# REINFORCEMENT LEARNING FOR BRAIN IMAGE CAPTIONING BY IMPOSING VISUAL POSTERIOR MATCHING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Decoding visual semantics from brain activity can provide a natural-language account of perceived stimuli. While existing approaches mainly focus on image classification, retrieval, or reconstruction, brain image captioning provides a compositional readout of objects, attributes, and relations. However, powerful vision-language priors can turn incomplete and noisy neural evidence into plausible yet stimulus-inaccurate descriptions. We propose Visual Posterior Matching (VPM), a reinforcement learning framework for brain image captioning. VPM first aligns neural representations with image tokens to initialize a brain-conditioned language policy, then trains a conditional flow model on paired image and caption features. During reinforcement learning, VPM constructs shared noisy anchors from the paired stimulus and rewards agreement between the clean visual-feature estimates induced by generated and reference captions at the same anchors. This target-anchored reward is designed to keep generation tied to stimulus-specific evidence while allowing diverse phrasings. Experiments on fMRI and EEG show that VPM improves core captioning metrics over supervised initialization and outperforms existing brain-captioning systems on the evaluated NSD dataset.

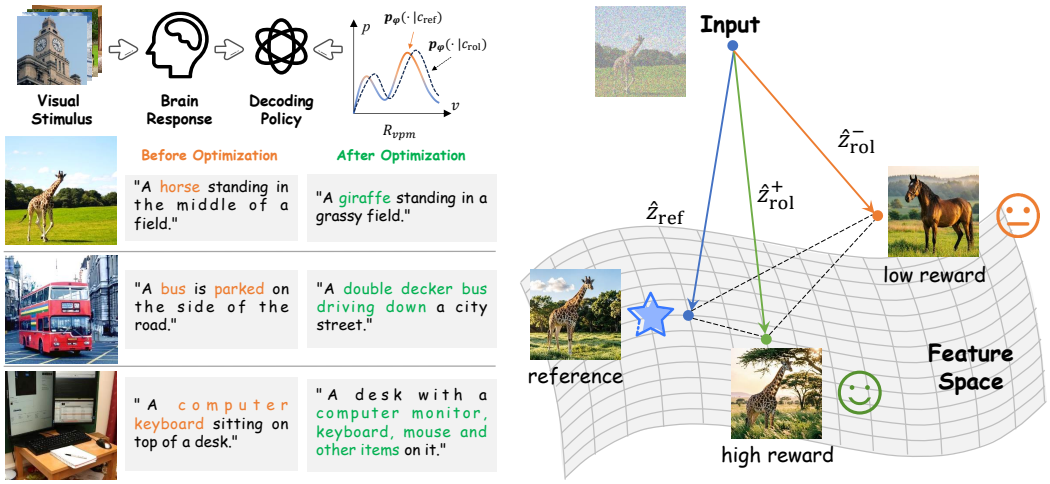
## INTRODUCTION

Translating visual experience reflected in human brain activity into natural language is a fundamental problem at the intersection of neural representation, visual understanding, and language generation. Unlike approaches that recover a coarse image, retrieve a candidate stimulus, or classify a visual category, brain image captioning produces readable, compositional, and interactive semantic content. It therefore provides a flexible interface for probing what visual information is represented in brain activity, studying neural representations in cognitive neuroscience, communicating visually grounded intent through brain–computer interfaces, and supporting assistive communication. Early visual brain decoding relied primarily on linear mappings, hand-crafted features, and discriminative models for stimulus identification or low-level reconstruction. More recently, deep networks and generative priors have mapped brain signals to visual representations and improved the semantic content and detail of reconstructed images (Takagi & Nishimoto, 2023; Lu et al., 2023). Building on pretrained vision–language models (VLMs), recent systems further extend this direction from producing a plausible image to describing a visual stimulus with open vocabulary (Ferrante et al., 2023; Chen et al., 2025b; Xia et al., 2024).

Large VLMs create an important opportunity for brain image captioning. Existing approaches typically project neural activity into visual or multimodal token spaces, then use a pretrained language model to read out an open-ended description (Ferrante et al., 2023; Chen et al., 2025b; Xia et al., 2024). This paradigm reuses powerful visual-semantic priors, allowing a language model to interpret brain signals that are limited, noisy, and highly subject-specific. At the same time, it makes faithfulness harder to preserve. A brain response is only an incomplete and indirect observation of a visual stimulus, and the same underlying visual evidence may admit multiple valid linguistic descriptions that emphasize different aspects of the scene. Maximum-likelihood training further rewards reproducing one reference sentence token by token. Under weak neural evidence and strong language priors, a policy can thus produce a common and well-formed sentence while omitting the



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.




