UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Improving Brain-to-Image Reconstruction via Fine-Grained Text Bridging

Permalink

https://escholarship.org/uc/item/9hd6k1gb

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 47(0)

Authors

Xia, Runze
Feng, Shuo
wang, renzhi
et al.

Publication Date

2025

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Improving Brain-to-Image Reconstruction via Fine-Grained Text Bridging

Runze Xia, Shuo Feng, Renzhi Wang, Congchi Yin, Xuyun Wen, Piji Li*
{xiarunze, fengshuo, rzhwang, congchiyin, wenxuyun, pjli}@nuaa.edu.cn
College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing, China
MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing, China
The Key Laboratory of Brain-Machine Intelligence Technology, Ministry of Education, Nanjing, China

Abstract

Brain-to-Image reconstruction aims to recover visual stimuli perceived by humans from brain activity. However, the reconstructed visual stimuli often missing details and semantic inconsistencies, which may be attributed to insufficient semantic information. To address this issue, we propose an approach named Fine-grained Brain-to-Image reconstruction (**FgB2I**), which employs fine-grained text as bridge to improve image reconstruction. FgB2I comprises three key stages: detail enhancement, decoding fine-grained text descriptions, and text-bridged brain-to-image reconstruction. In the detail-enhancement stage, we leverage large vision–language models to generate fine-grained captions for visual stimuli and experimentally validate its importance. We propose three reward metrics (object accuracy, text-image semantic similarity, and image-image semantic similarity) to guide the language model in decoding fine-grained text descriptions from fMRI signals. The fine-grained text descriptions can be integrated into existing reconstruction methods to achieve fine-grained Brain-to-Image reconstruction.

Keywords: Brain Decoding; fMRI; Brain-to-Image Reconstruction

Introduction

Human possesses sophisticated visual and cognitive systems that allow us to easily comprehend visual scenes and guide actions (Kandel et al., 2000; Goodale & Milner, 1992). However, the specific mechanisms underlying this remarkable ability still remain unknown, posing a significant challenge in cognitive science, while also captivating the neuroscience community's interest in exploring these mysteries further. The goal of Brain-to-Image reconstruction is to recover visual stimuli perceived by humans from brain activity signals. Recent advances in functional Magnetic Resonance Imaging (fMRI), particularly its high spatial resolution, have driven progress in decoding and reconstructing visual information from brain activity patterns (Glover, 2011).

Unlike cameras that record every pixel, the human brain processes visual cognition differently (Chen, Qi, & Pan, 2023). It concurrently handles linguistic concepts and semantic information associated with visual scenes (Popham et al., 2021; Zhang, Han, Worth, & Liu, 2020), as depicted in the top section of Figure 1. For example, as shown in the image, when we observe a bathroom scene, our focus may be on the main objects, forming semantic concepts in the brain's representation space such as *mirror*, *sink*, *towels*, and *toilet*.

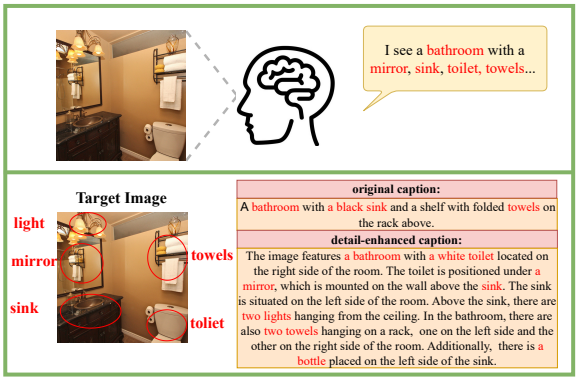


Figure 1: The top section is the illustration of cognitive assumptions during scene observation. The bottom section demonstrates a comparison of the granularity between the detail-enhanced and the original caption.

Therefore, semantics plays a vital role in Brain-to-Image reconstruction. Some studies (Lu, Du, Zhou, Wang, & He, 2023; Takagi & Nishimoto, 2023), have utilized the original image captions from the Natural Scenes Dataset (NSD) (Allen et al., 2022) as semantic targets. These studies use the CLIP (Radford et al., 2021) text encoder to obtain the semantic representation of image captions and train regression models to map fMRI signals to the semantic representation space, thereby achieving semantic reconstruction of images.

However, in more complex visual stimuli, the simple captions often fail to adequately describe the content of the image. As a result, objects observed in the image may not appear in the semantic target, limiting the potential of semantic reconstruction. For instance, the original caption corresponding to the scene in the Figure 1 is *A bathroom with a black sink and a shelf with folded towels on the rack above*. This caption omits obvious objects like "toilet" and "mirror", though information about these prominent objects is likely present in the brain activity, leading to a lack of decoding details.

