Mind captioning: Evolving descriptive text of mental content from human brain activity

Tomoyasu Horikawa[1*]

[1] NTT Communication Science Laboratories, Kanagawa, Japan.

* Correspondence: horikawa.t@gmail.com

Running title: Evolving descriptions of mental content from the brain

Summary

A central challenge in neuroscience is decoding brain activity to uncover the mental content comprising multiple components and their interactions. Despite progress in decoding language-related information from human brain activity[1–4], generating comprehensive descriptions of intricate and structured mental content linked to visual semantics remains elusive[5–12]. Here, we present a method that progressively generates descriptive text mirroring brain representations via semantic features computed by a deep language model. We constructed linear decoding models to decode brain activity, measured by functional magnetic resonance imaging (fMRI) while subjects viewed videos, into semantic features of corresponding video captions. We then iteratively optimized candidate descriptions by aligning their semantic features with the brain-decoded features through word replacement and interpolation. This process resulted in the evolution of increasingly well-structured descriptions that faithfully captured the viewed content. Remarkably, comprehensible descriptions were generated even when excluding the fronto-temporal language areas from the analysis, highlighting explicit representations of structured semantic information outside the typical language network. Additionally, our method generalized to generate descriptions of imagined content, providing a means to interpret intricate mental content by translating brain signals into linguistic descriptions. These findings pave the way for non-verbal thought-based brain-to-text communication, potentially aiding individuals facing difficulties in language expression.

Introduction

Humans can recognize and recall intricate visual content comprising multiple semantic components, including objects, places, actions, and events, along with their interactions and relationships. These elaborate and structured mental representations form the foundation for translating thoughts into language and communicating experiences. Recently, substantial progress has been made in brain decoding of language-related information, enabling the direct production of linguistic outputs, such as text, from the human brain[1–4]. However, decoding the perceptual—and not to mention mental—content associated with visual semantics to generate comprehensive descriptions of subjective experiences remains challenging. Translating brain activity linked to non-linguistic semantic information, or thoughts, into verbal descriptions could significantly enhance our ability to interpret diverse mental states, opening up numerous possibilities for applications, particularly with text-prompt-based systems (e.g., ChatGPT[13] and Gemini[14]), as well as for scientific research.

Prior research on decoding visual semantics using human fMRI data has focused on individual components or static images. This narrow focus made it difficult to decode complex content involving interactions between multiple elements, obscuring how the brain represents rich and structured visual semantics. While studies have decoded individual components in viewed[5–7], imagined[5], and dreamed[8] content using object- or word-level features, they fall short of capturing neural representations of interactions and relationships[15], which are crucial for recognizing actions and social interactions[16–20]. Some researchers have incorporated caption databases[9] or deep neural network (DNN)-based modules, such as non-linear image captioning models[10–12], to produce linguistically structured outputs. However, these approaches can suffer from dependencies on specific resources or overfitting, potentially restricting their ability to generate flexible and precise descriptions of diverse content. Moreover, outputs from non-linear methods may spuriously reflect information not explicitly represented in the brain, rendering them unsuitable for inferring brain functions through decoding techniques[21–23].

To overcome these limitations, we introduce a new generative decoding method called mind captioning, which generates descriptive text mirroring the semantic information represented in the brain (Fig. 1). Our method combines linear feature decoding analysis[5,6,9], using semantic features computed by a deep language model (LM), with a novel optimization method that generates text based on these features. Semantic features serve as intermediate representations for decoding (translating) semantic information from the brain into text. They can act as a bridge for decoding both perceptual and mental content, as shared representations exist between visual perception and mental imagery, particularly for high-level information[5,8,24–26].

Additionally, deep LMs offer the advantage of effectively capturing contextual meanings, which are crucial for delineating intricate interrelationships[27–32].

The challenge lies in linguistically interpreting the information in semantic features decoded from brain activity. Although the ideal approach would be examining all possible word sequences to identify the best description whose semantic features match the decoded features, it is not feasible because the number of candidates is infinite. We thus developed an iterative optimization method that generates descriptive text from scratch by progressively aligning the semantic features of candidate descriptions with the target brain-decoded features through word replacement and interpolation, searching for the best description (Fig. 1b). Crucially, we leveraged an LM pre-trained for masked language modeling (MLM)[33] to constrain the search space during optimization. By directly optimizing word sequences to match brain-decoded features, our method is less susceptible to potential biases from external resources or additional modules.

To demonstrate the effectiveness of our method, we first validated our method for perceptual content by constructing decoding models (decoders) from stimulus-induced brain activity and then tested their generalizability to activity during mental imagery. Specifically, we measured brain activity in six subjects using fMRI while they viewed or imagined video clips[34] (Extended Data Fig. 1), training decoders to predict semantic features computed from corresponding video captions using an LM (DeBERTa-large)[35]. These decoders were then used to translate brain activity of viewed and imagined content into semantic features using novel test videos not used during training. Finally, we used the decoded features to optimize the text by employing an MLM model (RoBERTa-large)[36]. To enhance the quality of the fMRI data, we averaged brain activity over five repetitions for each stimulus or imagery in the test phase. See Extended Data Fig. 2 for details on the model and parameter validations.

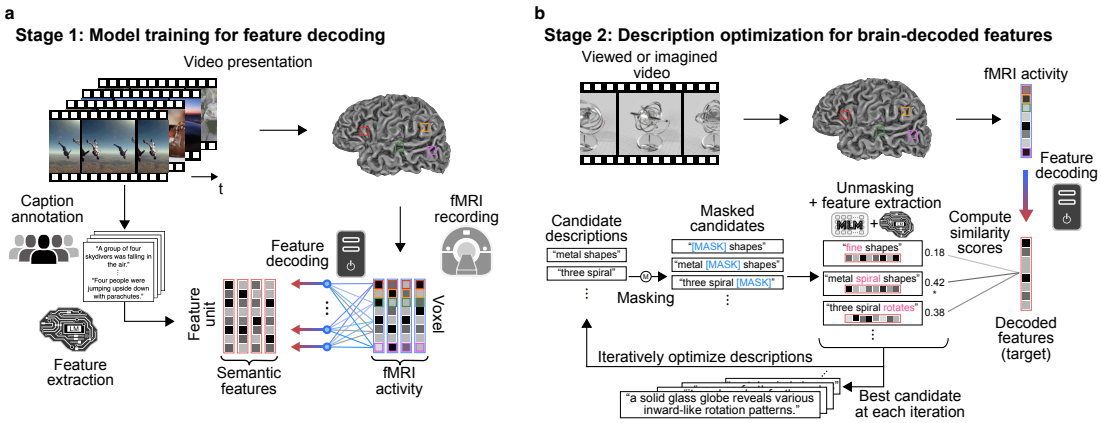


Fig. 1 Mind captioning. Our method consisted of two stages. **a**, In the first stage, we trained linear decoding models (decoders) to decode the whole-brain fMRI activity, measured while each subject viewed video clips, into semantic features computed from the captions of the viewed videos using an LM (frozen). **b**, In the second stage, we used the trained decoders to decode stimulus- or imagery-induced fMRI activity for novel video stimuli. We then optimized candidate descriptions iteratively by aligning their semantic features with the brain-decoded features through word replacement and interpolation, leveraging another LM pre-trained for masked language modeling (MLM; frozen). The optimization process consisted of three stages: masking, unmasking, and candidate selection. In the masking stage, we randomly applied masks by either replacing words with a mask token or interpolating mask tokens into the candidate word sequences. In the unmasking stage, the MLM model was used to create new candidates by filling in the masks in the masked candidates based on the context of surrounding words. In the candidate selection stage, we computed the semantic features of all new and original candidates using an LM for feature extraction. We then evaluated the similarity based on Pearson correlation coefficients between those candidate features and the target brain-decoded features to score the candidate descriptions and select the top candidates for further optimization. These optimization stages were repeated 100 times. We initiated the optimization process from a non-informative word ("<unk>") to avoid incorporating any prior assumptions for description generation.

Generating viewed content descriptions

The optimization of text, based on decoded features from stimulus-induced brain activity, resulted in a progressive evolution of descriptive texts (Fig. 2a). Initially, the descriptions were fragmented and lacked clear meaning. However, through iterative optimization, these descriptions naturally evolved to have a coherent structure and effectively capture the key aspects of the viewed videos. Notably, the resultant descriptions accurately depicted the content, including the dynamic changes in the viewed events (Fig. 2b). Additionally, even when specific objects were not correctly identified, the descriptions still successfully conveyed the presence of interactions among multiple objects (e.g., Fig. 2b bottom left).

Throughout the optimization process, the features of the evolved descriptions exhibited increasingly stronger correlations with the target features and, consequently, with the features of the reference captions annotated on the viewed videos (Fig. 2c). In contrast, the same optimization performed with an untrained MLM model, which randomly suggests candidate words during unmasking—a process analogous to a genetic algorithm—did not yield comparable results (Extended Data Fig. 3b). These results emphasize the significance of contextual information from a pre-trained MLM model to efficiently explore descriptions closely aligned with brain representations and to enhance description quality.

To assess decoding performance, we used multiple similarity metrics to evaluate the similarity between the generated descriptions and the reference captions. We calculated similarity scores for both the captions of the viewed (correct) and irrelevant (incorrect) videos and defined discriminability as the difference between them. The generated descriptions showed significantly high discriminability across all metrics and subjects (Fig. 2d; Wilcoxon signed-rank test, one-tailed, $p < 0.01$, FDR corrected across metrics and subjects), indicating that these descriptions were accurate enough to differentiate video content.

To gain an intuitive understanding of the performance, we evaluated the accuracy of video identification by comparing the generated descriptions to both the correct and incorrect captions across various numbers of candidates. The performance consistently exceeded chance levels for all set sizes, with approximately 50% accuracy among 100 distinct possibilities for all subjects when using feature correlation (chance = 1%; Fig. 2e), demonstrating the effectiveness of our method in translating detailed information from the brain into text through semantic features.

To further highlight the effectiveness of our method, it is worth noting that it surpasses the previous approach and holds promise for future improvements. Our method outperformed the

database search method[9] in capturing subjective perception more accurately and flexibly (Extended Data Fig. 5). Additionally, we showed that our method robustly generated descriptions that faithfully reflected the viewed content, regardless of the LMs employed, and found correlations between brain encoding performance—a metric assessing alignment between the brain and models[30,37]—and the text generation performance of the LMs (Extended Data Fig. 6). These results suggest that employing LMs more closely resembling the brain (e.g., GPT3[13], OPT[38], and LLaMA[39]) could further improve the effectiveness of our method.

Capturing visual relational information

A key advantage of structured descriptions over word lists is their flexibility in organizing words to convey contextual meaning, including relationships between components[40]. To investigate whether our generated descriptions accurately captured visual relational information, we examined the effect of shuffling their word order—either for all words or nouns only, reasoning that if the original word order accurately conveyed visual relationships, shuffling would reduce discriminability. While the shuffled descriptions retained reasonably high accuracy in identifying videos, indicating that simple word lists can provide informative cues, shuffling all words and even only nouns significantly impaired discriminability (Fig. 2f; Wilcoxon signed-rank test, one-tailed, $p < 0.01$, FDR corrected across subjects). These results demonstrate that our method can generate descriptions that capture more detailed information than a simple word list. Additionally, the effect of shuffling was more pronounced when using features from deeper LM layers to generate descriptions (Extended Data Fig. 7), underscoring the importance of the deep structure of the LMs in constructing contextual semantic representations and accurately decoding relational information.

Importantly, we confirmed that the coherent structure of the generated descriptions, crucial for depicting visual relations, was not artificially created by the MLM model. Specifically, the shuffling of word orders significantly reduced the feature correlations between the descriptions and the target brain-decoded features, especially when the shuffling substantially altered the original meanings (Fig. 2g). This suggests that the generated word sequences were optimized to authentically reflect the semantic information in the brain. Additionally, when decoders trained with semantic features from word-order shuffled captions were used—a method designed to prevent the decoders from learning structured semantics—incoherent word sequences were generated (Extended Data Fig. 3c). These results suggest that the generation of coherent descriptions is not solely reliant on the guidance of the MLM model.

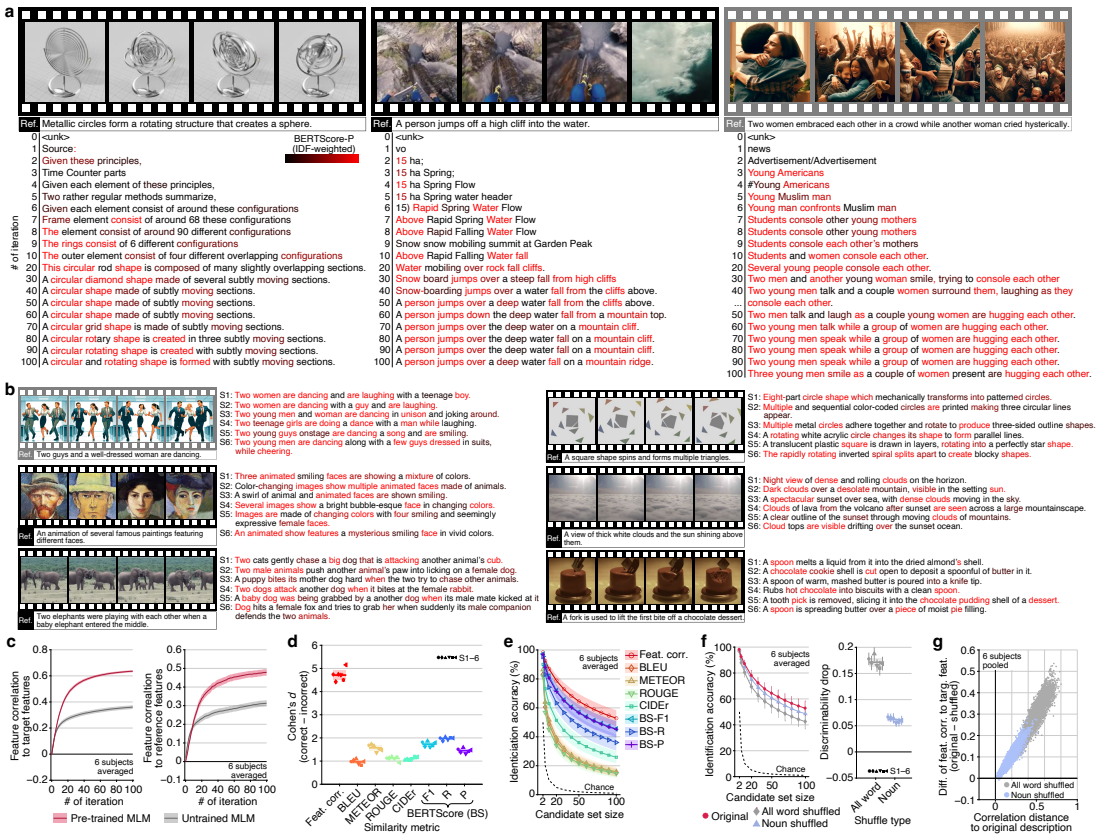


Fig. 2 | Generation of descriptive texts of viewed content.

[Note: To comply with bioRxiv's policy on displaying human faces, frames containing real human faces in this figure (framed by gray) have been replaced with images synthesized by DALL·E3 (https://openai.com/dall-e-3), based on captions annotated by humans for the video.]

Descriptions were generated using features from all layers of the DeBERTa-large model (24 layers) decoded from whole-brain activity. **a**, Evolved descriptions during the optimization process. **b,** Descriptions generated after 100 optimization iterations for all subjects (Extended Data Fig. 3a for more examples). In **a**, **b**, the color indicates precision-based accuracy (BERTScore-P weighted by the inverse of document frequency, IDF). The text below frames shows a reference caption of the video. **c**, Feature correlations between semantic features of generated descriptions and those decoded from the brain (left), as well as those computed from correct references (right) through optimization. **d,** Cohen's *d* of discriminability. Results are shown as effect size to allow for comparison across metrics (Extended Data Fig. 4b for raw scores). **e,** Video identification accuracy with varying numbers of candidates. **f,** Effects of word-order shuffling on video identification accuracy and discriminability. The feature correlation was used as the similarity metric. **g,** Scatterplot of the correlation distances (one minus feature correlation) between the original and shuffled descriptions against the difference in feature correlations to target features between the original and shuffled descriptions Each dot indicates a result of a shuffled description. Shades in **c**, **e**, and error bars in **d**, **f** indicate 95% confidence intervals (C.I.) across samples (n = 72). Shades in **d, f** indicate 95% C.I. across subjects (n = 6). For **c**, **e**, **f**, and **g**, see Extended Data Fig. 4a, c, d, and f for individual results.