**a. VPM System Overview** **b. Posterior-Matching Reward**

Figure 1: **Visual Posterior Matching (VPM) for stimulus-faithful brain image captioning.** *(a)* **VPM system overview and qualitative examples.** A visual stimulus evokes a brain response that is decoded by a brain-conditioned policy. VPM aligns neural and visual representations and optimizes the policy with a target-anchored visual-posterior reward. Relative to the initialized policy, the optimized policy recovers stimulus-defining content, including *giraffe* rather than *horse*, *double-decker bus* rather than a generic *bus*, and a *computer monitor* in the desk scene. *(b)* **Target-anchored posterior-matching reward.** At a shared noisy visual anchor, the frozen flow critic maps the reference and rollout captions to clean-feature $x$-predictions. Cosine agreement with the reference prediction assigns high reward to a compatible rollout and low reward to a mismatched rollout.

identity-defining content of the current stimulus. Representation alignment and generative readout address whether a language model can read brain signals; they do not fully explain why free generation should remain within the semantic range supported by the paired visual evidence. The first challenge is therefore to preserve stimulus identity during open-ended language generation.

Sequence-level reinforcement learning (RL) appears to offer a direct mechanism for addressing this challenge. It bridges the gap between teacher-forced likelihood training and sentence-level evaluation, and has improved image captioning through sequence-level objectives and self-critical learning (Ranzato et al., 2015; Rennie et al., 2017). Recent reward designs further leverage visual–language similarity, learned vision–language rewards, and cycle-consistency signals to evaluate generated captions (Bahng et al., 2025). However, these rewards typically measure linguistic quality, global image–text relevance, or generic preferences, rather than whether a candidate caption still explains the visual evidence associated with a specific training sample. As a result, the policy may drift while achieving a higher proxy score, for example, as illustrated in Fig. 1, changing from "a red car" to a more common but inaccurate description such as "a car on the street." Moreover, jointly updating the brain encoder and language policy with a mismatched reward may propagate noisy gradients back into an already aligned brain–visual interface. The second challenge is therefore to construct a sequence-level reward that is directly anchored to the paired visual stimulus, tolerates valid linguistic reformulations, and remains stable during policy optimization.

In this paper, we propose *Visual Posterior Matching* (VPM), a framework that uses the visual evidence of the paired stimulus to guide sequence-level brain image captioning. We first establish a brain-side interface that a frozen VLM can read: a trainable brain encoder aligns its tokens to image tokens while a frozen language policy is supervised to decode the reference caption. We then train a conditional flow critic on paired image and caption features. The critic learns the local velocity of noisy visual features conditioned on a caption and directly recovers a conditional denoised visual feature (Lipman et al., 2023). During RL, VPM constructs shared noisy visual anchors from the target image at multiple noise scales and compares the denoised features induced by a rollout caption and the reference caption at the same anchors. A rollout receives higher reward when its $x$-predictions agree more closely with the reference. The paired image determines the noisy an-

chors, while the reference caption specifies the comparison target. The reward measures agreement between the resulting visual-feature estimates rather than token-level correspondence between captions. We freeze the brain encoder and visual posterior critic, and use KL-regularized group relative policy optimization (GRPO) to refine only the language readout policy (Guo et al., 2025). This design confines RL to the stage where it can improve generation without disrupting the established brain–visual alignment.

We evaluate VPM using two VLM backbones on four NSD subjects and one THINGS-EEG subject. Across the reported settings, VPM increases BLEU-4 and CIDEr relative to supervised initialization, while changes in other metrics depend on the setting. Comparisons with DPO and CLIP-RL start from the same Qwen2.5-VL base policy. Additional experiments examine x-prediction versus ODE integration, critic pretraining scale, and image–text encoder choices.

In summary, our contributions are as follows:

- We formulate sequence-level brain image captioning as a visual-posterior preservation problem, in which the training image defines the acceptable semantic variation of a caption.
- We introduce VPM, which uses a conditional-flow critic and target-anchored $x$-prediction comparison to guide captions toward the visual evidence of the paired stimulus.
- We demonstrate that VPM yields consistent improvements over supervised initialization and strong results against existing brain-captioning baselines across subjects.