To address this issue, we propose detail-enhancement for image captions using large vision-language models with strong text-image comprehension capabilities to generate fine-grained captions that comprehensively capture all the main information. In the bottom section of Figure 1, we illustrate the difference between the two types of captions, showing how the detail-enhanced caption better describes the main

*Corresponding author.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

content of the image.

Based on this, we propose a novel Brain-to-Image reconstruction method called Fine-grained Brain-to-Image (**FgB2I**), which aims to improve Brain-to-Image reconstruction by decoding fine-grained semantic descriptions from brain signals as a bridge for image reconstruction. We first train an unified brain-to-text model for all subjects. To decode consistent and fine-grained text descriptions for visual stimuli, we design three reward metrics to guide the model: object accuracy, text-image semantic similarity, and image-image semantic similarity. However, these metrics are non-differentiable and cannot be directly used for model training. Inspired by (Mnih et al., 2015), we employ the reinforce algorithm (Sutton, McAllester, Singh, & Mansour, 1999) to enable the model to be normally trained using these reward metrics. Through these steps, we are able to decode fine-grained textual descriptions from brain signals. We can easily integrate the decoded fine-grained textual descriptions into the high-level reconstruction of existing methods to improve Brain-to-Image reconstruction.

In summary, our study makes the following contributions:

- We identify the issue of missing target details in textual descriptions for Brain-to-Image reconstruction and propose the detail-enhancement stage to remedy this issue.
- We introduce a novel approach named Fine-grained Brain-to-Image reconstruction (**FgB2I**), which employs detail-enhanced text as supplements for Brain-to-Image reconstruction. We use three rewards to guide the decoding of text descriptions and employ Reinforce Algorithm to address indifferentiable problems in model co-training stage.
- We conduct extensive experiments on multiple existing Brain-to-Image reconstruction methods, demonstrating the necessity of detail enhancement and the effectiveness of different reward metrics in our proposed framework.

Related Work

Brain-to-Text Decoding

Recent advances in deep learning have propelled brain decoding research, particularly in reconstructing language and visual stimuli from cognitive signals. Notable progress includes EEG2Text (Wang & Ji, 2022) achieving open-vocabulary sentence-level decoding from EEG, while UniCoRN (Xi et al., 2023) and PREDFT (Yin, Ye, & Li, 2024) successively advanced fMRI-to-text translation through NLP-inspired architectures and predictive coding mechanisms, highlighting the growing sophistication of Brain-Computer Interfaces.

Brain-to-Image Reconstruction

Visual reconstruction research increasingly adopts diffusion models for their generation capabilities. Key approaches include mapping fMRI data to CLIP text features and Stable Diffusion's VAE latent space (Takagi & Nishimoto, 2023), and Ozcelik et al.'s two-stage framework (Ozcelik & VanRullen, 2023) combining VDVAE (Child, 2021) for low-level attributes with Versatile Diffusion (Xu, Wang, Zhang, Wang, & Shi, 2023) guided by CLIP semantics. MindEye (Scotti et al., 2023) further demonstrates how diffusion priors optimize reconstruction accuracy. Selective attention modulates the precision of neural representations during visual encoding by suppressing unattended features' fidelity, while working memory constraints further shape object-level reconstruction through their limited capacity to maintain task-relevant signals during post-perceptual processing stages (Chun & Johnson, 2011; Luck & Ford, 1998; Xia, Yin, & Li, 2024).

Method

Detail Enhancement for image captions via LVLMs

As shown in Figure 2(a), we address the limited descriptive capacity of original image captions through visual-language augmentation. Standard captions often omit critical details (e.g., wall paintings and metallic objects in Figure 2(a)) that humans naturally perceive. This leads to the loss of information, which the details enhancement process aims to address.

We employ the large visual-language model (LVLM) LLaVA (Liu, Li, Wu, & Lee, 2023) to generate fine-grained captions for this stage. LLaVA demonstrates powerful image understanding capabilities, allowing us to extract granular information from images. We employ it with prompt "*Outline the main content of the image. Describe the color, shape, and size of an object. Use plain language to accurately describe key visual information.*" to generate text descriptions, effectively capturing object attributes and spatial relationships.

Brain-to-Text Decoding Model

We develop an unified model for all participants to decode text descriptions from their brain signals. Inspired by the concept of prefix tuning (Li & Liang, 2021), we employ a transformer-based network to align fMRI signals with the textual space, obtaining a prefix embedding that guides the GPT2 (Radford et al., 2019) model to generate text. Similar to the approach in (Scotti et al., 2023), we use the predefined template `nsdgeneral` in the dataset to obtain fMRI signals $F_i \in \mathbb{R}^{N_s}$ for each participant s.