Contributions from different brain areas

Having confirmed the ability to generate accurate and well-structured descriptions from whole-brain activity, we next investigated the contribution of different brain regions to this decoding, with a particular focus on the involvement of the typical language network in representing intricately structured visual semantics, a core component of thought[41–43]. To begin, we analyzed how the brain encodes semantic information in videos by constructing two encoding models: one based on the semantic features used in our decoding analysis and the other based on visual features from a video recognition DNN[44]. The semantic encoding model effectively predicted brain activity in the language-related areas, as well as regions involved in object, action, and interaction recognition in the higher visual cortex (HVC) and parietal cortex (Fig. 3a)[16–20]. This model progressively outperformed the visual encoding model from the lower visual cortex (LVC) through the HVC and language network (Fig. 3b−e). Furthermore, we found that while voxels in the HVC and the language network were better predicted by features from deeper layers of the visual DNN and the LM (Fig. 3f), the LM layers yielding the highest performance for language network voxels were much deeper than those for the HVC (Wilcoxon rank-sum test, one-tailed, $p < 0.01$, FDR corrected across subjects), highlighting a significant link between the language network and contextualized semantic information. These results demonstrate the involvement of the language network, along with distributed brain regions, in encoding contextual semantics, consistent with previous studies on its activation by non-verbal visual semantics[42,43].

We then assessed decoding performance using video identification analysis based on descriptions generated from these brain regions with varying selectivity (Fig. 3g). Focusing on voxels more accurately predicted by the semantic model, we achieved higher performance, approaching the results from whole-brain activity for some subjects (Extended Data Fig. 4g). Conversely, voxels better predicted by the visual model showed weaker performance, indicating a lower contribution from the posterior parts of the visual cortex. Using only the language network, despite its involvement in encoding contextual semantics, did not yield high performance, contrary to previous studies on linguistic semantic decoding[3]. Remarkably, ablating the language network did not profoundly impact performance, achieving almost 50% accuracy from 100 possibilities (chance = 1%; six subjects averaged), showing a decrease in accuracy and discriminability by word order shuffling (Fig. 3h; Wilcoxon rank-sum test, one-tailed, $p < 0.01$, FDR corrected across subjects), and even generating intelligible descriptions (Extended Data Fig. 3d). These results suggest that we can still generate accurate descriptions reflecting structured semantics without relying on the language network, indicating that structured semantic information exists in regions extending from the anterior portions of the

visual cortex to areas outside the language network. The representations in these regions may support the comprehension of complex visual events by individuals with global aphasia[42], providing insight into how the brain represents complex non-verbal semantics.

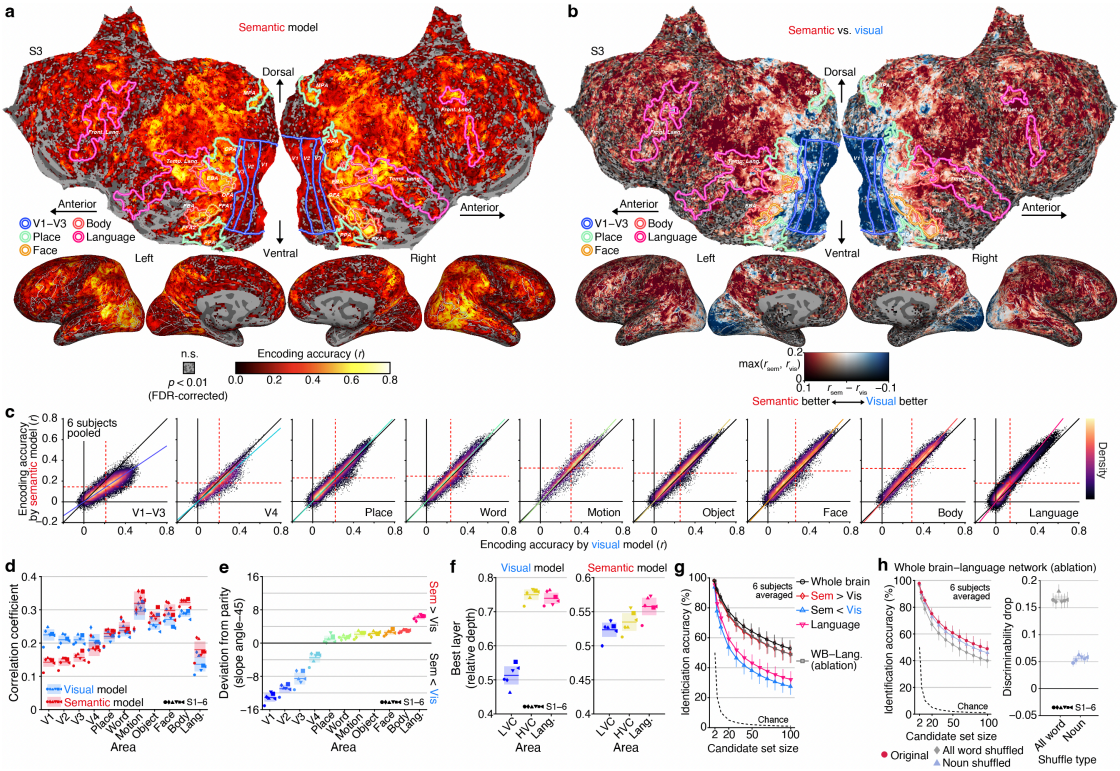


Fig. 3 Contributions from different brain areas. The encoding analysis was conducted using cross-validation within the training data that contained fMRI responses to diverse videos (n = 2,108). Encoding models were trained using features from each layer to generate predictions from multiple layers. The final predictions were then constructed by selecting the best layer per voxel based on performance on the training data using nested cross-validation. **a**, Encoding performance of the semantic model. **b**, Differences in performance between the semantic and visual models. **c**, Density heatmap showing the encoding accuracy (six subjects pooled; solid lines, best linear fit; dashed lines, mean accuracy). **d,** Mean encoding accuracy within each functional area. **e**, Deviation from parity based on the slope angles of the best linear fit. **f**, Mean of the best layers. The indices of layers with the highest performance were averaged across voxels in each area and then converted into relative depth. The LVC and HVC encompassed V1−V3 and category-selective areas, respectively. **g**, Video identification accuracy from different brain areas. Decoders were trained using voxels with higher encoding accuracy according to the semantic or visual models, voxels within the language network, or voxels from the entire brain except for the language network. **h**, Effects of word-order shuffling on video identification accuracy and discriminability when not using the language network. In **d**−**h**, error bars and shades indicate 95% C.I. across samples and subjects, respectively. For **a** and **b**, **g**, and **h**, see Extended Data Fig. 8, 4g, and e for individual results.

Generating imagined content descriptions

Finally, we investigated whether the decoders trained on stimulus-induced brain activity could be used to generate descriptions of imagined content from activity induced by mental imagery, using the same evaluation procedures as for the perception data analysis. The analysis successfully generated descriptions that accurately reflected the content of the imagined videos, although accuracy varied among individuals (Fig. 4a). These descriptions were more similar to the captions of the imagined videos than to irrelevant ones (Extended Data Fig. 9a and b), with proficient subjects achieving nearly 40% accuracy in identifying imagined videos from 100 candidates (Fig. 4b; chance = 1%). Shuffling word order resulted in a notable drop in video identification accuracy and discriminability (Fig. 4c; Wilcoxon signed-rank test, one-tailed, $p < 0.01$, FDR corrected across subjects). Excluding the language network from the analysis slightly, but not substantially, decreased video identification accuracy (Fig. 4d). Taken together, these results confirm the generalizability of the decoders to generate descriptions of imagined content, showcasing its ability to effectively verbalize imagined content even without relying on the language network. Furthermore, we were able to obtain comprehensible descriptions of the imagined content with reasonable identification accuracy from single-trial fMRI activity (Fig. 4e and Extended Data Fig. 10), demonstrating the potential applicability of our method to experiences that are difficult to reproduce, such as dreams[25].

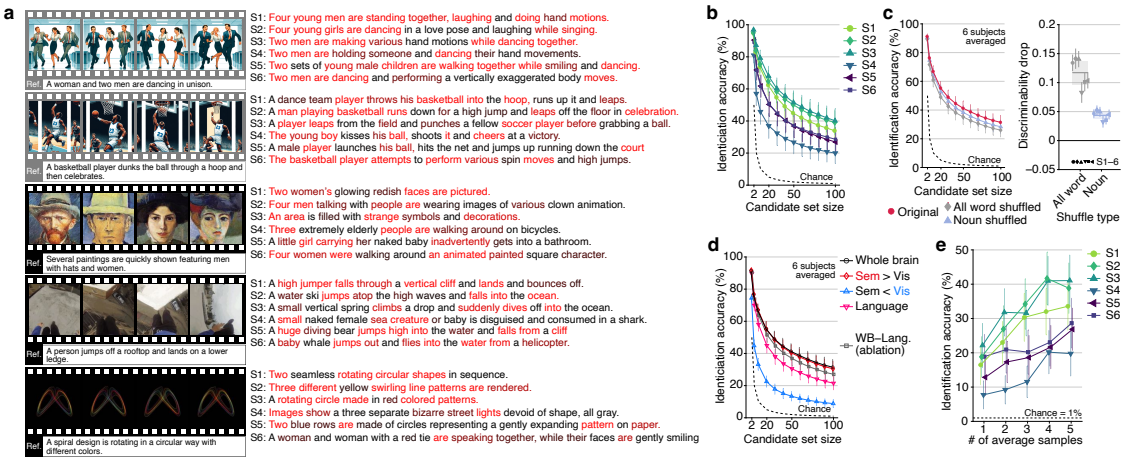


Fig. 4 Generation of descriptive text of imagined content.
[Note: To comply with bioRxiv's policy on displaying human faces, frames containing real human faces in this figure (framed by gray) have been replaced with images synthesized by DALL·E3 (https://openai.com/dall-e-3), based on captions annotated by humans for the video.]
We used the decoders trained on stimulus-induced brain activity (DeBERTa-large; decoded from the whole-brain activity, unless otherwise stated) to analyze the brain activity of subjects engaged in visual mental imagery (recall) of videos based on verbal descriptions (Extended Data Fig. 1b). The feature correlation was used as the similarity metric. **a**, Descriptions of imagined content generated after 100 optimization iterations for all subjects (see Extended Data Fig. 3e for more examples). **b**, Identification accuracy of imagined videos from individual subjects. **c**, Effects of word-order shuffling on identification accuracy and discriminability of imagined videos. **d**, Video identification accuracy of imagined videos from different brain areas. **e**, Video identification accuracy of imagined videos with a varying number of averaged samples. For **a**, **c**, and **d**, conventions are the same with Fig. 2b, f, and Fig. 3g. See Extended Data Fig. 9 and 10 for more data.

Discussion

We have successfully generated descriptive text representing perceived and imagined visual content by aligning semantic features of text with those decoded from human brain activity. Our success is attributed to two key factors: the advancement of deep LMs that provide contextualized semantic representations similar to those in the human brain[30,45] and our innovative text optimization method utilizing the MLM model for word-level optimization while efficiently constraining the search space. Together, these factors facilitate the direct translation of brain representations into text, resulting in faithful descriptions of visual semantic information in the brain. These descriptions were well structured, accurately capturing individual components and their relationships without using the language network, thus highlighting the explicit representation of structured semantic information outside this network. Our method enables the intelligible interpretation of internal thoughts, demonstrating the feasibility of non-verbal, thought-based brain-to-text communication.

While our method enables generating linguistic outputs through brain decoding, it differs from previous language decoding attempts[1–4] as it does not rely on brain activity associated with language processing. It successfully generated English descriptions even from our non-native English-speaking subjects, without involving the language network, demonstrating that our method can directly translate non-linguistic semantics in the brain into linguistic descriptions, regardless of the subject's native language or language proficiency. Therefore, this method can be applied to decode brain activity in non-linguistic subjects, including infants and animals, potentially providing insights into how they develop the neural basis for processing complex visual semantics. Additionally, it could assist individuals with language expression difficulties, such as aphasia.

A limitation of this study is its reliance on video captions provided by independent human annotators. These captions may not align with our subjects' perceptions and may contain ambiguities and inaccuracies, potentially affecting the decoding performance. Furthermore, we instructed the annotators to focus on the visual content rather than their emotional reactions[46]. As a result, the generated descriptions were predominantly concrete and rarely captured abstract aspects such as the subjects' impressions and emotions. Depending on how well the annotations accurately reflect and encompass various dimensions of subjective experience, our method may capture the content of our mind more comprehensively.

Our method has successfully generated descriptions of imagined content, even from single-trial fMRI activity. However, this success raises ethical concerns regarding potential invasions of

privacy. Key issues include the risk of unintentionally disclosing primitive thoughts before the individual has chosen to verbalize them. Additionally, unwanted biases inherent in the LMs[47] could distort the results within feasible optimization limits. Moreover, although requiring intensive data collection from willing participants may currently ensure consent[3], advances in inter-individual alignment technology could reduce this requirement[48,49]. Therefore, it is imperative to establish regulations that promote the ethical use of these technologies[50] while ensuring explicit informed consent and safeguarding subjects' mental privacy and autonomy in deciding which thoughts to disclose.

Data availability

All experimental data that support the findings of this study will be made publicly available at OpenNeuro before publication.

Code availability

All code that supports the findings of this study will be made publicly available on GitHub before publication.

Acknowledgments

The author thanks Yuki Honda for assistance with MRI data collection and cleaning caption data, Scinob Kuroki, Shimpei Yamagishi, Hiromi Narimatsu, and Yuta Suzuki for assistance with scanner operation, Mitsuaki Tsukamoto, Misato Tanaka, and Yukiyasu Kamitani for assistance with preliminary investigations, and Shogo Kajimura for valuable discussions. We appreciate the support of the WPI-IRCN Human fMRI Core, the University of Tokyo Institutes for Advanced Studies. This research was supported by grants from JST PRESTO Grant Number JPMJPR185B Japan and JSPS KAKENHI Grant Number JP21H03536.

Author contributions

Conceptualization, T.H.; Methodology, T.H.; Validation, T.H.; Formal Analysis, T.H.; Investigation, T.H.; Resources, T.H.; Writing – Original Draft, T.H.; Visualization, T.H.; Funding Acquisition, T.H..

Competing interests

The author declares no competing interests.

Correspondence and requests for materials should be addressed to Tomoyasu Horikawa.

References

1. Défossez, A., Caucheteux, C., Rapin, J., Kabeli, O., & King, J. Decoding speech perception from non-invasive brain recordings. *Nat. Mach. Intell.* **5,** 1097–1107 (2023).
2. Metzger, S. L. et al. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature* **620,** 1037–1046 (2023).
3. Tang, J., LeBel, A., Jain, S. & Huth, A. G. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nat. Neurosci*. **26,** 858–866 (2023).
4. Willett, F. R. et al. A high-performance speech neuroprosthesis. *Nature* **620,** 1031–1036 (2023).
5. Horikawa, T. & Kamitani, Y. Generic decoding of seen and imagined objects using hierarchical visual features. *Nat. Commun.* **8**, 15037 (2017a).
6. Huth, A. G. et al. Decoding the Semantic Content of Natural Movies from Human Brain Activity. *Front. Syst. Neurosci.* **10**, 81 (2016).
7. Nishida, S., & Nishimoto, S. Decoding naturalistic experiences from human brain activity via distributed representations of words. *NeuroImage*, **180** 232–242 (2018).
8. Horikawa, T., & Kamitani, Y. Hierarchical Neural Representation of Dreamed Objects Revealed by Brain Decoding with Deep Neural Network Features. *Front. Comput. Neurosci.* **11**, 4 (2017b).
9. Doerig, A. et al. Semantic scene descriptions as an objective of human vision. Preprint at https://arxiv.org/abs/2209.11737 (2022).
10. Ferrante, M., Ozcelik, F., Boccato, T., Rullen, R.V. & Toschi, N. Brain Captioning: Decoding human brain activity into images and text. Preprint at https://arxiv.org/abs/2305.11560 (2023).
11. Huang, W. et al. A dual-channel language decoding from brain activity with progressive transfer training. *Hum. Brain map.* **42**, 5089–5100 (2021a).
12. Huang, W. et al. A neural decoding algorithm that generates language from visual activity evoked by natural images. Neural Netw. **144**, 90–100 (2021b).
13. Brown, T.B. et al. Language Models are Few-Shot Learners. Preprint at https://arxiv.org/abs/2005.14165 (2020).
14. Anil, G.T. et al. Gemini: A Family of Highly Capable Multimodal Models. Preprint at https://arxiv.org/abs/2312.11805 (2023).
15. Baldassano, C., Beck, D. M. & Fei-Fei, L. Human-Object Interactions Are More than the Sum of Their Parts. *Cereb. Cortex* **27**, 2276–2288 (2017).
16. Karakose-Akbiyik, S., Caramazza, A. & Wurm, M. F. A shared neural code for the physics of actions and object events. *Nat. Commun.* **14**, 3316 (2023).
17. McMahon, E. & Isik, L. Seeing social interactions. *Trends Cogn. Sci.* **27**, 1165–1179 (2023).