## Related Work

### Brain Decoding

Brain-to-image decoding began with fMRI-based identification and reconstruction from linear mappings or hand-crafted representations (Kay et al., 2008; Miyawaki et al., 2008). Later work introduced deep reconstruction and self-supervised feature learning (Shen et al., 2019; Beliy et al., 2019). Recent methods align neural activity with pretrained visual or vision-language representations and use generative priors, especially diffusion models, for reconstruction and retrieval-augmented decoding (Takagi & Nishimoto, 2023; Lu et al., 2023; Scotti et al., 2023; 2024). EEG studies similarly use cross-modal alignment and guided generation for visual retrieval and reconstruction (Du et al., 2023; Li et al., 2024; Wu et al., 2025; Zheng et al., 2026).

Brain-to-text decoding targets semantic readout rather than pixel recovery. Early systems projected neural activity into visual or multimodal latent spaces and used pretrained captioning models or joint image–text pipelines to generate sentences (Ferrante et al., 2023; Mai & Zhang, 2023). More recent work explores dedicated neural language decoders, unified multimodal decoders, and models that couple visual reconstruction with language interaction (Shen et al., 2024; Chen et al., 2025b;a; Xia et al., 2024; Qiu et al., 2025; Ye et al., 2026). Most methods align brain signals with pretrained multimodal representations and rely on external language priors for open-ended caption generation.

### Reinforcement Learning for Image Captioning

Reinforcement learning is widely used in image captioning to bridge token-level likelihood training and sequence-level evaluation. Early work optimized non-differentiable caption metrics through sequence-level, self-critical, and actor-critic training (Ranzato et al., 2015; Rennie et al., 2017; Ren et al., 2017; Zhang et al., 2017), and later extensions considered off-policy updates and weaker-supervision settings (Shi et al., 2021; Chen et al., 2017). More recent studies focus on reward design, using visual-semantic similarity, CLIP alignment, learned vision-language reward models, cycle consistency, and completeness-oriented objectives, while direct preference optimization provides a reference-policy alternative (Cho et al., 2022; Gaur et al., 2024; Dzabraev et al., 2024; Bahng et al., 2025; Tang et al., 2026; Jia et al., 2026; Rafailov et al., 2023). These methods differ in how they assign credit over a sequence and what signals they optimize, but each relies on the reward to specify the behavior that receives reinforcement. The central issue remains reward mismatch, since proxy rewards may improve fluency or benchmark scores without fully constraining fine-grained faithfulness to image content.