The specific process is illustrated in Figure 3. It is important to note that the number of fMRI voxels varies across participants, so we construct a linear layer for each participant to map the fMRI signals to a unified dimensionality. For the fMRI signals of participant s, they are first passed through a linear layer to obtain the fMRI embedding $Z \in \mathbb{R}^{l \times d}$, where l denotes the length of the fMRI prefix and d is the dimension of the word embedding in the GPT2 model. Additionally, a learnable sequence constant embedding $Z' \in \mathbb{R}^{l \times d}$ is introduced to capture semantic information carried in Z fully. The Z and Z' embedding are concatenated and inputted into a multi-layer Transformer for comprehensive interaction, generating the final fMRI prefix embedding $p \in \mathbb{R}^{2l \times d}$ as input to the GPT2 model.

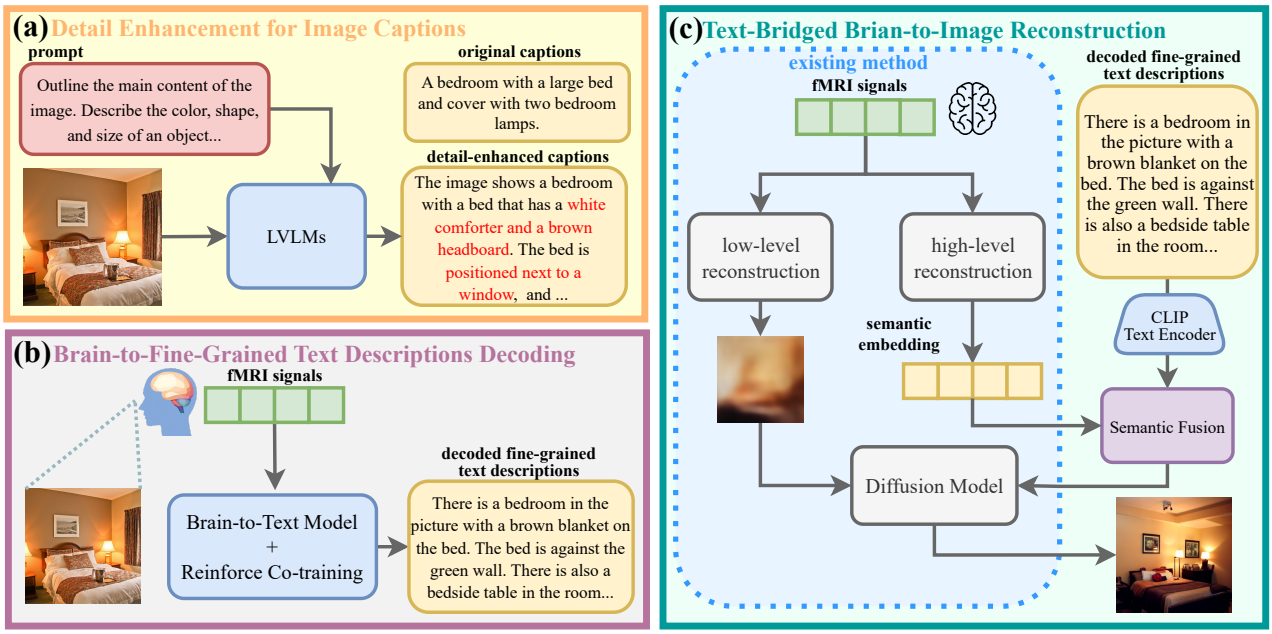


Figure 2: Overview of FgB2I. **(a)** Detail enhancement of visual stimuli captions through LVLMs. **(b)** The process of decoding fine-grained text descriptions from brain signals, with the details inside the blue box further illustrated in Figure 3. **(c)** The workflow for combining fine-grained text descriptions with existing methods, including semantic fusion through weighted average of text semantic embedding and the integration of text and image embeddings.

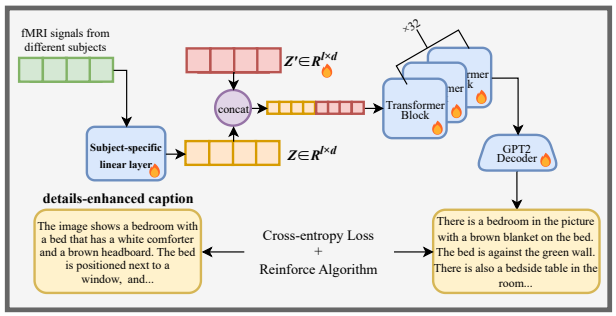


Figure 3: Diagram of the brain-to-text model structure and training. The flame represents the trainable components.