18. Tarhan, L. & Konkle, T. Sociality and interaction envelope organize visual action representations. *Nat. Commun.* **11**, 3002 (2020).
19. Wurm, M. F., Caramazza, A. & Lingnau, A. Action categories in lateral occipitotemporal cortex are organized along sociality and transitivity. *J. Neurosci*. **37**, 562–575 (2017).
20. Wurm, M. F. & Caramazza, A. Two 'what' pathways for action and object recognition. *Trends Cogn. Sci.* **26**, 103–116 (2022).
21. Kamitani, Y. & Tong, F. Decoding the visual and subjective contents of the human brain. *Nat. Neurosci.* **8**, 679–685 (2005).
22. Kriegeskorte, N. & Kievit, R. A. Representational geometry: integrating cognition, computation, and the brain. *Trends in Cogn. Sci.* **17**, 401–412 (2013).
23. Ritchie, J. B., Kaplan, D. M. & Klein, C. Decoding the Brain: Neural Representation and the Limits of Multivariate Pattern Analysis in Cognitive Neuroscience. *Br. J. Philos. Sci.* **70**, 581–607 (2019).
24. Dijkstra, N., Bosch, S. E. & van Gerven, M. A. J. Shared Neural Mechanisms of Visual Perception and Imagery. *Trends Cogn. Sci.*, **23**, 423–434 (2019).
25. Horikawa, T., Tamaki, M., Miyawaki, Y. & Kamitani, Y. Neural decoding of visual imagery during sleep. *Science* **340**, 639–642 (2013).
26. Pearson, J., Naselaris, T., Holmes, E. A. & Kosslyn, S. M. Mental Imagery: Functional Mechanisms and Clinical Applications. *Trends Cogn. Sci.* **19**, 590–602 (2015).
27. Caucheteux, C. & King, J.-R. Brains and algorithms partially converge in natural language processing. *Commun. Biol.* **5**, 134 (2022).
28. Jain, S. & Huth, A. G. Incorporating context into language encoding models for fMRI. *NeurIPS* **31**, 6629–6638 (2018).
29. LeBel, A., Jain, S. & Huth, A. G. Voxelwise encoding models show that cerebellar language representations are highly conceptual. *J. Neurosci.* **41**, 10341–10355 (2021).
30. Schrimpf, M. et al. The neural architecture of language: integrative modeling converges on predictive processing. *Proc. Natl. Acad. Sci. USA* **118**, e2105646118 (2021).
31. Toneva, M. & Wehbe, L. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *NeurIPS*, **32** 14928–14938 (2019).
32. Toneva, M., Mitchell, T. M. & Wehbe, L. Combining computational controls with natural text reveals aspects of meaning composition. *Nat. Comput. Sci.* **2**, 745–757 (2022).
33. Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *In Proc. Conference of the North American Chapter of the Association for Computational Linguistics* **1**, 4171–4186 (2019).
34. Cowen, A. S. & Keltner, D. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proc. Natl. Acad. Sci. USA* **114**, E7900–E7909 (2017).

35. He, P., Liu, X., Gao, J. & Chen, W. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. Preprint at https://doi.org/10.48550/arXiv.2006.03654 (2020).
36. Liu, Y. et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. Preprint at https://doi.org/10.48550/arXiv.1907.11692 (2019).
37. Schrimpf, M. et al. Brain-Score:Which Artificial Neural Network for Object Recognition is most Brain-Like? Preprint at https://www.biorxiv.org/content/10.1101/407007v1 (2018).
38. Zhang, S. et al. OPT: Open Pre-trained Transformer Language Models. https://arxiv.org/abs/2205.01068 (2022).
39. Touvron, H. et al. LLaMA: Open and Efficient Foundation Language Models. Preprint at https://arxiv.org/abs/2302.13971 (2023).
40. Thrush, T. et al. Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality. *CVPR*, 5228–5238 (2022).
41. Fedorenko, E. & Varley, R. Language and thought are not the same thing: evidence from neuroimaging and neurological patients. *Ann. N. Y. Acad. Sci.* **1369**, 132–153 (2016).
42. Ivanova, A. A. et al. The Language Network Is Recruited but Not Required for Nonverbal Event Semantics. *Neurobiol. Lang. (Camb)* **2**, 176–201 (2021).
43. Sueoka, Y., Paunov, A. M., Ivanova, A. A., Blank, I. A. & Fedorenko, E. The language network reliably 'tracks' naturalistic meaningful non-verbal stimuli. Neurobiol. Lang. (2024).
44. Bertasius, G., Wang, H. & Torresani, L. Is Space-Time Attention All You Need for Video Understanding? Preprint at https://arxiv.org/abs/2102.05095 (2021).
45. Antonello, R., Vaidya, A.R. & Huth, A.G. Scaling laws for language encoding models in fMRI. Preprint at https://arxiv.org/abs/2305.11863 (2023).
46. Achlioptas, P., Ovsjanikov, M., Haydarov, K., Elhoseiny, M. & Guibas, L. J. ArtEmis: Affective Language for Visual Art. *CVPR* 11564–11574 (2021).
47. Navigli, R., Conia, S. & Ross, B. Biases in Large Language Models: Origins, Inventory, and Discussion. ACM Journal of Data and Information Quality, **15**, 1–21 (2023).
48. Scotti, P. S. et al. MindEye2: Shared-Subject Models Enable fMRI-To-Image With 1 Hour of Data. Preprint at https://arxiv.org/abs/2403.11207 (2024).
49. Wang, H. et al. Inter-individual and inter-site neural code conversion and image reconstruction without shared stimuli. Preprint at https://arxiv.org/abs/2403.11517 (2024).
50. Goering, S. et al. Recommendations for Responsible Development and Application of Neurotechnologies. *Neuroethics* **14**, 365–386 (2021).

Methods

Subjects

Six healthy subjects with normal or corrected-to-normal vision participated in our experiments: Subject 1 (male, aged 37–38), Subject 2 (female, aged 37–38), Subject 3 (male, aged 35–36), Subject 4 (female, aged 35–36), Subject 5 (male, aged 29–30), and Subject 6 (male, aged 22–23). All subjects were non-native English speakers, and Subject 6 had limited proficiency in English. All subjects provided written informed consent, and the study protocol was approved by the Ethics Committee of NTT. The sample size was determined based on prior fMRI studies with similar behavioral protocols[5,51,] rather than relying on statistical methods. Data from each subject were collected over multiple scanning sessions spanning approximately six months. We determined the experimental parameters and analytical pipelines based on data from a preliminary experiment with Subject 1 using the same stimulus set. As a result, Subject 1 was exposed to the same stimuli multiple times, which could have influenced their brain responses.

Visual stimuli

Visual stimuli consisted of emotionally evocative short videos collected in a previous study[34]. The original video dataset comprised a total of 2,196 videos (https://goo.gl/forms/XErJw9sBeyuOyp5Q2) with durations ranging from approximately 0.15 s to around 90 s. These videos covered diverse visual content, including objects, scenes, actions, and events. Each video was resized to fit within a 16-degree visual angle while preserving the original aspect ratio. They were presented at the center of the display against a gray background (no sound was delivered, although some of the original videos included audio). To eliminate redundancy, we excluded 16 duplicate or similar videos, resulting in a final set of 2,180 unique videos for use in our experiment.

Experimental design

We conducted two types of main experiments: a video presentation experiment and an imagery experiment. All visual stimuli were presented on a liquid crystal display monitor (NNL MRI InroomViewingDevice) located at the rear of the fMRI scanner bore, and all audio stimuli (e.g., beep sounds in the imagery experiment) were delivered using S14 insert earphones (Sensimetrics). On each experimental day, one consecutive session was conducted for a maximum of 2 hours. Subjects were given adequate time for rest between runs (every 8–10 minutes) and were allowed to take a break or stop the experiment at any time. The total duration of the video presentation experiment and imagery experiment was approximately 17.1 hours (averaged across runs and subjects).

Video presentation experiment. The video presentation experiment consisted of two distinct types of sessions: training sessions and test sessions, each comprising 60 and 10 runs, respectively. Each run included 36 or 37 stimulus blocks and randomly inserted 3 or 4 additional evaluation blocks. The duration of each run varied depending on the length of the video presented and the duration of the evaluations (695.2 s per run, averaged across runs and subjects). In total, each subject spent approximately 11.8 hours on training sessions and 1.8 hours on test sessions.

In each stimulus block where videos shorter than 10 s were presented, the same video stimulus was repeated until the total presentation duration exceeded 10 s. In cases where stimulus blocks contained videos longer than 10 s, the video stimulus was presented once, followed by less than 1 s of rest, allowing the total durations of stimulus blocks to be divisible by 1 s (TR). To facilitate subjects to recognize any details of the visual content in the videos, they were permitted to view stimuli without fixation.

In each evaluation block, five descriptions (both in English and their translations into Japanese) depicting the visual contents of videos were presented to ask subjects to rate the degree to which each description was consistent with what they had perceived in the preceding stimulus block. Those five descriptions were from twenty captions collected for the video presented immediately before the evaluation block (see Material and Methods: "Caption annotation for visual stimuli"), but occasionally containing one from those for another video. Subjects held two button boxes, each with two buttons, in their hands and used the right and left button boxes to select and rate sentences, respectively. They were required to assign a score on a five-point scale to a description that appeared to be associated with the preceding video (with higher scores indicating closer alignment with their perception; random initial score), while they were also allowed to assign "x" if they detected a description that they thought was from another video. The rating was performed at the subjects' own pace. After completing the rating, subjects could proceed to the next block by selecting the "Proceed?" sentence located at the bottom of the other descriptions and pressing a button on the left button box.

Each block was followed by an additional 2-s rest period, and additionally 32- and 8-s rest periods were added to the beginning and end of each run, respectively. During rest periods, subjects were requested to maintain a steady fixation on the central fixation spot, which comprised a combination of a bull's eye and crosshairs, with outer and inner circles measuring 0.3/0.05 degrees, respectively. This instruction aimed to sustain their attention on the screen.

In the training session, a total of 2,180 unique videos were each presented only once. The presentation orders of the videos and evaluation blocks were pseudo-randomized throughout the

entire training session, and this order remained consistent for all subjects. Furthermore, the descriptions used for evaluations also remained consistent for all subjects. In the test session, 72 videos used in the last two runs of the training session were each presented five times (see Methods "MRI data preprocessing" for the data handling of the overlapping data in the training session). These 72 videos were randomly divided between two runs, ensuring that both consecutive runs contained all videos, and they were presented in a pseudo-random order within each run.

Imagery experiment. The imagery experiment comprised a total of 30 runs, each of which included 12 trials. Each trial consisted of a preparation block, an imagery block, a video presentation block, and an evaluation block. Subjects were required to visually imagine (recall) one of the 72 videos presented in the test session of the video presentation experiment. The duration of each run varied depending on the length of the preparation blocks and the imagined and presented videos (434.6 s per run, averaged across runs and subjects). In total, each subject spent approximately 3.6 hours on the imagery experiment.

During a preparation block, a verbal description (both in English and their translations into Japanese) of a target video was presented to subjects to prepare mental imagery of the visual content of the target. Descriptions for target videos were selected from each set of twenty captions collected for each video (see Material and Methods: "Caption annotation for visual stimuli"), and a single caption was used for each video consistently across trials and subjects. To enhance the similarity between the visual experiences in the subsequent imagery block and those during perception in the video presentation experiment, subjects were encouraged to prepare to imagine all details, even those not explicitly described in the description. The description served merely as a guide, enabling subjects to vividly imagine the complete visual content of the target, and their mental imagery did not need to be constrained by the specific details of the descriptions. Subjects were asked to press a button after they got ready to start an imagery period, and the onset was notified by a beep sound with less than 1 s of rest after their button press to align the onset of the imagery period and TR.

During the imagery period, subjects were required to imagine the visual content of the target video with their eyes closed as if they were actually watching the video. The duration of the imagery block was set to match the length of the stimulus blocks for the same target videos in the video presentation experiment (i.e., recalling visual scenes repeatedly for videos shorter than 10 s or recalling them only once for videos longer than 10 s; see Methods: "Video presentation experiment"). An additional 2 s was added to these durations to allow subjects to fully replay

467 the entire scenes of the target videos. The end of the imagery period was notified by another
468 beep sound, prompting subjects to open their eyes.

469 Following the imagery block, a stimulus block followed, and the target video was presented in
470 the same manner as the video presentation experiment. This block allowed subjects to assess the
471 fidelity of their mental imagery against the actual video content. The stimulus block was then
472 followed by two sequential 3-s blocks (one for accuracy and the other for vividness), in which a
473 score was presented at the center of the display, prompting subjects to evaluate the accuracy and
474 vividness of their mental imagery on a five-point scale (ranging from "very accurate/vivid" to
475 "cannot accurately/vividly imagine the target"). Subjects adjusted the score from its random
476 initial setting using the left and right buttons of a button box placed in their right hand.

477 Each of the imagery, stimulus, and evaluation blocks was followed by an additional 2-s rest
478 period, and additionally 32- and 8-s rest periods were added to the beginning and end of each
479 run, respectively. Subjects were requested to maintain a steady fixation on the central fixation
480 spot as in the video presentation experiment.

481 Prior to the imagery experiment, subjects underwent adequate practice to establish associations
482 between each target video and its corresponding verbal description. This practice aimed at
483 enabling them to recall a target video based on its corresponding description. Additionally, to
484 assist in remembering these associations, subjects were allowed to view the description and
485 video pairs during every inter-run-rest period.

486 In the imagery experiment, the trials for the 72 videos were randomly distributed among six
487 runs, ensuring that each set of six consecutive runs contained all the videos. Within each run,
488 the order of target videos was determined in a pseudo-random manner.

489 **Retinotopy and functional localizer experiments**. In addition to the main experiments, we
490 conducted a retinotopy experiment and three functional localizer experiments (a visual category
491 localizer, an MT+ localizer, and a language area localizer) to separately delineate visual areas
492 and to localize known regions of interest (ROIs) in each subject.

493 *Retinotopy*. The retinotopy experiment followed the protocol of the Human Connectome Project
494 7T Retinotopy Dataset[52] (eight 300-s runs), in which dynamic colorful textures were presented
495 through slowly moving apertures in various shapes (wedge, ring, and bar). The data were used
496 to delineate the borders between each visual cortical area and to identify the retinotopic map
497 (V1, V2, V3, V3A, V3B, V4, and V7) on the cortical surfaces of individual subjects based on

the results of the *fsfast* retinotopy analysis (Freesurfer[53]) and the population receptive field analysis (code is available at https://kendrickkay.net/analyzePRF/)[54].

Visual category localizer. The visual category localizer experiment followed the protocol of the fLoc functional localizer experiment[55] (stimuli and code are available at http://vpnl.stanford.edu/fLoc/; eight 300-s runs; 48 blocks, each consisting of a 6-s stimulus or blank, with initial and final 6-s rest periods for each run), in which visual images from word, body, face, place, and object categories were presented. The data were used to localize brain areas that selectively respond to these categories. Because object-selective areas were not clearly localized using the original stimulus set, we additionally included two stimulus types (intact and scrambled object conditions) using images from The Object Databank (original images are available at tarrlab, https://sites.google.com/andrew.cmu.edu/tarrlab/stimuli). The contrasts between word/body/face/place and others (from these four categories) were used to define visual category-selective areas (word: visual word form area [VWFA], occipital word form area [OWFA]; body: extrastriate body area [EBA], fusiform body area [FBA]; face: fusiform face area [FFA]; occipital face area [OFA]; place: parahippocampal place area [PPA], occipital place area [OPA], medial place area [MPA] consisted of the retrosplenial cortex [RSC] and parieto-occipital sulcus [POS]). The contrast between intact and scrambled objects was used to define an object-selective area (lateral occipital complex [LOC]).

