(40), 14565-14570.

Cohen, M. A., Alvarez, G. A., Nakayama, K., & Konkle, T. (2017). Visual search for object categories is predicted by the representational architecture of high-level visual cortex. Journal of Neurophysiology, 117(1), 388-402.

Cohen, M. A., Konkle, T., Rhee, J. Y., Nakayama, K., & Alvarez, G. A. (2014). Processing multiple visual objects is limited by overlap in neural channels. Proceedings of the National Academy of Sciences of the United States of America, 111(24), 8955-8960.

Cohen, M. A., Nakayama, K., Konkle, T., & Stantic, M. (2015). Visual awareness is limited by the representational architecture of the visual system. Journal of Cognitive Neuroscience, 27(11), 2240-2252.

Collell, G., Moens, M. F. (2016). Is an image worth more than a thousand words? On the fine-grain semantic differences between visual and linguistic representations. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers,* 2807-2817.

Contier, O., Baker, C. I., & Hebart, M. N. (2024). Distributed representations of behaviour-derived object dimensions in the human visual system. Nature Human Behaviour, 1-15.

Dils, A. T., & Boroditsky, L. (2010). Visual motion aftereffect from understanding motion language. Proceedings of the National Academy of Sciences of the United States of America, 107(37), 16396-16400.

Doerig, A., Kietzmann, T. C., Allen, E., Wu, Y., Naselaris, T., Kay, K., & Charest, I. (2024). Visual representations in the human brain are aligned with large language models. bioRxiv.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: long-term memory structure and sentence processing. Journal of Memory and Language, 41(4), 469-496.

Federmeier, K. D., & Kutas, M. (2001). Meaning and modality: influences of context, semantic memory organization, and perceptual predictability on picture processing. Journal of Experimental Psychology. Learning, Memory, and Cognition, 27(1), 202-224.

Gayet, S., & Peelen, M. V. (2022). Preparatory attention incorporates contextual expectations. Current Biology, 32(3), 687-692.

Harder, M., & Tylen, K. (2019). What is up and down in embodied language processing? An experimental study on semantic priming of visual perception. Cognitive Semiotics, 12(2).

Haufe, S., Meinecke, F., Gorgen, K., Dahne, S., Haynes, J. D., Blankertz, B., & Biessmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage, 87, 96-110.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science, 293(5539), 2425-2430.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. Nature Human Behaviour, 4(11), 1173-1185.

Hebb, D. O. (1949). The Organization of Behavior. A Neuropsychological Theory. John Wiley & Sons, Inc, London.

Heilbron, M., Armeni, K., Schoffelen, J. M., Hagoort, P., & de Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences of the United States of America, 119*(32).

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature, 532*(7600), 453-458.

Ishai, A., Ungerleider, L. G., Martin, A., Schouten, J. L., & Haxby J. V. (1999). Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Sciences of the United States of America, 96*(16), 9379-9384.

Kaiser, D., & Cichy, R. M. (2018). Typical visual-field locations enhance processing in object-selective channels of human occipital cortex. *Journal of Neurophysiology, 120*(2), 848-853.

Kemmerer, D. (2023). Grounded cognition entails linguistic relativity: a neglected implication of a major semantic theory. *Topics in Cognitive Science, 15*(4), 615-647.

Kiani, R., Esteky, H., Mirpour, K., & Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology, 97*(6), 4296-4309.

Kiela, D., & Bottou, L. (2014). Learning image embeddings using convolutional neural networks for improved multi-modal semantics. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 36-45.

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences, 17*(8), 401-412.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008a). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2.*

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P. A. (2008b). Matching categorical representations in inferior temporal cortex of man and monkey. *Neuron, 60*(6), 1126-1141.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems 25 (NIPS 2012)*.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Reviews of Psychology, 62*(1), 621-647.

Lahner, B., Dwivedi, K., Iamshchinina, P., Graumann, M., Lascelles, A., Roig, G., Gifford, A. T., Pan, B., Jin, S., Ratan Murty, N. A., Kay, K., Oliva, A., & Cichy, R. (2024). Modeling short visual events through the BOLD moments video fMRI dataset and metadata. *Nature Communications, 15*(1).