Fine-Grained Text Descriptions Decoding via Reinforced Co-Training

To achieve more fine-grained text decoding in the fine-grained semantic reconstruction process, a two-stage co-training approach is employed. In the first stage, we use cross-entropy loss (CE) to train the model to generate text descriptions from brain signals. This stage is crucial for training the GPT2 model to understand and decode the corresponding text descriptions. Here, our goal is to ensure that the decoded text could recognize the objects within the image and to evaluate the semantic similarity between the text and the image. Since the decoded text cannot provide gradient information for model training, we utilize the reinforce algorithm (Sutton et al., 1999) to design a loss function that would achieve these objectives. This algorithm is applied exclusively during the reinforce co-training stage.

The specific process involves applying softmax to the logits outputted by the GPT2 model to generate a probability distribution p. We then sample text sequences based on this distribution, and calculate reward scores for these sequences. These reward scores are used to guide the model towards to generate text that closely match the semantic content of the image. The loss function, which incorporates these reward scores, is formulated as follows:

$$\mathcal{L}(\theta) = \sum_{t=1}^{T} r_t \log p(a_t | s_t; \theta) \tag{1}$$

where T represents the length of the generated text, a_t represents the token sampled at step t, s_t represents the corresponding state, θ represents the model parameters, and r_t represents the reward obtained for the current text.

Three complementary metrics guide the decoding process (Fig.4): (1) object accuracy, ensuring the inclusion of correct objects; (2) text-image semantic similarity, and (3) image-image semantic similarity.

object accuracy : The extraction of nouns via TextBlob[1] from enhanced/decoded texts, followed by abstract noun filtering (*image*, *photo*, etc.) yields sets A and B. Inter-set congruence is quantified through Jaccard scoring (Niwattanakul, Singthongchai, Naenudorn, & Wanapu, 2013):

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

[1] https://textblob.readthedocs.io/en/dev/

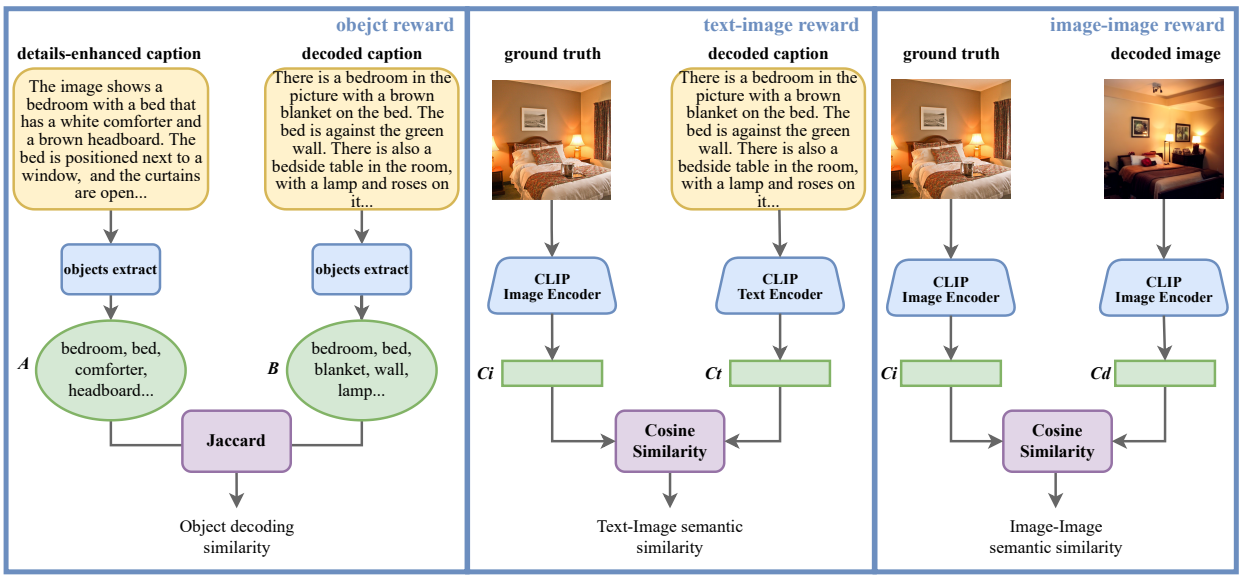


Figure 4: Three reward function calculation diagrams. **(Left)** Reward for evaluating the accuracy of decoded objects. **(Middle)** Semantic similarity between decoded text and visual stimuli, where C_i and C_t represent the corresponding CLIP embedding of the image and text. **(Right)** Semantic similarity between the reconstructed image and visual stimuli.

text-image semantic similarity : The semantic representations of images and textual descriptions C_i and C_t are obtained using CLIP text and image encoders. Subsequently, we measure the similarity between the two representations using cosine similarity:

$$r_{text-image} = \text{Cosine Similarity}(C_i, C_t) \tag{3}$$

image-image semantic similarity : CLIP image-space comparison enforces semantic coherence between reconstructed/original stimuli.