MT+ localizer. The MT+ localizer experiment followed the protocol reported by Tootell et al. (1995[56]; four 232-s runs; 13 blocks, each consisting of a 12-s stimulus, with initial and final 12-s rest periods for each run), in which random dot stimuli were presented in three conditions of stimulus presentation (moving, dynamic, and static). The contrast between moving and dynamic/static conditions was used to define the visual motion area MT+.

Language area localizer. The language area localizer experiment followed the protocol reported by Fedorenko et al. (2010)[57] and Scott et al. (2017)[58] while being modified to include both visual and auditory stimuli in each run and to use Japanese stimuli (eight 358-s runs; 19 blocks, each consisting of an 18-s stimulus or 14-s blank, with 14-s initial and final rest periods for each run). Each run included two visual conditions (sentence or nonword-sequence conditions) and two auditory conditions (intact or degraded conditions) for the stimuli. In the visual conditions, subjects read sentences or nonword sequences, each of which consisted of a string of approximately 18 words or nonwords. In auditory conditions, subjects listened to intact or acoustically degraded versions of short auditory passages (excerpt from audiobooks). The contrast between sentence/intact and nonword/degraded conditions was used to define language-sensitive areas around temporal and frontal cortices.

532 Voxels from V1, V2, and V3 were combined as the lower visual cortex (LVC), voxels from
533 VWFA, OWFA, EBA, FBA, FFA, OFA, PPA, OPA, MPA, LOC, and MT+ were combined as
534 the higher visual cortex (HVC), and voxels from temporal and frontal language areas were
535 combined as the language network. Voxels overlapping with LVC were excluded from HVC.

536 **MRI acquisition**

537 MRI data were collected using a 3.0-Tesla Siemens MAGNETOM Prisma scanner located at
538 the WPI-IRCN Human fMRI Core, the University of Tokyo Institutes for Advanced Studies. An
539 interleaved T2*-weighted gradient-echo echo planar imaging (EPI) scan was performed to
540 acquire functional images covering the entire brain (TR, 1000 ms; TE, 30 ms; flip angle, 65
541 deg; FOV, 192 × 192 mm; voxel size, 2 × 2 × 2 mm; slice gap, 0 mm; number of slices, 72;
542 multiband factor, 6). T1-weighted (T1w) magnetization-prepared rapid acquisition gradient-
543 echo (MP-RAGE) fine-structural images of the entire head were also acquired (TR, 2000 ms;
544 TE, 1.97 ms; TI, 900 ms; flip angle, 10 deg; FOV, 256 × 256 mm; voxel size, 1.0 × 1.0 × 1.0
545 mm).

546 **MRI data preprocessing**

547 For anatomical data, we first used SPM12 to preprocess each of the T1w anatomical images of
548 individual subjects for the bias-field correction and for re-defining its origin and orientation to
549 be set on the anterior commissure and the anterior commissure - posterior commissure (AC-PC)
550 line, respectively. Then, cortical surface meshes were generated from the processed T1w images
551 using *recon-all* from Freesurfer (version 7.3.2)[53], in which anatomical surface segmentations
552 were carefully checked and manually corrected. For visualization of analytical results on the
553 whole cortical areas, we visualized results using flattened cortical surfaces reconstructed from
554 the T1w images of individual subjects. Relaxation cuts were made into the surface of each
555 hemisphere to make flattened cortical surfaces for individual subjects. Functional data were
556 aligned and projected onto the surface for visualization using Pycortex[59].

557 For functional data of each run, we performed the MRI data preprocessing through the pipeline
558 provided by fMRIPrep (version 20.2.7)[60]. First, a BOLD reference image was generated using a
559 custom methodology of fMRIPrep. A field map (B0-nonuniformity map) estimated based on a
560 phase-difference map was used to estimate susceptibility distortion and to correct the BOLD
561 reference for a more accurate co-registration with the anatomical reference. The BOLD
562 reference was then co-registered to the T1w reference using bbregister (FreeSurfer; version
563 7.3.2) which implements boundary-based registration[61]. BOLD runs were slice-time corrected
564 using 3dTshift from AFNI 20160207[62], and the BOLD time-series were resampled onto their
565 original, native space (2 × 2 × 2 mm voxels) by applying a single, composite transform to

correct for head-motion and susceptibility distortions using antsApplyTrainsforms from ANTs (version 2.3.3) with Lanczos interpolation.

To create data samples, we first discarded the first 8-s scans of the preprocessed BOLD signals from each run to avoid MRI scanner instability and then regressed out nuisance parameters from each voxel amplitude for each run, including a constant baseline, a linear trend, and temporal components proportional to a total of 24 head-motion parameters (six base motion parameters, consisting of three rotations and three translations; temporal derivatives of the six base motion parameters; quadratic terms of the base motion parameters and their temporal derivatives) and 12 global signals (three types of mean voxel amplitudes within brain tissues around cerebrospinal fluid, white-matter, and gray-matter; temporal derivatives of the three base global signals; quadratic terms of the base global signals and their temporal derivatives). The data samples were temporally shifted by 4 s (4 volumes) to compensate for hemodynamic delays, were despiked to reduce extreme values (beyond ± 3 SD for each run), and were then averaged within each stimulus block (training and test sessions in video presentation experiment) and each imagery block (imagery experiment). The amplitudes of each voxel were then z-scored within each run to eliminate potential non-stationarities and scanner-specific signal biases and to equalize units across voxels.

For data from the training session of the video presentation experiment (training perception data), we discarded samples corresponding to the last two runs, in which videos used in the test session and the imagery experiment were presented, to ensure that our results generalize to new stimuli. For data from the test session of the video presentation experiment (test perception data) and those from the imagery experiment (test imagery data), samples corresponding to the identical video (5 repetitions) were averaged across repetitions to increase the signal-to-noise ratio of the fMRI signals unless otherwise stated.

Caption annotation for visual stimuli

We conducted an online experiment using Amazon Mechanical Turk (AMT) to collect written caption descriptions for the stimulus video clips. The collection procedure was based on the approach used in collecting Microsoft COCO Captions[63]. In the experiment, multiple workers viewed each video clip to provide a single sentence encapsulating all the essential aspects of the visual scenes. The captions were required to be detailed, containing more than eight words to enable accurate scene visualization for readers. To ensure the quality, the collected data underwent manual checks by an experimenter and were further proofread with the assistance of ChatGPT (https://chat.openai.com/) for typo correction and removal of incorrect or unintelligible sentences. We collected twenty non-duplicated sentences for each video,

matching the number of captions per video to a representative video caption dataset (MSR-VTT)[64].

Feature computation by deep neural network models

We used deep neural network models pre-trained for language or vision tasks to compute semantic and visual features. In all the following analyses, to mitigate potential biases arising from baseline differences across model units, feature values of individual units underwent z-score normalization using means and standard deviations of feature values of each unit estimated from respective training data for each analysis.

Semantic features. To extract semantic features from the stimulus video captions, we used a total of 42 pre-trained language models (LMs; available at Hugging Face's Transformers library)[65], which were selected to cover a wide range of model architectures (e.g., BERT and GPT2) and sizes (e.g., base, large, and xlarge). For each sequence of words (or sentence), we first tokenized the input sequence and processed the tokenized sequence using an LM to produce a set of vector embeddings, each corresponding to each token, in multiple hidden layers. Following the procedures in Reimers & Gurevych (2019)[66], we constructed a vector embedding for a sequence by averaging vector embeddings over tokens except for the special tokens (e.g., <cls> token) in each layer of a model. We then used the averaged vector embeddings from multiple layers as a set of semantic features for the input sequence of the model. For each video stimulus, we computed semantic features for a total of 20 captions annotated to the videos and averaged these features over the multiple captions to construct semantic features for the video.

The 42 LMs used in the present study were the following (l and u indicate the numbers of layers and units, respectively): BERT[33] (n = 6; base-uncased [BBU; l = 12, u = 768], base-cased [BBC; l = 12, u = 768], large-uncased [BLU; l = 24, u = 1024], large-cased [BLC; l = 24, u = 1024], large-uncased-whole-word-masking [BLUwwm; l = 24, u = 1024], and large-cased-whole-word-masking [BLCwwm; l = 24, u = 1024]), RoBERTa[36] (n = 2; base [RB; l = 12, u = 768] and large [RL; l = 24, u = 1024]), DeBERTa[35] (n = 2; base [DB; l = 12, u = 768], large [DL; l = 24, u = 1024], and xlarge [DXL; l = 48, u = 1024]), ALBERT[67] (n = 8; base-v1 [ABv1; l = 12, u = 768], large-v1 [ALv2; l = 24, u = 1024], xlarge-v1 [AXLv1; l = 24, u = 2048], xxlarge-v1 [AXXLv1; l = 24, u = 4096], base-v2 [ABv2; l = 12, u = 768], large-v2 [ALv2; l = 24, u = 1024], xlarge-v2 [AXLv2; l = 24, u = 1024], and xxlarge-v2 [AXXLv2; l = 12, u = 4096]), OpenAI-GPT[68] (n = 1; [GPT; l = 12, u = 768]), GPT2[69] (n = 4; base [GPT2; l = 12, u = 768], medium [GPT2M; l = 24, u = 1024], large [GPT2L; l = 36, u = 1280], and xlarge [GPT2XL; l = 48, u = 1600]), Sentence GPT[70] (n = 1; [sGPT; l = 12, u = 768]), XLnet[71] (n = 2; base-cased

[XLB; l = 12, u = 768] and large-cased [XLL; l = 24, u = 1024]), DistilBERT[72] (n = 2; base-uncased [dBBU; l = 6, u = 512] and base-cased [dBBC; l = 6, u = 768]), DistilGPT2[65] (n = 1; [dGPT2; l = 6, u = 768]), T5[73] (n = 3; small [T5S; l = 6, u = 512], base [T5B; l = 12, u = 768], and large [T5L; l = 24, u = 1024]), BART[74] (n = 2; base [BaB; l = 6, u = 768] and large [BaL; l = 12, u = 1024]), CTRL[75] (n = 1; [CTRL; l = 48, u = 1280]), XLM[76] (n = 2; mlm-17-1280 [XLM17; l = 16, u = 1280] and mlm-100-1280 [XLM100; l = 16, u = 1280]), XLM-RoBERTa[77] (n = 2; base [XLMRB; l = 12, u = 768] and large [XLMRL; l = 24, u = 1024]), ELECTRA[78] (n = 1; [ELECTRA; l = 12, u = 256]), CLIP[79] text encoder (n = 1; [CLIP; l = 12, u = 512]).

We hypothesized that an LM whose semantic features exhibited higher prediction performance of brain responses to visual stimuli would provide more effective intermediate representations for translating visual semantic information in the brain into text descriptions. According to this hypothesis, we performed the cross-validation encoding analysis within the training perception data using semantic features from each of the 42 LMs to identify the most suitable model. The results consistently showed that the DeBERTa-large model[35] performed well in the validation analysis (Extended Data Fig. 2a). Therefore, we decided to primarily use this model in the decoding analysis.

To construct semantic features that do not contain structured semantic information (Extended Data Fig. 3c), we used captions with randomly shuffled word orders. We created up to 1,000 shuffled variants for each caption by rearranging the order of the words. From these shuffled variants, we computed semantic features and then calculated the average across all the variants. These features were further averaged across twenty captions for each video to obtain semantic features for each video.

Vision model features. To extract visual features from the presented video stimuli, we used a TimeSformer model[44] (model and code are available at https://github.com/facebookresearch/TimeSformer) that was pre-trained in a sequential manner for both object recognition, using ImageNet[80], and action recognition, using Kinetics-400[81]. This model has shown high performance in predicting video-induced brain activity[82]. For each video, we first resized its spatial size to match the model's input size (224 pixels) while preserving the original aspect ratio. We then selected frames at intervals of 32 within the video. If the video had fewer frames than the model's temporal input size (eight temporal positions), we uniformly selected eight frames to cover the entire video length. We computed feature vectors for each layer from these eight temporal positions, ensuring comprehensive coverage of the entire video without any temporal overlaps. Following the procedure outlined in Nishimoto (2021)[82], we averaged the feature vectors over the spatial encoding dimension within each layer. For each

video, we repeated this procedure for three spatial crops (left-center-right or top-center-bottom). Finally, we further averaged the visual features over the temporal dimension and the three spatial positions. This averaging process resulted in 768 features for each of the 12 layers, approximately matching the number of features with those of LMs. These visual features were z-normalized using features from respective training data for each analysis.

Voxel-wise encoding modeling analysis

We performed voxel-wise encoding modeling analysis by constructing encoding models that predict signal amplitudes of individual voxels from a feature vector in each layer of each model (e.g., 1024 features of the first layer of the DeBERTa-large model) using the L2 regularized linear regression algorithm (i.e., ridge regression). The analysis was performed in both a cross-validation and a generalization manner. In the cross-validation analysis, we used 6-fold cross-validation on the training perception data (58 runs data were divided into five sets of 10 runs and one set of 8 runs data). In each fold of the cross-validation analysis, models for individual voxels were trained with the five sets of data, and the estimated models were used to predict voxel activities for the left-out test set. This process was repeated by rotating training-test assignments six times to produce model predictions for all data samples. In the generalization analysis, models were trained using all training perception data and tested on the test perception data (Extended Data Fig. 6). Model performance for each voxel was evaluated by calculating Pearson correlation coefficients between across-sample profiles of measured and predicted brain activities of individual voxels.

The regularization parameters of the ridge regression models were optimized based on the model performance on the training data. Regularization coefficients were estimated separately for each layer, model, and subject based on the performance of all voxels. In the cross-validation analysis, we used a 5-fold cross-validation procedure (inner loop) nested within the training perception data for the outer loops of the 6-fold cross-validation. In each fold of the nested (or inner) cross-validation loop, models for individual voxels were trained with four sets of the data (within the training perception data) using each of the ten possible regularization coefficients (log spaced between 10 and 10000). The estimated models were then used to predict voxel activities for the left-out test set (one out of five sets). This procedure was repeated by rotating training-test assignments five times to produce model predictions for all data samples within each fold of the outer loops of the cross-validation. These procedures were also repeated for each outer loop of the cross-validation to estimate model performance for each fold of the outer loops. Finally, the regularization parameters that produced the maximal model performance (mean correlation coefficients averaged over all voxels) in the nested (inner) cross-validation were used for predictions in each fold of the outer cross-validation loops. In the

generalization analysis, we conducted the same parameter estimation procedure using a 6-fold cross-validation procedure within the entire training perception data to determine the regularization parameters.

For each model, predictions from multiple layers were integrated to construct final predictions for the model by selecting the best layer for each voxel based on the model performance in the training data. For each fold of the cross-validation analysis, we determined the best layer per voxel based on the model performances within each of the 5-fold nested cross-validation loops (cf., Fig. 3f). We then aggregated predictions from the best layers determined for each left-out set to construct predictions for all data samples. For the generalization analysis, we determined the best layer per voxel based on the model performance estimated through 6-fold cross-validation within the entire training perception data and used predictions from the best layers as predictions for the test perception data.

Comparisons of encoding accuracies from two models (semantic and visual models) were performed based on the slope angles of best linear fit between two sets of model prediction accuracies of individual voxels. The best linear fit was estimated by Deming regression[83] to accounts for observation errors on both x- and y-axis. The slope estimates were converted to angles by first calculating the arctangent of the slopes to obtain angles in radians, and then converting it to degrees. The calculated angles were further subtracted from 45 (degree) to obtain deviations from the parity (Fig. 3e).