LaTourrette, A. S., & Waxman, S. R. (2020). Naming guides how 12-month-old infants encode and remember objects. *Proceedings of the National Academy of Sciences of the United States of America, 117*(35), 21230-21234.

LaTourrette, A., Chan, D. M., & Waxman, S. R. (2023). A principled link between object naming and representation is available to infants by seven months of age. *Scientific Reports, 13*(14328).

Lazaridou, A., Pham, N. T., & Baroni, M. (2015). Combining language and vision with a multimodal skip-gram model. *Proceedings of the 2015 Conference of the North America Chapter of the Association for Computational Linguistics: Human Language Technologies,* 153-163.

Lupyan, G. (2012). Linguistically modulated perception and cognition: the label-feedback hypothesis. *Frontiers in Psychology, 3.*

Lupyan, G., & Spivey, M. J. (2010). Making the invisible visible: verbal but not visual cues enhance visual detection. *PLoS One, 5*(7).

Lupyan, G., & Swingley, D. (2012). Self-directed speech affects visual search performance. *Quarterly Journal of Experimental Psychology, 65*(6), 1068-1085.

Meteyard, L., Bahrami, B., & Vigliocco, G. (2007). Motion detection and motion verbs: language affects low-level visual perception. *Psychological Science, 18*(11), 1007-1013.

Meyer, A. S., Belke, E., Telling, A. L., & Humphreys, G. W. (2007). Early activation of object names in visual search. *Psychonomic Bulletin & Review, 14*(4). 710-716.

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A., & Kriegeskorte, N. (2013). Human object-similarity judgements reflect and transcend the primate-IT object representation. *Frontiers in Psychology, 4.*

Polk, T. A., & Farah, M. J. (1995). Brain localization for arbitrary stimulus categories: a simple account based on Hebbian learning. *Proceedings of the National Academy of Sciences of the United States of America, 92*(26), 12350-12373.

Popham, S. F., Huth, A. G., Bilenko, N. Y., Deniz, F., Gao, J. S., Nunez-Elizalde, A. O., & Gallant, J. L. (2021). Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature Neuroscience, 24*(11), 1628-1636.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: integrative modelling converges in predictive processing. *Proceedings of the National Academy of Sciences of the United States of America, 118*(45).

Shepard, R. N., & Chipman, S. (1970). Second-order isomorphism of internal representations: shapes of states. *Cognitive Psychology, 1*(1), 1-17.

Silberer, C., & Lapata, M. (2014). Learning grounded meaning representations with autoencoders. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 1*, 721-732.

Smith, L. B., Suanda, S. H., & Yu, C. (2014). The unrealized promise of infant statistical word-referent learning. *Trends in Cognitive Sciences, 18*(5), 251-258.

Sperber, R. D., McCauley, C., Ragain, R. D., & Weil, C. M. (1979). Semantic priming effects on picture and word processing. *Memory & Cognition, 7*(5), 339-345.

Stanfield, R. A., & Zwaan, R. A. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science, 12*(2), 153-156.

Tang, J., Du, M., Vo, V. A., Lal, V., & Huth, A. G. (2023). Brain encoding models based on multimodal transformers can transfer across language and vision. *arXiv*.

Tang, J., LeBel, A., & Huth, A. G. (2021). Cortical representations of concrete and abstract concepts in language combine visual and linguistic representations. *bioRxiv*.

Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science, 10*(2), 172-175.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*(4), 327-352.

Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. M. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science, 383*(6682), 504-511.

Wang, A. Y., Kay, K., Naselaris, T., Tarr, M. J., & Wehbe, L. (2023). Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence, 5*, 1415-1426.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences of the United States of America, 104*(19), 7780-7785.

Yeh, L. C., & Peelen, M. V. (2022). The time course of categorical and perceptual similarity effects in visual search. *Journal of Experimental Psychology. Human Perception and Performance, 48*(10), 1069-1082.

Yeh, L. C., Throat, S., & Peelen, M. V. (2024). Predicting cued and oddball visual search performance from fMRI, MEG, and DNN neural representational similarity. *The Journal of Neuroscience, 44*(12).

Zwaan, R. A., Stanfield, R. A., & Yaxley, R. H. (2002). Language comprehenders mentally represent the shapes of objects. Psychological Science, 13(2), 168-171.