The final loss function is formulated as:

$$\mathcal{L} = \mathcal{L}_{CE} + \alpha\mathcal{L}_1 + \beta\mathcal{L}_2 + \gamma\mathcal{L}_3 \tag{4}$$

Here, α, β, and γ are trade-off factors that balance the importance of each reward in the training process, with $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$ calculated using formula (1).

Text-Bridged Brian-to-Image Reconstruction

Fine-grained text decoded from brain signals enhances diffusion-model reconstruction via semantic control but can compromise color or structural fidelity. To mitigate this, we integrate our approach into three fMRI-to-image pipelines: LDM, based on the Stable Diffusion model; and BrainDiffuser and MindEye, both built on Versatile Diffusion (Takagi & Nishimoto, 2023; Ozcelik & VanRullen, 2023; Scotti et al., 2023; Xu et al., 2023). These frameworks employ two-phase reconstruction:

Low-level. fMRI signals are mapped to VDVAE/VQVAE latent space (Child, 2020; Van Den Oord, Vinyals, et al., 2017), and then decoded into an initial guess image. **High-level.** Simultaneously, fMRI signals are projected into CLIP space to guide the semantic aspects of image generation. To balance the semantic information from existing methods with our fine-grained descriptions, we fuse the CLIP embeddings of decoded text with original high-level features. The resulting fused representation replaces the original high-level semantics during image reconstruction.

Experimental Settings

Dataset

In our study, we utilize a subset of the Natural Scenes Dataset (NSD) (Allen et al., 2022), a comprehensive collection of neuroimaging data obtained from eight participants using a 7-Tesla fMRI scanner. The dataset encompasses a total of 30–40 scanning sessions, during which each participant viewed three repetitions of 9000-10,000 distinct images sourced from the Microsoft COCO dataset. For our analysis, we focus on data from subjects 1, 2, 5, and 7, who completed the full set of imaging sessions. 982 images are common across all four subjects and are designated as the test dataset, while the remaining trials served as the training dataset. For the test dataset, we average the responses across the three trials associated with each image, whereas for the training dataset, we use the individual trials without averaging.

Implementation Details

In our experiments, we employ the CLIP ViT-B/32 model, and the GPT2-Base model. We utilize an 8-head and 32 layers transformer network, and set the fMRI prefix length l to 10, as well as the maximum length for generated text to 77 (the maximum encoding length of CLIP text encoder). The LLaVA-1.5-7b model is employed for details enhancement for image captions. For quantitative evaluation, we employed a range of metrics, including SSIM (Wang, Bovik, Sheikh,

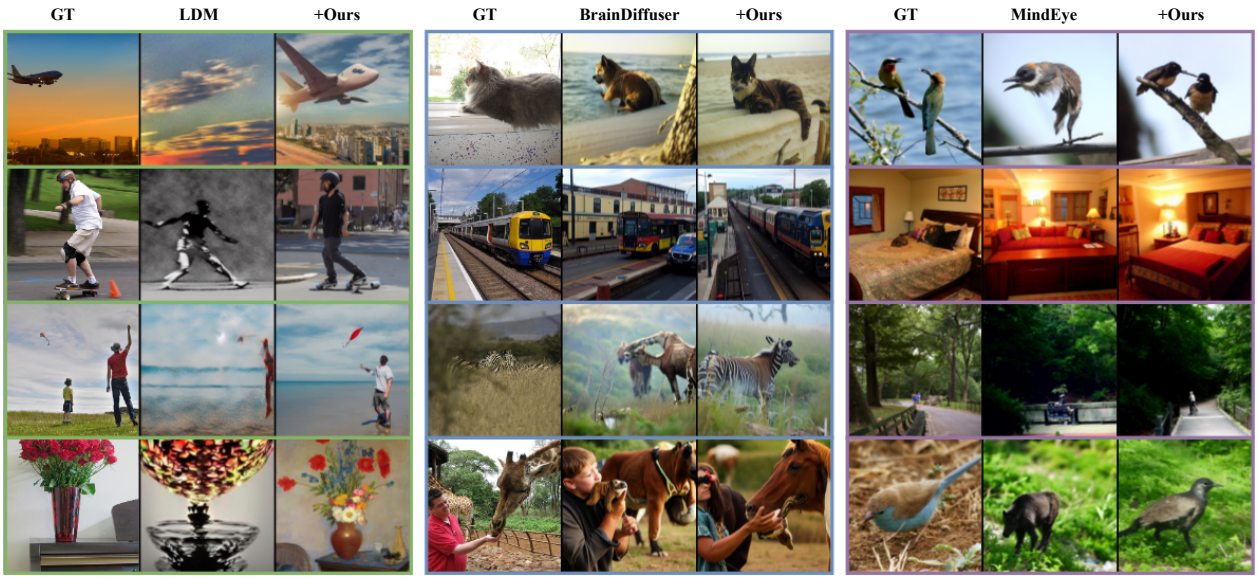


Figure 5: A comparison of the reconstructed image results between existing methods (LDM (Takagi & Nishimoto, 2023), BrainDiffuser (Ozcelik & VanRullen, 2023), and MindEye (Scotti et al., 2023)) and the results obtained when these methods are combined with our fine-grained text descriptions. GT denotes the corresponding ground truth stimulus image.