Feature decoding analysis

We performed feature decoding analysis by constructing a set of L2 regularized linear regression models (decoders) that predict a pattern of feature values for each layer of a model from a pattern of fMRI voxel activities (one decoder for each feature unit). We used the L2 regularized linear regression algorithm and performed both cross-validation and generalization analyses to produce predictions for all data samples in the training perception data as well as the test perception and imagery data. The decoders were trained using fMRI voxel patterns from the whole brain (WB; unless otherwise stated), and voxels whose signal amplitudes were best predicted by the target feature set were selected based on cross-validation (or nested cross-validation) encoding analysis in the respective training data (with a maximum of 50,000 voxels). The performance of the feature decoding analysis was evaluated by calculating Pearson correlation coefficients between the across-sample profiles of feature values computed from an LM and those predicted from the brain activity for each model unit. In the cross-validation analysis, we used 6-fold cross-validation on the training perception data, and the model training and the regularization parameter estimation were conducted using a 5-fold cross-validation

procedure nested within the training perception data for the outer loops of the 6-fold cross-validation, similar to the encoding analysis (see Methods: "Voxel-wise encoding modeling analysis"). In the generalization analysis, models were trained using all training perception data and tested on the test perception and test imagery data, and the regularization parameters were determined using a 6-fold cross-validation within the entire training perception data. In both cross-validation and generalization analyses, we optimized the regularization parameters of the ridge regression models based on the model performance on respective training data, and the regularization coefficients were estimated separately for each layer, model, and subject based on the performance of all units.

Text generation analysis

To generate descriptive text based on a set of semantic features (target features; e.g., features for multiple layers decoded from the brain), we conducted an iterative optimization of descriptions, in which semantic features of candidate descriptions were progressively aligned with target features through the iterative replacement and interpolation of tokens (referred to as "words" for simplicity) within the candidate descriptions. Each step of the optimization process consisted of three stages: the masking stage, the unmasking stage, and the candidate selection stage.

In the masking stage, for each of the current candidate descriptions (e.g., "metal shapes"), we first generated an exhaustive list of masked candidates by replacing each word or a sequence of words (three words at the maximum) with a mask token (e.g., "<MASK> shapes," "metal <MASK>," and "<MASK>") or interpolating a mask token between words or at the top or bottom of the description (e.g., "metal <MASK> shapes," "<MASK> metal shapes," and "metal shapes <MASK>"). This masking procedure was repeatedly applied on the newly generated masked candidates (two times at the maximum) to generate masked candidates with multiple masks (e.g., "<MASK> <MASK> shapes" and "metal <MASK> <MASK>"). From all the generated masked candidates for each original candidate, we randomly selected five masked candidates to process them in the next unmasking stage.

In the unmasking stage, we used an LM pre-trained for masked language modeling (MLM model; not necessarily the same as the model used for the feature computation) to generate alternative words to fill in the masks in the masked candidates within the context of surrounding words. For each mask within a masked candidate description, we generated five alternative words for the mask to create five new candidates by random sampling from a categorical distribution likelihood estimated by the MLM model. In cases where a masked candidate had multiple masks, we processed the masks sequentially from the top until all masks were updated. These procedures yielded five new candidates from one masked candidate. In the main analysis,

we used the pre-trained RoBERTa-large model (vocabulary size, 50265; subword segmentation based on byte pair encoding) for the guide of the text generation because this model consistently demonstrated stable performance in optimizing descriptions in a validation analysis with the training data (Extended Data Fig. 2b).

In the candidate selection stage, we computed semantic features of all new and original candidate descriptions using an LM from which the target features originated. We then calculated Pearson correlation coefficients between those candidate features and target features for all layers and averaged those correlations over layers to score the candidate descriptions. To enhance the conciseness of generated descriptions, reduce the computational cost of handling long descriptions, and avoid overfitting to noises on brain-decoded features, we added an exponential penalty to the length of candidate descriptions as described by $s = \frac{r}{l^{\alpha}}$, where s is the similarity score used to rank candidates, r is the mean correlation coefficient between candidate features and target brain-decoded features averaged over layers, l is the length (or the number of tokens) of the candidate, and α is the parameter for the length penalty. We chose the parameter α = 0.1 based on the validation analyses using six possible penalty parameters (0, 0.05, 0.1, 0.15, 0.2, and 0.25) with a randomly selected subset of 50 samples from the training perception data. The mean length of generated descriptions with this parameter (mean l = 15.56) was comparable with the mean length of reference captions in the training perception data (mean l = 15.94) under the same analytical setting as the main analysis (Extended Data Fig. 2c−e). After computing similarity scores for all candidates, we ranked them and selected the top five candidates to proceed with further optimization.

During each step of the optimization process, the maximum search width was 130 (five new candidates for each of the five masked candidates derived from five original candidates, in addition to the five original candidates themselves).

We repeated these optimization stages 100 times, and the obtained description was taken to be the text describing the semantic information represented in the target features, or the brain. Since the masking and unmasking stages involve randomness in the optimization process, to avoid the local optima, we repeated the same process five times for each data sample to select the description showing the highest similarity scores with the target features as the final prediction. For all the analysis, we began the optimization process from a non-informative initial state (i.e., unknown token, <unk> for the tokenizer of the RoBERTa-large) to avoid incorporating any prior assumptions for description generation.

Unlike autoregressive LMs with causal attention often used to support linguistic information decoding[3,4], MLM models with bidirectional attention have the significant advantage of

incorporating contextual information from all surrounding words[33]. This characteristic makes our optimization process more suitable for decoding visual information, which lacks specific directionality, in contrast to linguistic information.

Database-search-based description prediction analysis

One of the key strengths of our method is its flexibility to optimize descriptions at the word (or token) level. To assess its effectiveness, we used the database-search (DB-search) method[9] to set the baseline performance for description prediction from the brain. In this analysis, predictions were produced by searching for the best captions whose semantic features exhibited the highest correlations with target features from multiple large databases of image and video captions (MS-COCO[84], Microsoft common objects and context; GCC[85], Google Conceptual Captions; MSR-VTT[86], Microsoft Research Vision to Text). We computed semantic features for all captions in the databases (~4.1M) using an LM used in the feature decoding analysis (i.e., the DeBERTa-large model). For each fMRI data sample, we predicted semantic features at multiple layers using trained decoders and computed correlation coefficients for all layers between the decoded features and the features computed from all database captions. Finally, we selected the caption with the highest mean correlations averaged over all layers as the prediction for the sample.

Evaluation of the similarity between generated descriptions and references

We used multiple metrics to quantitatively evaluate the similarity of generated descriptions to reference captions annotated to the viewed or imagined videos. These metrics include the Pearson correlation coefficient (feature correlation), BLEU, METEOR, ROUGE-L, CIDEr, and three variants of BERTScore. In the evaluation of each description predicted for a viewed or imagined video, we computed scores for each metric between the description and twenty reference captions annotated to the corresponding video. Subsequently, we selected the highest matching score among the twenty references as the final score. For the computation of BLEU and METEOR, we used a Python implementation in the NLTK toolbox[87]. For ROUGE-L and CIDEr, we used the code available at https://github.com/salaniz/pycocoevalcap. For BERTScore, we used the code available at https://github.com/Tiiiger/bert_score.

Feature correlation. We defined the feature correlation as the mean of Pearson correlation coefficients between semantic features of a reference caption or target features and those of a generated description averaged over multiple layers. In this evaluation, we computed semantic features using the same LM as the one used in the decoding analysis, unless otherwise stated (i.e., DeBERTa-large in the main analysis). The feature correlation was used for all analyses where the similarity metric is not explicitly specified.

BLEU. BLEU computes the precision scores by comparing predicted n-grams with reference captions while considering a brevity penalty. We used the 4-gram variant (BLEU-4) with a smoothing method[88].

METEOR. METEOR computes scores based on unigram matching in predicted and reference sequences using precision and recall while considering word variations like stemming and synonymy.

ROUGE-L. ROUGE-L computes scores by emphasizing recall, measuring the overlap in words between the prediction and reference, and primarily focusing on their longest shared sequence.

CIDEr. CIDEr measures the similarity between predicted descriptions and references through n-gram-based comparison while considering consensus across multiple references.

BERTScore. BERTScore computes the similarity based on contextualized embeddings computed from individual tokens by a bidirectional transformer LM. We used the 17th layer of the RoBERTa-large model with the baseline rescaling (default of the official implementation) to compute scores and used three variants of the BERTScore, P (precision), R (recall), and F1 for the evaluation. To evaluate the token-wise precision of generated descriptions, we used the precision measure P weighted by the inverse of document frequency (IDF) estimated from captions in multiple databases (MSCOCO, GCC, and MSR-VTT) and all captions collected in the present study. The IDF-weighted P was only used to highlight tokens with high precision in generated descriptions, and not used in any quantitative evaluations.

Based on these metrics, we further evaluated the discriminability of the generated descriptions. For each description, we calculated similarity scores between the generated description and the reference captions annotated to the corresponding viewed or imagined videos (correct) and those annotated to irrelevant videos (incorrect; n = 2,179). We then used the difference between the score of the correct video and the mean score averaged over all incorrect videos as the discriminability.

Video identification analysis

To evaluate the performance of description predictions, we performed the video identification analysis to identify videos based on the similarity between predicted descriptions and reference captions annotated to videos, using multiple metrics (see Methods "Evaluation of the similarity between generated descriptions and references"). For each data sample, we first computed similarity scores between the predicted description and all the reference captions of all videos using a specific metric. We then compared the similarity scores to the correct reference captions

(twenty captions of the target video) with those to the incorrect reference captions (a total of 43,580 captions of 2,179 irrelevant videos) to see if the similarity scores to the correct references were higher than those to the incorrect references. For the analysis with the feature correlation and BERTScore, we used mean similarity scores averaged within multiple captions for each video. For the analysis with BLEU, METEOR, ROUGE-L, and CIDEr, we used the highest matching scores among multiple captions for each video. We performed the analysis with variable numbers of candidates that ranged from 2 (chance level, 50%) to 100 (chance level, 1%) and selected the video with the highest similarity score among candidates as the prediction for identification. We defined the accuracy by the proportions of correct video identification. When the number of candidates was two (one correct and the other incorrect), we conducted identification for all combinations of the correct candidate and each of the other incorrect candidates. In cases where the number of candidates was more than two (one correct and a specified number of incorrect), we randomly selected a set of incorrect candidates from all available incorrect candidates, repeating the analysis 100 times to estimate the mean accuracy averaged over repetitions.

Evaluation of the fidelity of the relational information in generated descriptions

To investigate whether the generated descriptions faithfully represent the visual relationships among individual components in viewed or imagined videos (e.g., differences between "a bird eats a snake" and "a snake eats a bird")[40], we evaluated the effect of shuffling word orders of generated descriptions on the discriminability and video identification performance. For this evaluation, we randomly shuffled the word order of each generated description to create word-shuffled variants of the original descriptions. The shuffling was performed at the word level rather than the token level to maintain minimal coherence within individual words. For each original description, we created up to 1,000 shuffled variants using all words or only nouns, in which descriptions that contained only one noun were excluded from the noun-shuffling analysis. To identify nouns in each description, we performed the part-of-speech tagging using spaCy[89] (version 2.2.4). We then computed semantic features for these shuffled descriptions.

If generated descriptions accurately captured the visual relations among individual components of the viewed and imagined videos, higher similarities to correct video captions should be observed with the original rather than shuffled descriptions while maintaining differences to irrelevant video captions. We thus computed feature correlation scores between reference captions and the original and shuffled variants of generated descriptions to examine if the original descriptions exhibited higher discriminability than the shuffled ones (e.g., Fig. 2f right). Furthermore, we performed the video identification analysis using either original or shuffled

variants of the generated description to quantify how the shuffling impacts the accuracy of correctly identifying viewed and imagined contents (e.g., Fig. 2f left).

We also examined the possibility that the word orders of the generated descriptions were artificially created by the MLM model used for description generation while allowing for deviations from the information in the brain. If a generated description best matches the brain representation, its semantic features should exhibit the highest similarity to the brain-decoded features used for the description generation than features of the word-order shuffled variants. We then computed feature correlation scores between the target brain-decoded features and features of the original and the shuffled descriptions to examine if the original description showed a greater score than the shuffled descriptions. Additionally, to take into consideration the degree of meaning changes of the generated descriptions caused by shuffling, we computed feature correlation scores between the original and shuffled descriptions, defining the correlation distance as one minus the feature correlation (Fig. 2g and Extended Data Fig. 4f).

Evaluation of the diversity in generated descriptions

For the evaluation of the diversity of generated descriptions, we used Self-BLEU[90] to compute the sentence similarity either across subjects for each video or across videos for each subject. For each video/subject, we can compute a BLEU score by regarding one description from a subject/video as a hypothesis and descriptions from the other subjects/videos as references. To evaluate the diversity across subjects/videos, we computed BLEU scores for every generated description from different subjects/videos and defined the average BLEU scores across subjects/videos as the Self-BLEU of the video/subject, respectively. A higher Self-BLEU score indicates less diversity in the descriptions generated for the video/subject. In this analysis, we used the same analytical settings with the use of the BLEU for the similarity evaluation (BLEU-4 with smoothing; see Methods: "Evaluation of the similarity between generated descriptions and references").

Evaluation of the consistency between generated descriptions and subjective ratings

The reference captions used in this study were collected from subjects in an independent online experiment and not from the subjects of our fMRI experiment. Therefore, it is possible that not all captions precisely matched the subjective perception of our fMRI subjects, even though they were used as "correct" references in the evaluation of the generated descriptions. To examine whether descriptions generated from video-induced brain activity were consistent with the subjective perceptions of individual subjects, we investigated the relations between subjective ratings from the video presentation experiment and the similarity of generated descriptions to evaluated captions (see Methods: "Video presentation experiment"). We focused on 212 videos

evaluated during the training perception data collection. We used the generated descriptions obtained from the cross-validation decoding analysis, employing feature correlations as a metric for the similarity. For each generated description for an evaluated video, we computed feature correlations against five evaluated captions. Aggregating these correlation scores across all 212 evaluated videos yielded 1060 scores for each subject. We classified these scores according to ratings by individual subjects to explore whether higher-rated captions exhibited correspondingly higher scores. Additionally, we calculated Pearson correlation coefficients between the feature correlation scores and subjective ratings to ascertain the presence of positive correlations between them (Extended Data Fig. 5).

Statistics and reproducibility

Statistical analysis was performed on an individual basis unless otherwise stated, and the results from multiple subjects (n = 6) can be considered as replications of the analysis[91]. We thus reported quantitative results for each subject, together with averages across subjects, except for the results of the validation analysis (i.e., Extended Data Fig. 2c, d). To account for multiple comparisons, we used the Benjamini-Hochberg method[92] to control the false discovery rate (FDR) and provided this information when applicable (e.g., Figs. 2d, f, 3f, h, and 4c and Extended Data Figs. 4a, 5b, 6b, and 9a).

To test the statistical significance of the performance difference between the pre-trained and untrained MLM models, we used a one-tailed Wilcoxon signed-rank test on the feature correlations between the generated descriptions (after 100 optimization iterations) and both the target brain-decoded features and the correct reference captions (n = 72; FDR corrected across subjects; Fig. 2c and Extended Data Fig. 4a). To test the statistical significance of the decoding accuracy based on the generated descriptions, we used a one-tailed Wilcoxon signed-rank test on the discriminability (differences in similarity scores between correct and incorrect reference captions) calculated for multiple metrics (n = 72; FDR corrected across subjects and metrics; Fig. 2d, and Extended Data Fig. 9a). All results of the video identification analysis were presented with 95% confidence intervals (C.I.) across samples (n = 72) to determine if the mean identification accuracy across samples exceeded the chance level (Figs. 2e, f, 3g, h, 4b−e, and Extended Data Figs. 4c−e, g 9c−e, g, and 10a, b). To test the effect of word-order shuffling on discriminability, we also used a one-tailed Wilcoxon signed-rank test evaluated based on the feature correlation (n = 72; FDR corrected across subjects; Figs. 2f, 3h, and 4c, and Extended Data Figs. 5, and 9f).

To test the statistical significance of the correlation between layer depth and both discriminability (Extended Data Fig. 7a, b top) and its drop by shuffling (Extended Data Fig. 7a, b bottom), we used a one-tailed *t*-test (FDR corrected across subjects).

To test the statistical significance of the diversity differences in the generated descriptions, we used a one-tailed Wilcoxon signed-rank test on estimated self-BLEU scores (n = 72 for self-BLEU across subjects; n = 6 for self-BLEU across videos; Extended Data Fig. 5c, d).