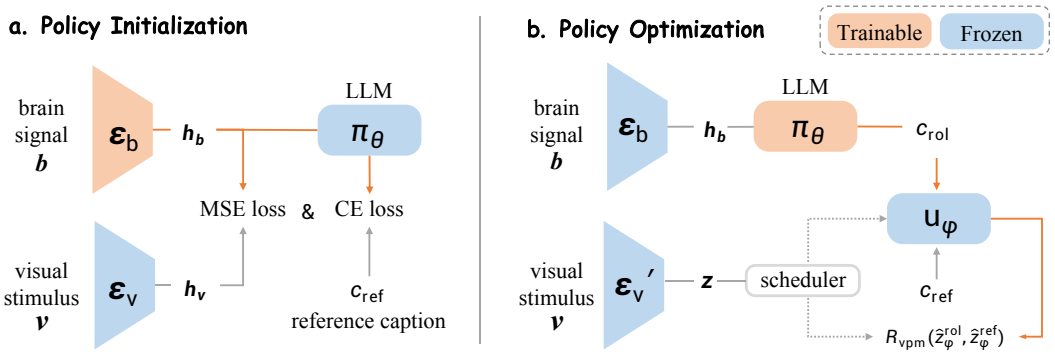



Figure 2: **Method overview. (a) Policy initialization.** The trainable brain encoder $\epsilon_b$ maps $b$ to $h_b$. MSE aligns $h_b$ with image tokens $h_v = \epsilon_v(v)$ from the frozen image encoder, while CE evaluates the frozen policy $\pi_\theta$ against $c_{\text{ref}}$. **(b) Policy optimization.** The brain encoder $\epsilon_b$, visual feature extractor $\epsilon_v'$, and posterior model $u_\varphi$ are frozen; only the LLM policy $\pi_\theta$ is trainable. The scheduler creates $K = 5$ target-derived noisy anchors. At each anchor, $u_\varphi$ produces clean-feature estimates $\hat{z}_\varphi^{\text{rol}}$ and $\hat{z}_\varphi^{\text{ref}}$ from the rollout and reference captions; their weighted cosine agreement, $R_{\text{vpm}}(\hat{\boldsymbol{z}}_\varphi^{\text{rol}}, \hat{\boldsymbol{z}}_\varphi^{\text{ref}})$, updates $\pi_\theta$.

### Conditional Diffusion and Flow Matching

Diffusion models represent complex implicit conditional distributions through a time-dependent denoising quantity along a Gaussian corruption path (Ho et al., 2020; Song et al., 2021b). At a noisy state, noise ($\epsilon$), clean-data ($z_0$), and velocity prediction are reparameterizations of the same conditional denoiser, whereas score prediction is its noise-scaled form; these choices change the regression target but not the represented distribution (Song et al., 2021a;b). Conditional variants specify distributions with side information (Shi et al., 2022; Chung et al., 2023). Flow matching and rectified flows directly regress a probability path's velocity, providing simulation-free targets, while stochastic interpolants unify flow and diffusion (Lipman et al., 2023; Liu et al., 2023; Albergo et al., 2025). These formulations separate local conditional estimation from numerical generation of a complete trajectory, and their practical behavior depends on the regression target, noise schedule, and inference procedure. In conditional multimodal settings, text can define a distribution over visual features; under an affine path, velocity predictions admit a $z_0$ form at common noisy anchors, providing a common feature space for comparing conditional denoised estimates.

## Method

### Problem Setup and Policy Initialization

We consider a training set $\mathcal{D} = \{(b, v, c_{\text{ref}})\}$, where $b$ is a brain response, $v$ is its paired visual stimulus, and $c_{\text{ref}}$ is a reference caption for that stimulus. The brain response can be an fMRI volume or an EEG trial, depending on the dataset. At inference time, the caption policy receives $b$ only. The paired image and reference caption are used during training: $v$ supplies image-token supervision and the target visual feature used by the reward, while $c_{\text{ref}}$ supervises policy initialization and conditions the posterior comparison.

We use a trainable brain encoder $\epsilon_b$, a frozen VLM image encoder $\epsilon_v$, and an LLM policy $\pi_\theta$. For both fMRI and EEG, $\epsilon_b$ is the one-layer Q-Former-style BrainXS encoder from UMBRAE (Xia et al., 2024) with a hidden size of 1,024. The brain encoder maps a neural observation to the VLM token space, so that its output has the same token structure as the image representation. Specifically,

$$h_b = \epsilon_b(b), \qquad h_v = \epsilon_v(v), \qquad \pi_\theta(c \mid h_b). \tag{1}$$

During policy initialization, we freeze $\epsilon_v$ and $\pi_\theta$ and train only $\epsilon_b$. The objective aligns brain tokens with image tokens and uses the frozen policy to decode the reference caption,

$$\mathcal{L}_{\text{init}} = \|h_b - h_v\|_F^2 - \alpha \log \pi_\theta(c_{\text{ref}} \mid h_b). \tag{2}$$

The first term follows prior brain-to-image-token alignment (Ferrante et al., 2023; Xia et al., 2024), and the second term is the caption negative log-likelihood. We use $\alpha = 0.25$. This stage defines the reference policy $\pi_{\text{ref}}(\cdot \mid h_b) = \pi_\theta(\cdot \mid h_b)$.

### Conditional Posterior Model

A caption can describe several compatible visual concepts and thus induces a conditional distribution over the corresponding visual features. Directly comparing these distributions is impractical; conditional flow matching instead predicts their clean-feature posterior means at shared noisy features under the MSE objective.