Method	High-Level				Low-Level			
	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓
LDM	/	/	83.0%	83.0%	76.0%	77.0%	/	/
LDM+DE	/	/	84.6%	85.1%	77.3%	79.4%	/	/
BrainDiffuser	.254	.356	**94.2%**	96.2%	87.2%	**91.5%**	.775	**.423**
BrainDiffuser+DE	**.255**	**.357**	93.6%	**96.3%**	**91.1%**	**91.5%**	**.737**	**.423**

Table 1: The reconstruction evaluation of detail-enhancement (denoted as DE in the table) on the LDM (Takagi & Nishimoto, 2023) and BrainDiffuser (Ozcelik & VanRullen, 2023) methods, using the same evaluation metrics as theirs. The values presented in the table are the mean of the assessment results from four participants.

Method	Low-Level				High-Level			
	PixCorr↑	SSIM↑	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑	Eff↓	SwAV↓
LDM	/	/	83.0%	83.0%	76.0%	77.0%	/	/
LDM+Ours	/	/	81.1%	88.3%	85.6%	86.2%	/	/
BrainDiffuser	.254	.356	94.2%	96.2%	87.2%	91.5%	.775	.423
BrainDiffuser+Ours	.260	.371	93.9%	96.4%	92.4%	92.2%	.710	.409
MindEye	**.309**	.323	94.7%	**97.8%**	93.8%	**94.1%**	.645	.367
MindEye+Ours	.305	**.354**	**94.8%**	**97.8%**	**94.3%**	93.8%	**.637**	**.360**

Table 2: A comparison of the reconstruction results for LDM (Takagi & Nishimoto, 2023), BrainDiffuser (Ozcelik & VanRullen, 2023), and MindEye (Scotti et al., 2023) methods combined with FgB2I's fine-grained text descriptions, using the same evaluation metrics as theirs. The values presented in the table are the mean of the assessment results from four participants.

& Simoncelli, 2004), PixCorr (Pixel-Wise Correlation), Eff (Tan & Le, 2019), SwAV (Caron et al., 2020), and Two-Way Identification Accuracy (Alexnet (Krizhevsky, Sutskever, & Hinton, 2012), and InceptionV3 (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016)), following the approach of (Scotti et al., 2023). SSIM and PixCorr assess low-level similarity, capturing structural and pixel-wise correlations, respectively. Eff and SwAV compute feature distances using pretrained net-

Method	Low-Level		High-Level	
	Alex(2)↑	Alex(5)↑	Incep↑	CLIP↑
LDM	77.7%	75.2%	67.4%	70.8%
+Ours (L_{CE})	85.1%	89.3%	85.4%	86.1%
+Ours (L_{CE},L_1)	84.8%	89.6%	85.9%	86.9%
+Ours (L_{CE},L_2)	84.4%	89.5%	85.4%	**87.1%**
+Ours (L_{CE},L_3)	85.2%	89.7%	**86.1%**	86.8%
+Ours (all L)	**85.2%**	**89.8%**	85.9%	86.9%

Table 3: The performance of FgB2I on the LDM method using different reward metrics for the loss function.

works to evaluate semantic consistency. Two-Way Identification Accuracy measures whether reconstructed images can be correctly matched to originals based on pretrained networks feature similarities.

During the Brain-to-fine-grained text training phase, we optimize the model using AdamW (Loshchilov & Hutter, 2019) with an initial learning rate of 1e-4 over 200 epochs in first stage. The second stage is conducted using the weights of this model from the first stage, with a reduced learning rate of 2e-6 for 10 epochs, maintaining the same other settings, and the parameters α, β, and γ are all set to 0.01.

Results and Analysis

Details Enhancement Results

To validate the effectiveness of details enhancement, we conduct experiments on two methods: LDM (Takagi & Nishimoto, 2023) and BrainDiffuser (Ozcelik & VanRullen, 2023). In this experiments, we substitute the original image descriptions with detail-enhanced captions generated by LVLMs, while maintaining all other experimental settings and parameters as reported in the original papers.