To test the statistical significance of differences in feature correlations across levels of subjective ratings, we used a one-tailed *t*-test after applying Fisher's z transform (FDR corrected across comparison pairs). To test the statistical significance of the correlation between ratings and feature correlations between generated descriptions and evaluated captions, we used a one-tailed *t*-test after applying Fisher's z transform (n = 6,360). To test the interaction effects between text generation methods and ratings, we used ANOVA.

To test the statistical significance of the correlation between discriminability and encoding accuracy across multiple LMs, we used a one-tailed *t*-test after applying Fisher's z transform (n = 42; FDR corrected across subjects and metrics; Extended Data Fig. 6b). For evaluating the similarity between the generated descriptions and evaluated captions, we pooled results from all six subjects to ensure sufficient data samples to detect differences, while also accounting for variations in the number of samples across rating levels and subjects (Extended Data Fig. 5e).

The performance of the encoding analysis was evaluated by the correlation between measured and predicted fMRI signals for each voxel, and the statistical significance was tested using a one-tailed *t*-test after applying Fisher's z transform (n = 2,108; FDR corrected across voxels; Fig 3a, b). The mean encoding accuracy within each brain area was presented with a 95% C.I. across voxels (Fig. 3d). To compare the encoding accuracies between the semantic and visual encoding models were performed based on the slope angles of best linear fit between the two sets of encoding accuracies for individual voxels. The slope estimates were converted to angles by calculating the arctangent of the slopes to obtain angles in radians, which were converted to degrees. The angles were then subtracted from 45 degrees to obtain deviations from parity (Fig. 3e). To test the statistical significance of the difference in the best layers that showed the highest encoding accuracy among multiple layers of the vision DNN and the LM (Fig. 3f), we used a one-tailed Wilcoxon rank-sum test (FDR corrected across subjects).

Reference

51. Horikawa, T., Cowen, A.S., Keltner, D. & Kamitani, Y. The Neural Representation of Visually Evoked Emotion Is High-Dimensional, Categorical, and Distributed across Transmodal Brain Regions. *iScience*, **23**, (2019).
52. Benson, N.C. et al. The Human Connectome Project 7 Tesla retinotopy dataset: Description and population receptive field analysis. *J. Vis*. **18**, (2018).
53. Fischl, B. FreeSurfer. *Neuroimage* **62**, 774–781 (2012).
54. Kay, K. N., Winawer, J., Mezer, A. & Wandell, B. A. Compressive spatial summation in human visual cortex. *J. neurophysiol.* **110**, 481–494 (2013).
55. Stigliani, A., Weiner, K. S. & Grill-Spector, K. Temporal processing capacity in high-level visual cortex is domain specific. *J. Neurosci.* **35**, 12412–12424 (2015).
56. Tootell, R. B. et al. Functional analysis of human MT and related visual cortical areas using magnetic resonance imaging. *J. Neurosci.* **15**, 3215–3230 (1995).
57. Fedorenko, E., Hsieh, P. J., Nieto-Castañón, A., Whitfield-Gabrieli, S. & Kanwisher, N. New method for fMRI investigations of language: defining ROIs functionally in individual subjects. *J. Neurophysiol*. **104**, 1177–1194 (2010).
58. Scott, T. L., Gallée, J., & Fedorenko, E. A new fun and robust version of an fMRI localizer for the frontotemporal language system. *Cogn. Neurosci.* **8**, 167–176 (2017).
59. Gao, J. S., Huth, A. G., Lescroart, M. D. & Gallant, J. L. Pycortex: an interactive surface visualizer for fMRI. *Front. Neuroinform.* **9**, 23. (2015).
60. Esteban, O. et al. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Methods* **16**, 111–116 (2019).
61. Greve, D. N. & Fischl, B. Accurate and robust brain image alignment using boundary-based registration. *NeuroImage* **48**, 63–72 (2009).
62. Cox, R. & Hyde, J. Software tools for analysis and visualization of fMRI data. *NMR in Biomedicine* **10** (1997).
63. Chen, X. et al. Microsoft COCO Captions: Data Collection and Evaluation Server. Preprint at https://arxiv.org/abs/1504.00325 (2015).
64. Xu, J., Mei, T., Yao, T. & Rui, Y. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. *CVPR*, 5288–5296 (2016).
65. Wolf, T. et al. Transformers: State-of-the-Art Natural Language Processing. Conference on Empirical Methods in Natural Language Processing. Preprint at https://arxiv.org/abs/1910.03771 (2019).
66. Reimers, N. & Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Conference on Empirical Methods in Natural Language Processing.* (2019).

67. Lan, Z. et al. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. Preprint at https://doi.org/10.48550/arXiv.1909.11942 (2019).
68. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving language understanding with unsupervised learning. *Technical report, OpenAI*. (2018).
69. Radford, A. et al. Language Models are Unsupervised Multitask Learners. (2019).
70. Muennighoff, N. SGPT: GPT Sentence Embeddings for Semantic Search. Preprint at https://doi.org/10.48550/arXiv.2202.08904 (2022).
71. Yang, Z. et al. XLNet: Generalized Autoregressive Pretraining for Language Understanding. Preprint at https://doi.org/10.48550/arXiv.1906.08237 (2019).
72. Sanh, V., Debut, L., Chaumond, J. & Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. Preprint at https://doi.org/10.48550/arXiv.1910.01108 (2019).
73. Raffel, C. et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.,* **21**, 1–140 (2019).
74. Lewis, M. et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *Annual Meeting of the Association for Computational Linguistics*. (2019).
75. Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C. & Socher, R. CTRL: A Conditional Transformer Language Model for Controllable Generation. Preprint at https://doi.org/10.48550/arXiv.1909.05858 (2019).
76. Lample, G. & Conneau, A. Cross-lingual Language Model Pretraining. Preprint at https://doi.org/10.48550/arXiv.1901.07291 (2019).
77. Conneau, A. et al. Unsupervised Cross-lingual Representation Learning at Scale. Annual Meeting of the Association for Computational Linguistics. (2019).
78. Clark, K., Luong, M., Le, Q. V. & Manning, C. D. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *ICLR* (2020).
79. Radford, A. et al. Learning Transferable Visual Models From Natural Language Supervision. International Conference on Machine Learning. (2021).
80. Deng, J. et al. ImageNet: A large-scale hierarchical image database. *Comput. Vis. Pattern. Recognit.* 248–255 (2009).
81. Carreira, J. & Zisserman, A. Quo Vadis, action recognition? A new model and the kinetics dataset. *CVPR*, 4724–4733 (2017).
82. Nishimoto, S. Modeling movie-evoked human brain activity using motion-energy and space-time vision transformer features. Preprint at https://www.biorxiv.org/content/10.1101/2021.08.22.457251v1 (2021).
83. Cornbleet, P. J., and Gochman, N. Incorrect Least-Squares Regression Coefficients in Method-

Comparison Analysis. *Clin. Chem.* **25**, 432–438 (1979).

84. Lin, T., et al. Microsoft COCO: Common Objects in Context. *European Conference on Computer Vision* (2014).

85. Sharma, P., Ding, N., Goodman, S. & Soricut, R. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. *Annual Meeting of the Association for Computational Linguistics.* (2018).

86. Xu, J., Mei, T., Yao, T. & Rui, Y. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. *CVPR*, 5288–5296 (2016).

87. Bird, S. & Loper, E. NLTK: The Natural Language Toolkit. *In Proc. ACL Interactive Poster and Demonstration Sessions (ed. Scott, D.)* 214–217 (2004).

88. Lin, C. & Och, F.J. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. *Annual Meeting of the Association for Computational Linguistics* (2004).

89. Honnibal, M. & Montani, I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. (2017).

90. Zhu, Y. et al. Texygen: A Benchmarking Platform for Text Generation Models. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval.* (2018).

91. Ince, R. A. A., Kay, J. W. & Schyns, P. G. Within-participant statistics for cognitive science. *Trends Cogn. Sci.* **26**, 626–630 (2022).

92. Benjamini Y. & Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. ser. B,* **57**, 298–300 (1995).

Extended Data

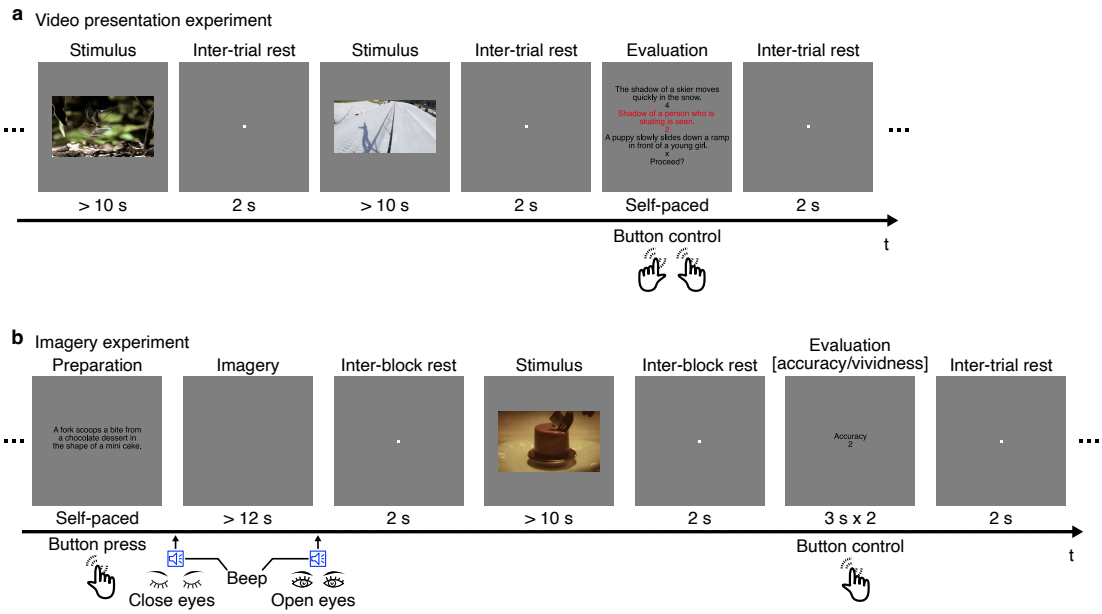
Mind captioning: Evolving descriptive text of mental content from human brain activity

Tomoyasu Horikawa[1*]

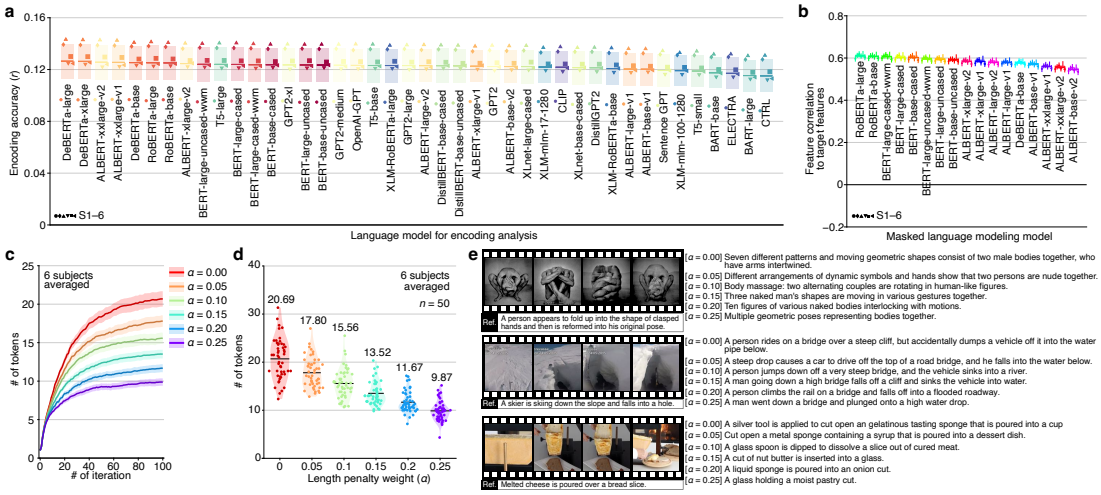
1. NTT Communication Science Laboratories, Kanagawa, Japan.
* Correspondence: horikawa.t@gmail.com

Extended Data

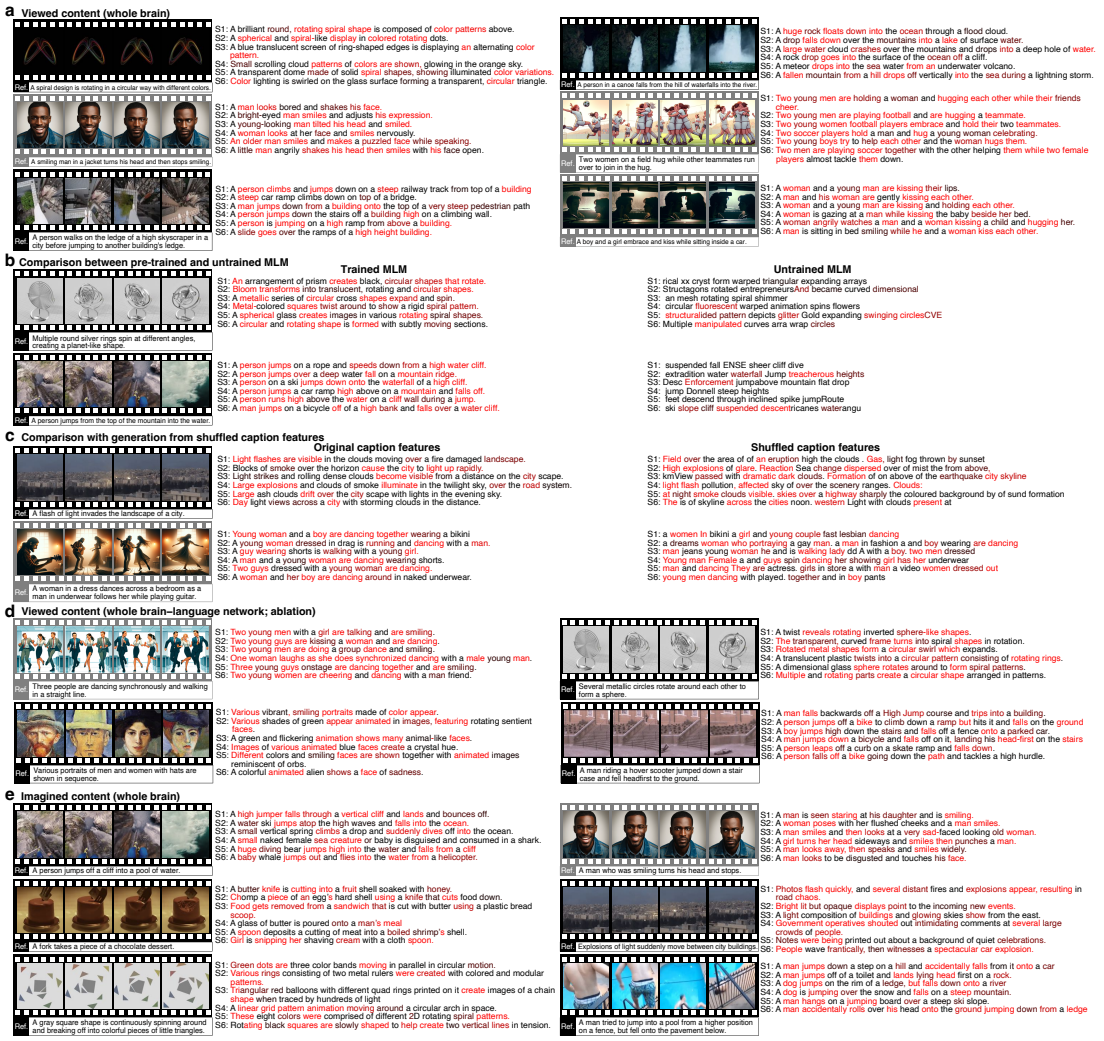
Extended Data Fig. 1 | Overview of experiments
Extended Data Fig. 2 | Validation of model and parameter determinations
Extended Data Fig. 3 | Examples of generated descriptions
Extended Data Fig. 4 | Text generation performance of viewed content for individual subjects
Extended Data Fig. 5 | Comparison of text generation performance with database-search method
Extended Data Fig. 6 | Relations between encoding accuracy and text generation performance of multiple language models
Extended Data Fig. 7 | Effects of shuffling on discriminability with a varying number of language model layers
Extended Data Fig. 8 | Voxel-wise encoding model performance
Extended Data Fig. 9 | Text generation performance of imagined content for individual subjects
Extended Data Fig. 10 | Performance of text generation with different numbers of averaged samples