We train a conditional posterior model on paired image and caption features. A fixed image feature extractor maps the paired image to

$$z = \epsilon'_v(v). \tag{3}$$

Let $p_0(z, c)$ be the joint distribution of visual features and captions. The posterior model represents the conditional distribution $p_0(z \mid c)$ along a data-to-noise affine path (Lipman et al., 2023),

$$\begin{aligned} z_\sigma &= (1-\sigma)z_0 + \sigma\xi, \\ \xi &\sim \mathcal{N}(0, I), \qquad \sigma \in (0, 1), \end{aligned} \tag{4}$$

which defines the noisy conditional posterior

$$p_\sigma(z \mid c) = \int \mathcal{N}\big(z;\, (1-\sigma)z_0, \sigma^2 I\big)\, p_0(z_0 \mid c)\, dz_0. \tag{5}$$

For a pair $(z_0, c) \sim p_0$, the path velocity is the tractable target $\xi - z_0$. We train the conditional posterior model $u_\varphi(z, c, \sigma)$ by flow matching (Lipman et al., 2023),

$$\begin{aligned} \mathcal{L}_{\text{CFM}}(\varphi) = \mathbb{E}_{(z_0,c)\sim p_0,\, \sigma\sim\rho,\, \xi\sim\mathcal{N}(0,I)} \\ \Big[\|u_\varphi(z_\sigma, c, \sigma) - (\xi - z_0)\|_2^2\Big]. \end{aligned} \tag{6}$$

During training, we shift a uniformly sampled base level $\tau$ according to

$$\sigma = \frac{s\tau}{1 + (s-1)\tau}, \qquad s = 3, \tag{7}$$

and evaluate Equation 6 at the shifted $\sigma$. This shift places greater emphasis on modeling the noisy regions of the visual-feature path. We implement $u_\varphi$ as a compact four-layer MLP with timestep- and text-conditioned adaptive RMS normalization to predict the velocity in Equation 6.

### Visual Posterior Matching Reward

Equation 6 learns a conditional velocity field. Under the linear path in Equation 4, it yields the following clean-feature estimate for a noisy feature $z_\sigma$ and caption $c$:

$$\hat{z}_\varphi(z_\sigma, c, \sigma) = z_\sigma - \sigma u_\varphi(z_\sigma, c, \sigma) \approx \mathbb{E}[z_0 \mid z_\sigma, c]. \tag{8}$$

For each conditional trajectory, $z_0 = z_\sigma - \sigma(\xi - z_0)$; consequently, the MSE-optimal velocity gives the conditional posterior mean shown above. We write this estimate as $\hat{z}_\varphi$. Because every estimate is expressed in the clean feature space, the reward compares the same variable, $z_0$, across noise levels.

Given the paired target feature $z$, the rollout caption $c_{\text{rol}}$, and the reference caption $c_{\text{ref}}$, we evaluate the shifted schedule at $K = 5$ fixed slots,

$$t_i \in \{0.1, 0.2, 0.3, 0.4, 0.5\}, \qquad i = 1, \ldots, K, \tag{9}$$

where the corresponding $\sigma_i$ follows Equation 7 with $\tau = 1 - t_i$. We draw $\xi_i \overset{\text{iid}}{\sim} \mathcal{N}(0, I)$ to form $z_i = (1-\sigma_i)z + \sigma_i\xi_i$. For each response, the same $\{z_i\}_{i=1}^K$ is used for the rollout and reference captions, and is shared by all $G$ rollouts. Thus, within a group, the anchor noise is held fixed when comparing caption conditions. Write $\hat{z}^{\text{ref}}_{\varphi,i} = \hat{z}_\varphi(z_i, c_{\text{ref}}, \sigma_i)$, $\hat{z}^{\text{rol}}_{\varphi,i} = \hat{z}_\varphi(z_i, c_{\text{rol}}, \sigma_i)$, $\hat{\boldsymbol{z}}^{\text{ref}}_\varphi = \{\hat{z}^{\text{ref}}_{\varphi,i}\}_{i=1}^K$, and $\hat{\boldsymbol{z}}^{\text{rol}}_\varphi = \{\hat{z}^{\text{rol}}_{\varphi,i}\}_{i=1}^K$. With exponentially decaying weights $w_i = 2^{-(i-1)}$, we define

$$R_{\text{vpm}}(\hat{\boldsymbol{z}}^{\text{rol}}_\varphi, \hat{\boldsymbol{z}}^{\text{ref}}_\varphi) = \sum_{i=1}^K w_i \frac{\big\langle \hat{z}^{\text{rol}}_{\varphi,i}, \hat{z}^{\text{ref}}_{\varphi,i} \big\rangle}{\big\|\hat{z}^{\text{rol}}_{\varphi,i}\big\|_2 \, \big\|\hat{z}^{\text{ref}}_{\varphi,i}\big\|_2}. \tag{10}$$