The results of these experiments are summarized in Table 1. The data clearly indicate that incorporating DE leads to improvements across various evaluation metrics, demonstrating that details enhancement is crucial. This suggests that the original captions may lack the necessary granularity to fully capture the intricate details present in the brain's representation of images. However, LVLMs inevitably hallucinate, producing inaccuracies that can sometimes limit the effectiveness of detail enhancement in image captions, though more advanced models can partially mitigate this issue.

Main Results

We apply fine-grained text descriptions to the reconstruction results of a total of three methods, LDM, BrainDiffuser, and MindEye. We first conduct a quantitative analysis and reported the comparison between the original methods and the methods enhanced with our approach, which are presented in Table 2. These results demonstrate the effectiveness of FgB2I in improving brain-to-image reconstruction.

The results clearly show that FgB2I leads to improvements across all methods, with the largest improvement observed in the LDM method, which solely uses text as the semantic control. The other two methods, which use both text and image for semantic control, show more limited improvements. A possible reason is that the combination of text and image semantic conditions in these methods makes the impact of improved text control on image reconstruction more limited.

To visually observe the fine-grained enhancement of image reconstruction by FgB2I, we present a comparison of some reconstructed images with those of the original methods in Figure 5. It can be visually observed from the figure that incorporating our improvements into various methods leads to more consistent reconstruction outcomes.

Ablation Analysis

During the reinforcement co-training phase for fine-grained text, we train the model using each individual reward loss. To clearly demonstrate the effect of each loss, we conduct the experiments on the LDM method for Subject 1, with the results presented in Table 3. The results show that L_1 and L_2 improve semantic alignment, as indicated by higher Incep and CLIP accuracy, but they slightly reduce visual similarity, as reflected in a lower Alex(2) accuracy. L_3 enhances high-level metrics such as Incep and CLIP while maintaining a good balance with low-level metrics.

Case Studies

The figure 5 illustrates a set of image reconstruction examples, contrasting the existing methods with those refined by incorporating our method for Subject 1. It can be observed from the results that the addition of our method to the LDM approach, which initially exhibited poorer reconstruction quality, yields clearer semantics and improved reconstruction outcomes. When applied to the higher-quality BrainDiffuser and MindEye methods, the integration of our fine-grained text facilitates the correction of semantic inconsistencies, leading to results that more closely align with the ground truth semantics. Furthermore, as evidenced by the third example from the MindEye method, our approach is capable of supplementing fine-grained details, such as including the railings on both sides of the road in our reconstruction, which were omitted in the original method.

Conclusion

We point out the limitations of using simple captions as semantic reconstruction targets in brain-to-image reconstruction tasks and propose employing LVLM to address this issue. Then, we propose FgB2I, a novel approach that enables the reconstruction of visual stimuli from fMRI signals with enhanced details, and fine-grained text descriptions. Experiments demonstrate FgB2I's capabilities in recovering fine-grained details lost by prior works, showcasing the potential of compensating for insufficient reference details and unifying multi-subject data through semantic decoding. We acknowledge the limitations of current fMRI data and the challenges in decoding neural signals, which inform our future work to improve the granularity of signal decoding. FgB2I provides new perspectives on understanding and reconstructing the intricate process of human visual cognition.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (No.62476127, No.62106105), the Natural Science Foundation of Jiangsu Province (No.BK20242039), the Basic Research Program of the Bureau of Science and Technology (ILF24001), the Fundamental Research Funds for the Central Universities (No.NJ2023032), the Scientific Research Starting Foundation of Nanjing University of Aeronautics and Astronautics (No.YQR21022), and the High Performance Computing Platform of Nanjing University of Aeronautics and Astronautics.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., ... others (2022). A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, *25*(1), 116–126.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, *33*, 9912–9924.

Chen, J., Qi, Y., & Pan, G. (2023). Rethinking visual reconstruction: Experience-based content completion guided by visual cues. , *202*, 4856–4866. Retrieved from https://proceedings.mlr.press/v202/chen23v.html

Child, R. (2020). Very deep vaes generalize autoregressive models and can outperform them on images. *arXiv preprint arXiv:2011.10650*.

Child, R. (2021). Very deep vaes generalize autoregressive models and can outperform them on images. In *9th international conference on learning representations, ICLR 2021, virtual event, austria, may 3-7, 2021*. OpenReview.net. Retrieved from https://openreview.net/forum?id=RLRXCV6DbEJ

Chun, M. M., & Johnson, M. K. (2011). Memory: Enduring traces of perceptual and reflective attention. *Neuron*, *72*(4), 520–535.

Glover, G. H. (2011). Overview of functional magnetic resonance imaging. *Neurosurgery Clinics*, *22*(2), 133–139.

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

Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S., Hudspeth, A. J., Mack, S., et al. (2000). *Principles of neural science* (Vol. 4). McGraw-hill New York.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In (Vol. 25).