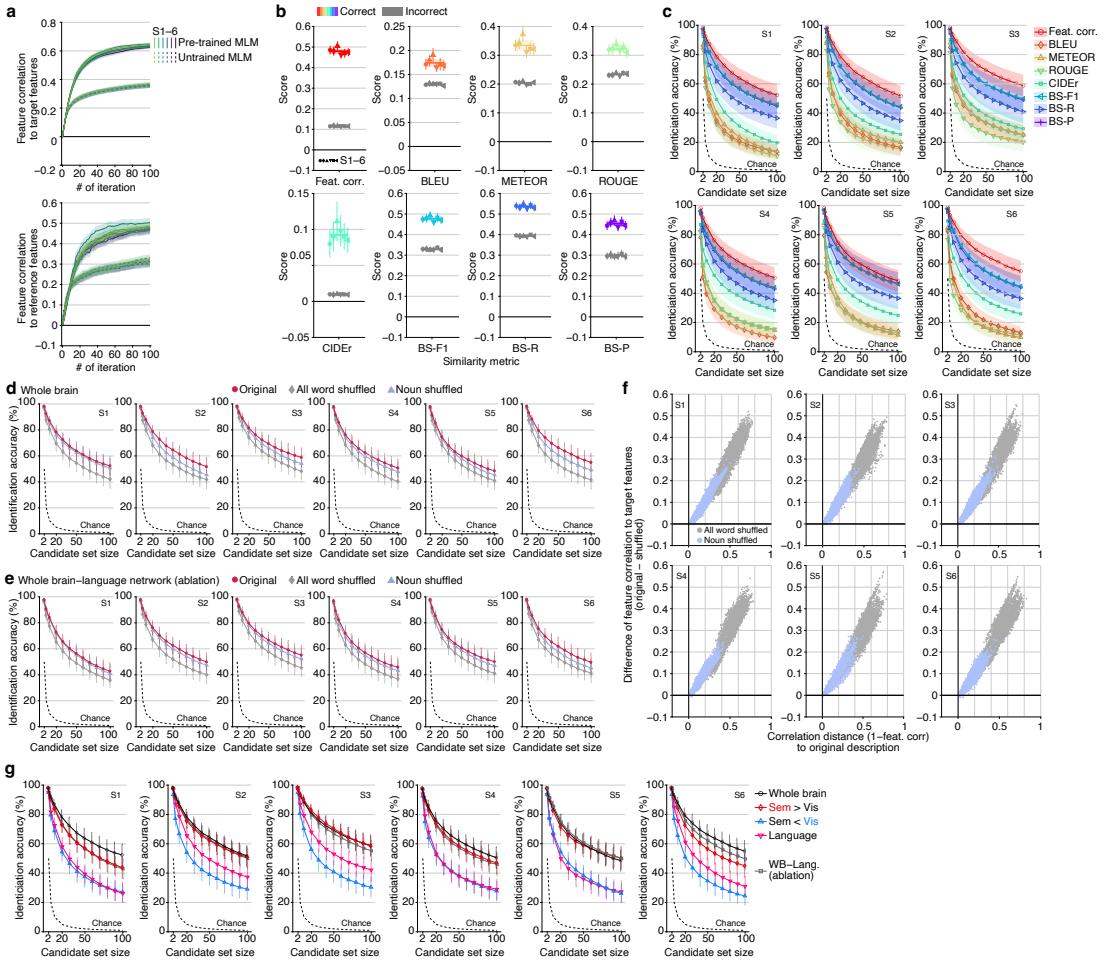
Extended Data Fig. 1 | Overview of experiments. We conducted two types of experiments: a video presentation experiment (including training and test sessions) and an imagery experiment. **a**, Video presentation experiment. Video stimuli were presented sequentially at the center of the display. Subjects were allowed to view the videos without fixating on a specific point. Occasionally, we presented five descriptions depicting the visual content of the preceding video to the subjects, asking them to rate the consistency between each description and their subjective perception. We used these ratings to evaluate the consistency between descriptions generated from the brain and the subjective perception of each individual (Extended Data Fig. 5e). **b**, Imagery experiment. Subjects were well-trained to associate verbal descriptions and target videos before the experiment. They were then required to imagine (recall) a video based on a description of the target video with their eyes closed according to auditory beeps. Each imagery block was followed by a stimulus presentation block, during which the target video was presented to allow the subjects to confirm the validity of their imagery. An evaluation block followed, asking the subjects to evaluate the accuracy and vividness of their imagery. Data samples were constructed by averaging fMRI volumes during each stimulus and imagery block in the video presentation and imagery experiments, respectively. The training data contained samples for 2,108 videos, each presented once. The test and imagery data contained samples for the same 72 videos (not used during training), averaged across five repetitions (unless otherwise stated).



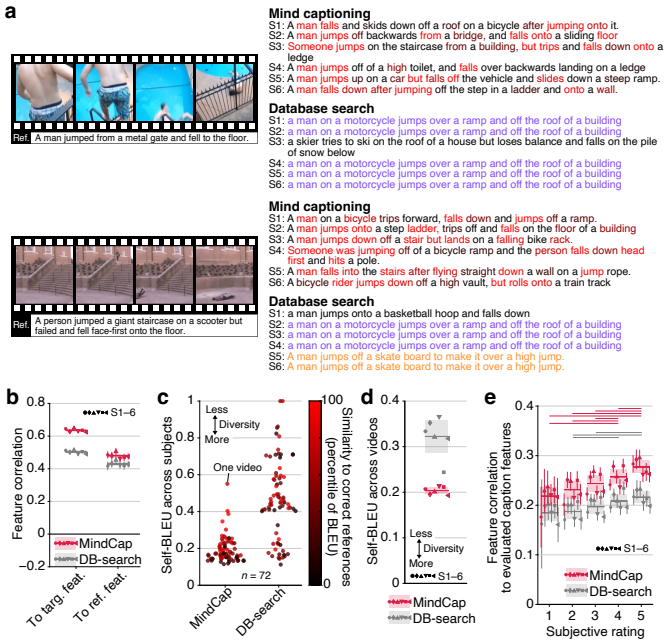
Extended Data Fig. 2 | Validation of model and parameter determinations. All results of this validation were obtained from cross-validation analysis within the training perception data. **a**, Encoding accuracy based on features of different language models (LMs; averaged over voxels in the whole brain). We used the DeBERTa-large model in the main analysis, as it consistently showed the highest performance across all subjects. **b**, Feature correlations between the target brain-decoded features and features of descriptions generated with the support of different MLM models. Different MLM models were used in the text generation analysis (DeBERTa-large was used for the decoding and performance evaluation). All models showed comparable performance in generating descriptions aligned with target decoded features, demonstrating the robustness of our method to the choice of MLM models. We used the RoBERTa-large model in the main analysis, as it showed the best performance. **c**, The number of tokens in generated descriptions through optimization. **d**, **e**, The number of tokens in generated descriptions (**d**) and the example descriptions (**e**) obtained with varying strength of length penalty. While the length of generated descriptions varied depending on the length penalty, the analysis consistently yielded accurate descriptions of the viewed video content. In **b**−**e**, the text generation analysis was performed for 50 videos randomly selected from the training perception data (DeBERTa-large; decoded from whole-brain activity). Shades in **a**, **b** indicate 95% confidence intervals (C.I.) across subjects (n = 6). Shades in **c** and error bars in **a**, **b** indicate 95% C.I. across samples (n = 50).



Extended Data Fig. 3 | Examples of generated descriptions.
[Note: To comply with bioRxiv's policy on displaying human faces, frames containing real human faces in this figure (framed by gray) have been replaced with images synthesized by DALL·E3 (https://openai.com/dall-e-3), based on captions annotated by humans for the video.]
a, Descriptions of viewed content generated from whole-brain activity. **b**, Descriptions of viewed content generated with the support of either the pre-trained or untrained MLM model. **c**, Descriptions of viewed content generated with features predicted by decoders trained based on original captions or those based on captions whose word order was randomly shuffled within each caption. **d**, Descriptions of viewed content generated from the activity of the whole brain except for the language network. **e**, Descriptions of imagined content generated from whole-brain activity. Conventions are the same in Fig. 2a.

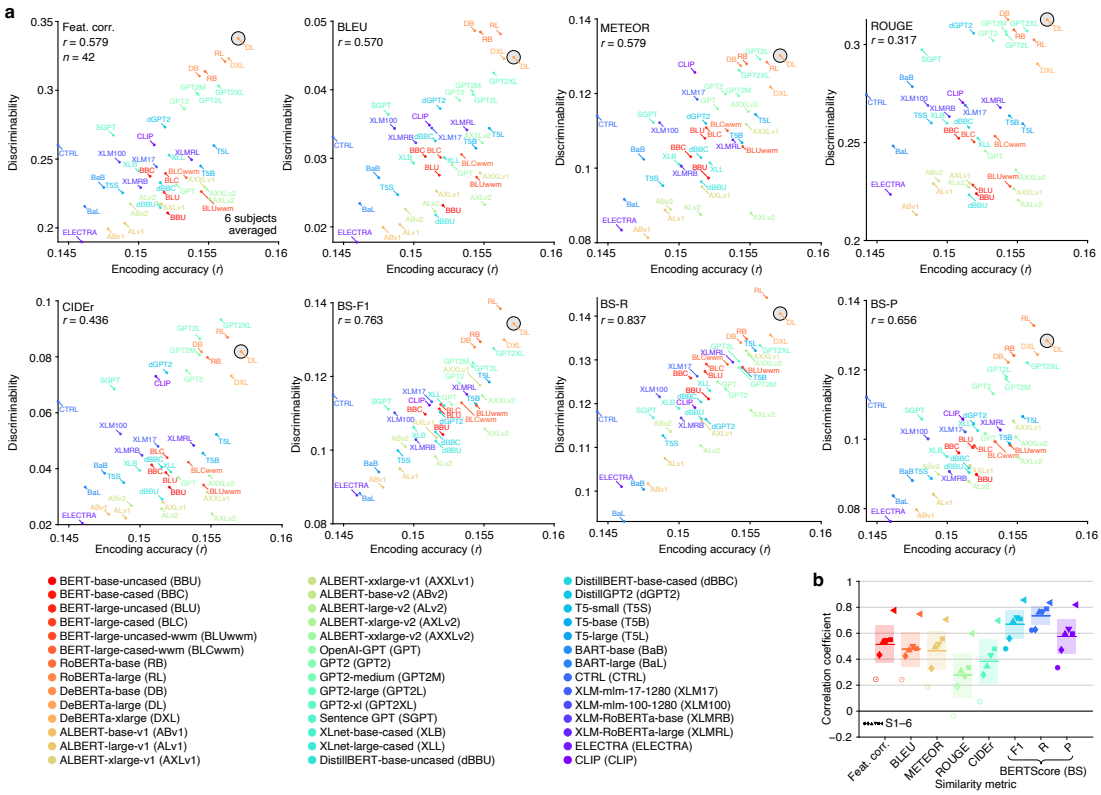


Extended Data Fig. 4 | Text generation performance of viewed content for individual subjects. a, Feature correlations between semantic features of generated descriptions and those decoded from the brain (top), as well as those computed from correct references (bottom) through optimization. The generated descriptions with the support of the pre-trained MLM models outperformed the results of the untrained MLM models for all subjects in both similarity measures (Wilcoxon signed-rank test, one-tailed, $p < 0.01$, FDR corrected across subjects; 100 iterations). **b**, Raw scores of the similarity between generated descriptions and captions for correct and incorrect references based on multiple similarity metrics. **c**, Video identification accuracy based on multiple metrics. **d**, **e**, Effects of word-order shuffling on video identification accuracy applied to descriptions generated from the whole-brain activity (**d**) and the activity of the whole brain except the language network (**e**). **f**, Scatterplot of the correlation distances (one minus feature correlation) between the original and shuffled descriptions against the difference in feature correlations to target features between the original and shuffled descriptions. **g**, Video identification accuracy obtained from different brain areas. Shades in **a**, **c**, and error bars in **b**, **d**, **g** indicate 95% C.I. across samples ($n = 72$). Shades in **b** indicate 95% C.I. across subjects ($n = 6$).

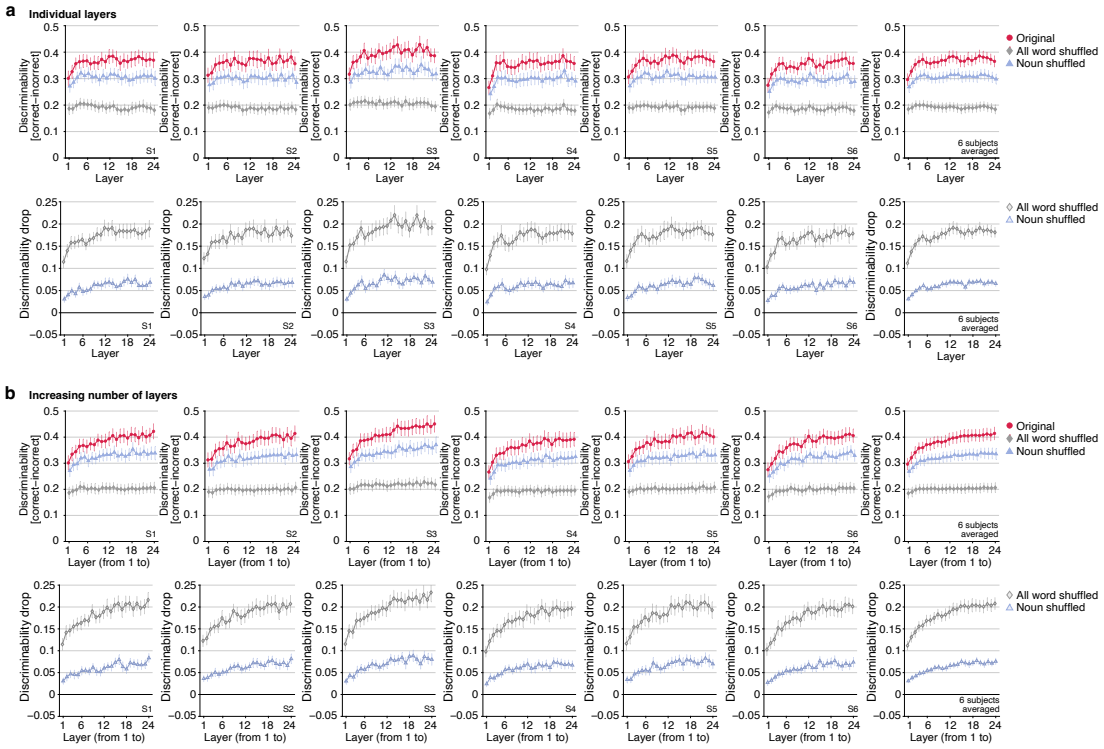


Extended Data Fig. 5 | Comparison of text generation performance with database-search method. The text generation performance for viewed content was compared between our method (mind captioning) and the database-search method (DB-search)[9]. **a**, Examples of generated descriptions. The DB-search produced moderately accurate descriptions, but it often produced identical ones for different individuals and videos. **b**, Feature correlation-based performance. Our method outperformed the DB-search in generating descriptions showing higher feature correlations with both target and reference features (Wilcoxon signed-rank test, one-tailed, $p < 0.01$, FDR corrected across subjects). **c**, **d**, Diversity across subjects (**c**) and videos (**d**). The diversity of generated descriptions was evaluated by self-BLEU[90] across subjects and videos. Our method exhibited greater diversity than the DB-search both across subjects and videos (Wilcoxon signed-rank test, one-tailed, $p < 0.05$). **e**, Feature correlations at each subjective rating. Both methods showed stronger feature correlations between the generated descriptions and the captions with higher ratings (lines above, $p < 0.01$, FDR corrected across pairs, six subjects pooled). Our method showed a significantly higher correlation between ratings and feature correlations than the DB-search (mean $r = 0.184$ and 0.134 for mind captioning and DB-search, respectively; six subjects averaged; $p < 0.01$, six subjects pooled), demonstrating higher flexibility in generating descriptions better aligned with subjective perceptions of individuals (ANOVA, interaction between methods and ratings, $p < 0.01$; six subjects pooled). Shades and error bars indicate 95% C.I. across subjects ($n = 6$) and samples ($n = 72$).