Li, X. L., & Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. *CoRR*, *abs/2304.08485*. Retrieved from https://doi.org/10.48550/arXiv.2304.08485 doi: 10.48550/ARXIV.2304.08485

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *7th international conference on learning representations, ICLR 2019, new orleans, la, usa, may 6-9, 2019*. OpenReview.net. Retrieved from https://openreview.net/forum?id=Bkg6RiCqY7

Lu, Y., Du, C., Zhou, Q., Wang, D., & He, H. (2023). Minddiffuser: Controlled image reconstruction from human brain activity with semantic and structural diffusion. In A. El-Saddik et al. (Eds.), *Proceedings of the 31st ACM international conference on multimedia, MM 2023, ottawa, on, canada, 29 october 2023- 3 november 2023* (pp. 5899–5908). ACM. Retrieved from https://doi.org/10.1145/3581783.3613832 doi: 10.1145/3581783.3613832

Luck, S. J., & Ford, M. A. (1998). On the role of selective attention in visual perception. *Proceedings of the National Academy of Sciences*, *95*(3), 825–830.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nat.*, *518*(7540), 529–533. Retrieved from https://doi.org/10.1038/nature14236 doi: 10.1038/NATURE14236

Niwattanakul, S., Singthongchai, J., Naenudorn, E., & Wanapu, S. (2013). Using of jaccard coefficient for keywords similarity. In *Proceedings of the international multiconference of engineers and computer scientists* (Vol. 1, pp. 380–384).

Ozcelik, F., & VanRullen, R. (2023). Natural scene reconstruction from fmri signals using generative latent diffusion. *Scientific Reports*, *13*(1), 15666.

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

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Scotti, P. S., Banerjee, A., Goode, J., Shabalin, S., Nguyen, A., Cohen, E., ... Abraham, T. M. (2023). Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *CoRR*, *abs/2305.18274*. Retrieved from https://doi.org/10.48550/arXiv.2305.18274 doi: 10.48550/ARXIV.2305.18274

Sutton, R. S., McAllester, D., Singh, S., & Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, *12*.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *2016 IEEE conference on computer vision and pattern recognition, CVPR 2016, las vegas, nv, usa, june 27-30, 2016* (pp. 2818–2826). IEEE Computer Society. Retrieved from https://doi.org/10.1109/CVPR.2016.308 doi: 10.1109/CVPR.2016.308

Takagi, Y., & Nishimoto, S. (2023). High-resolution image reconstruction with latent diffusion models from human brain activity. In *IEEE/CVF conference on computer vision and pattern recognition, CVPR 2023, vancouver, bc, canada, june 17-24, 2023* (pp. 14453–14463). IEEE. Retrieved from https://doi.org/10.1109/CVPR52729.2023.01389 doi: 10.1109/CVPR52729.2023.01389

Tan, M., & Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning, ICML 2019, 9-15 june 2019, long beach, california, USA* (Vol. 97, pp. 6105–6114). PMLR. Retrieved from http://proceedings.mlr.press/v97/tan19a.html

Van Den Oord, A., Vinyals, O., et al. (2017). Neural discrete representation learning. *Advances in neural information processing systems*, *30*.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, *13*(4), 600–612. Retrieved from https://doi.org/10.1109/TIP.2003.819861 doi: 10.1109/TIP.2003.819861

Wang, Z., & Ji, H. (2022). Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *Thirty-sixth AAAI conference on artificial intelligence, AAAI 2022, thirty-fourth conference on innovative applications of artificial intelligence, IAAI 2022, the twelveth symposium on educational advances in artificial intelligence, EAAI 2022 virtual event, february 22 - march 1, 2022* (pp. 5350–5358). AAAI Press. Retrieved from https://doi.org/10.1609/aaai.v36i5.20472 doi: 10.1609/AAAI.V36I5.20472

Xi, N., Zhao, S., Wang, H., Liu, C., Qin, B., & Liu, T. (2023). Unicorn: Unified cognitive signal reconstruction bridging cognitive signals and human language. *arXiv preprint arXiv:2307.05355*.

Xia, R., Yin, C., & Li, P. (2024). Decoding the echoes of vision from fmri: Memory disentangling for past semantic information. In *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 2040–2052).

Xu, X., Wang, Z., Zhang, G., Wang, K., & Shi, H. (2023). Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 7754–7765).

Yin, C., Ye, Z., & Li, P. (2024). Language reconstruction with brain predictive coding from fmri data. *arXiv preprint arXiv:2405.11597*.

Zhang, Y., Han, K., Worth, R., & Liu, Z. (2020). Connecting concepts in the brain by mapping cortical representations of semantic relations. *Nature communications*, *11*(1), 1877.