1121

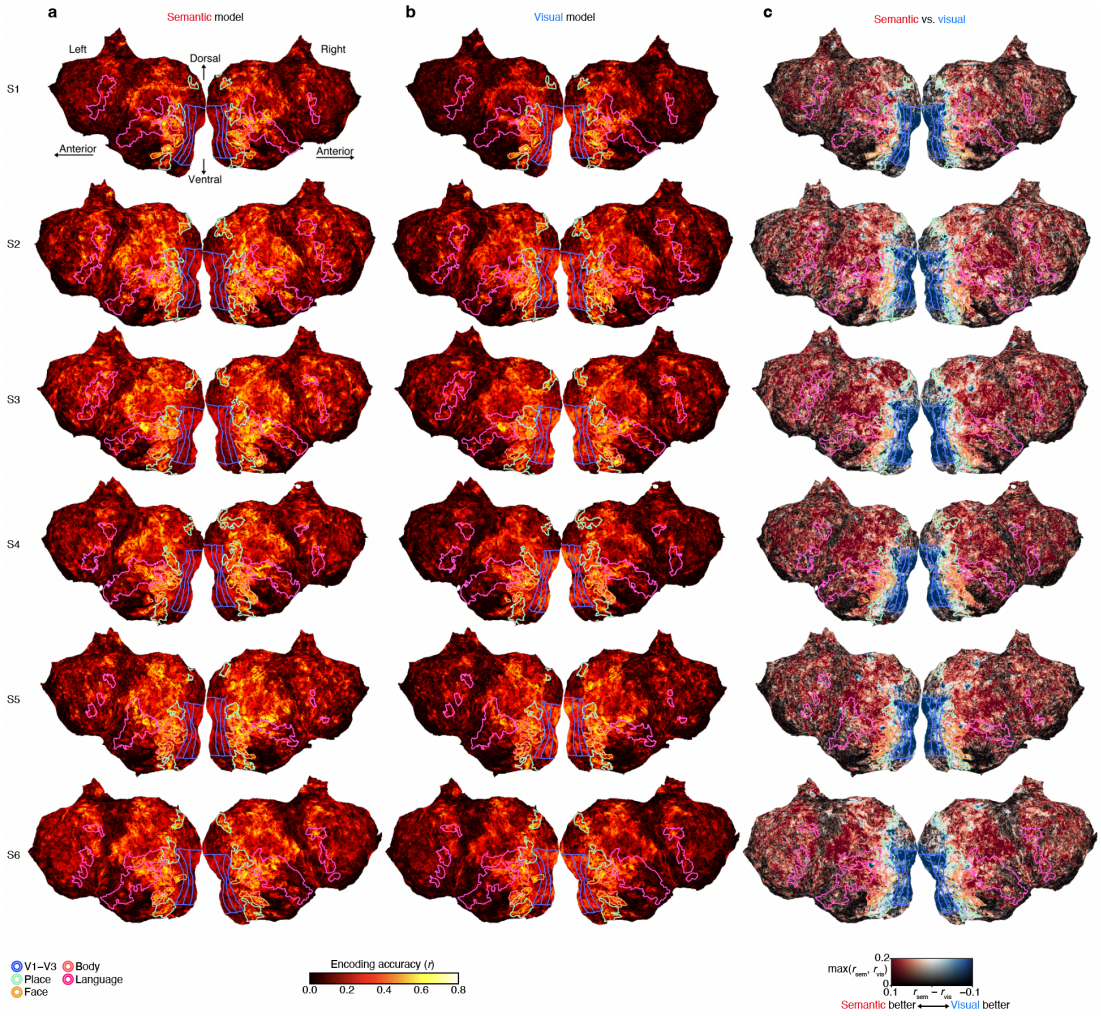


Extended Data Fig. 6 | Relations between encoding accuracy and text generation performance of multiple language models. We performed feature decoding and text generation analyses using features from 42 LMs. The generated description quality was evaluated based on discriminability using multiple metrics. In this analysis, we consistently used the DeBERTa-large model for feature correlation evaluation and the RoBERTa-large model for the text generation analysis to support text generation, respectively. The encoding analysis was performed by training models on the training perception data and testing them on the test perception data. **a**, Scatterplot of encoding accuracy against discriminability for multiple LMs ($n = 42$; six subjects averaged; dark circle, the model used in the main analysis, DeBERTa-large, DL). For each model, the discriminability scores were averaged across all samples, and encoding accuracy was averaged across all voxels in the whole brain. The discriminability scores were positive for all models, suggesting that our method can generate descriptions reflecting the viewed content regardless of the LMs used. **b**, Correlation coefficients between discriminability and encoding accuracy for all models. We found significantly positive correlations between encoding accuracy and discriminability of the generated descriptions for most metrics and subjects, except S1 (filled markers, *t*-test, one-tailed, $p < 0.01$, FDR corrected across metrics and subjects), suggesting that LMs with features more aligned with brain representations could be used to generate descriptions that accurately capture the viewed video content. Shades and error bars indicate 95% C.I. across samples ($n = 72$) and subjects ($n = 6$), respectively.



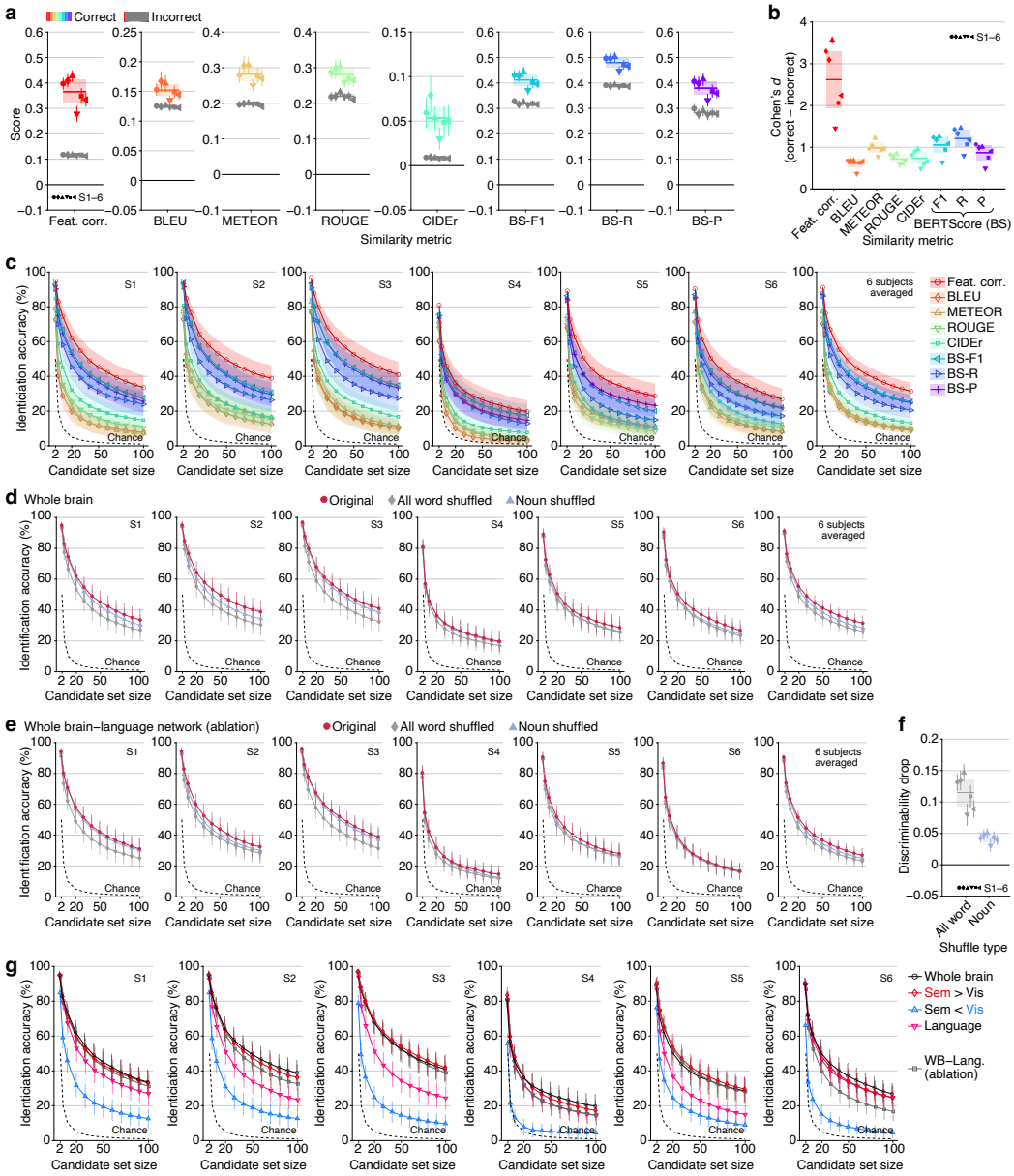
Extended Data Fig. 7 | Effects of shuffling on discriminability with a varying number of language model layers. The text generation analysis was conducted using decoded features of the DeBERTa-large model (24 layers) with a varying number of layers (individual layers or increasing number of layers). We evaluated the impact of shuffling on the discriminability of the descriptions generated with the varying number of layers. To compare the discriminability of the descriptions generated from different layer conditions, we consistently used the feature correlations computed using all layers of the DeBERTa-large model as a fixed metric. **a**, **b**, Discriminability and the drop in discriminability caused by shuffling for descriptions generated with individual layers (**a**) and with increasing number of layers (**b**). Overall, text generation with deeper layers tended to result in higher discriminability (mean r = 0.087 for the individual layers; mean r = 0.189 for the increasing number of layers; t-test, one-tailed, $p < 0.05$, FDR corrected across subjects). Furthermore, the drop in discriminability caused by shuffling became significantly larger when deeper layers were included in the analysis. Error bars indicate 95% C.I. across samples (n = 72; mean r = 0.180 [all] and 0.187 [noun] for the individual layers; mean r = 0.259 [all] and 0.214 for the increasing number of layers; t-test, one-tailed, $p < 0.05$, FDR corrected across subjects).

1123



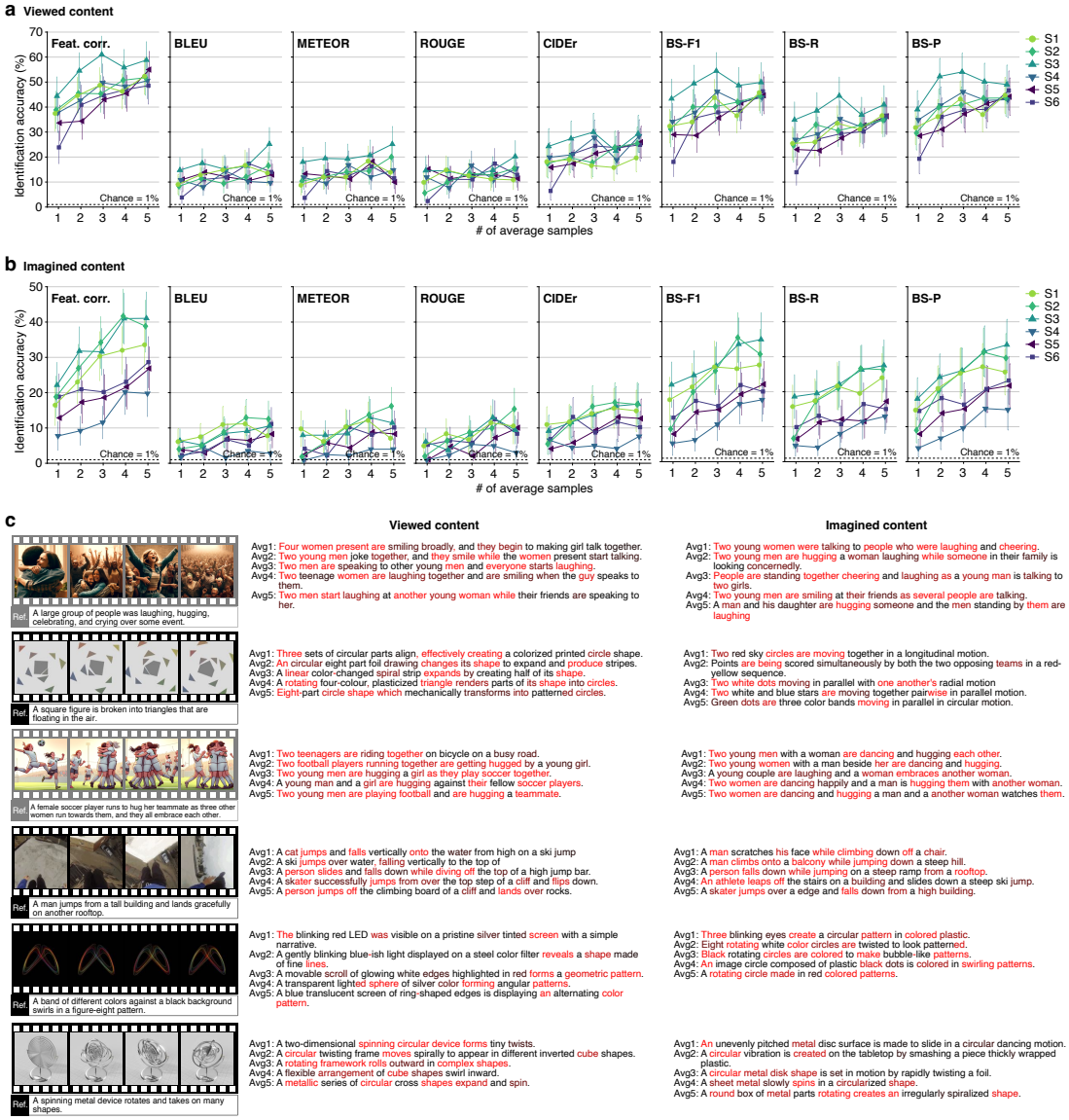
Extended Data Fig. 8 | Voxel-wise encoding model performance. a, Encoding accuracy by the semantic encoding model. **b**, Encoding accuracy by the visual encoding model. **c**, Comparison of encoding accuracy between the semantic encoding model and visual encoding model.

1124



Extended Data Fig. 9 | Text generation performance of imagined content for individual subjects. a, Raw scores of the similarity between generated descriptions and captions for correct and incorrect references based on multiple similarity metrics. **b**, Cohen's *d* of the discriminability. The generated descriptions showed significantly high discriminability across all metrics and subjects (Wilcoxon signed-rank test, one-tailed, $p < 0.01$, FDR corrected across metrics and subjects). **c**, Identification accuracy of imagined videos based on multiple metrics. **d**, **e**, Effects of word-order shuffling on identification accuracy of imagined videos applied to descriptions generated from the whole-brain activity (**d**) and the activity of the whole brain except the language network (**e**). **f**, Effects of word-order shuffling on the discriminability. The word-order shuffling resulted in significant drops in the discriminability even without using the language network (Wilcoxon signed-rank test, one-tailed, $p < 0.01$, FDR corrected across subjects). **g**, Identification accuracy of imagined videos from different brain areas. Shades in **c** and error bars in **a**, **d**–**g** indicate 95% C.I. across samples ($n = 72$). Shades in **a**, **b**, and **f** indicate 95% C.I. across subjects ($n = 6$).

1125



Extended Data Fig. 10 | Performance of text generation with different numbers of averaged samples.
[Note: To comply with bioRxiv's policy on displaying human faces, frames containing real human faces in this figure (framed by gray) have been replaced with images synthesized by DALL·E3 (https://openai.com/dall-e-3), based on captions annotated by humans for the video.]
The feature decoding and text generation analyses were performed with brain activity averaged across different numbers of samples. **a**, **b**, Video identification accuracy of viewed (**a**) and imagined (**b**) content with varying numbers of averaged samples. The video identification analysis was performed with 100 candidates (chance = 1%; error bars, 95% CI across samples). **c**, Descriptions generated using different numbers of averaged samples. For each video, generated descriptions from the same subjects are shown for both viewed and imagined content. The analysis revealed that the quality of the generated descriptions improved as the number of averaged samples increased, indicating that the performance was limited by noise in the fMRI signal of the test data. Meanwhile, the descriptions generated from fMRI activity in single trials reasonably captured the content of both viewed and imagined videos. These results suggest that our method is capable of generating moderately accurate descriptions of both viewed and imagined content, even when using single-trial fMRI activity.