Table 1: **Quantitative comparison on NSD subject 1.** Ours (Base) denotes the supervised policy initialization, while Ours (+VPM) applies visual posterior matching during policy optimization. All metrics are reported on a 0–100 scale, and higher is better. The best result in each column is shown in bold; – denotes an unreported result.

| Method | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE-L | CIDEr | SPICE | SentenceScore |
|---|---|---|---|---|---|---|---|---|---|
| SDRecon | 36.21 | 17.11 | 7.72 | 3.43 | 10.03 | 25.13 | 13.83 | 5.02 | 39.97 |
| OneLLM | 47.04 | 26.97 | 15.49 | 9.51 | 13.55 | 35.05 | 22.99 | 6.26 | 37.73 |
| BrainCap | 55.96 | 36.21 | 22.70 | 14.51 | 16.68 | 40.69 | 41.30 | 9.06 | 50.57 |
| MindEye2 | 54.82 | 38.60 | 26.49 | 18.16 | 17.54 | 43.77 | 55.70 | 10.97 | 56.85 |
| NeuroVLA | 57.19 | 37.17 | 23.78 | 15.85 | 18.60 | 36.67 | 49.51 | 12.39 | – |
| UMBRAE | 59.44 | 40.48 | 27.66 | 19.03 | 19.45 | 43.71 | 61.06 | 12.79 | 57.75 |
| VINDEX | 60.00 | 40.72 | 27.57 | 18.91 | 19.51 | 43.78 | 60.32 | 12.84 | – |
| MindLLM | 61.75 | 42.84 | 29.86 | 21.24 | 19.54 | 45.82 | 60.97 | 11.79 | – |
| DMM | 61.53 | 43.45 | 30.21 | 21.92 | 20.71 | 46.45 | 63.88 | – | – |
| Ours (Base) | 64.12 | 44.95 | 30.78 | 21.37 | 21.26 | 47.16 | 67.99 | 13.34 | 58.76 |
| **Ours (+VPM)** | **66.68** | **48.38** | **34.55** | **24.82** | **21.62** | **48.91** | **74.00** | **14.16** | **61.25** |

Table 2: **Reinforcement learning on the base model across fMRI and EEG.** Base denotes the policy after policy initialization only, and Ours denotes the VPM-optimized policy. fMRI results are means over NSD subjects 1, 2, 5, and 7; EEG results use THINGS-EEG subject 8. All metrics are reported on a 0–100 scale, and higher is better.

| Backbone | Bench | Method | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE-L | CIDEr | SPICE | SentenceScore |
|---|---|---|---|---|---|---|---|---|---|---|---|
| SmolVLM2 2.2B | NSD | Base | 61.71 | 43.47 | 29.99 | 20.79 | 20.25 | 46.28 | 63.37 | 12.56 | 57.11 |
| | | Ours | 63.01(↑1.3) | 44.72(↑1.3) | 31.12(↑1.1) | 21.75(↑1.0) | 19.96(↓0.3) | 46.44(↑0.2) | 64.44(↑1.0) | 12.54(↓0.02) | 58.38(↑1.3) |
| Qwen2.5 VL-3B | | Base | 63.53 | 44.22 | 30.41 | 21.17 | 20.80 | 46.71 | 66.66 | 12.97 | 57.96 |
| | | Ours | 65.78(↑2.3) | 47.44(↑3.2) | 33.85(↑3.4) | 24.33(↑3.2) | 21.22(↑0.4) | 48.36(↑1.7) | 72.26(↑5.6) | 13.73(↑0.8) | 60.17(↑2.2) |
| SmolVLM2 2.2B | THINGS -EEG | Base | 28.19 | 15.06 | 7.12 | 3.74 | 9.14 | 26.51 | 17.58 | 2.46 | 30.87 |
| | | Ours | 28.99(↑0.8) | 16.03(↑1.0) | 8.57(↑1.5) | 5.16(↑1.4) | 9.79(↑0.7) | 27.41(↑0.9) | 24.58(↑7.0) | 3.65(↑1.2) | 31.05(↑0.2) |
| Qwen2.5 VL-3B | | Base | 31.89 | 17.99 | 9.03 | 4.86 | 10.38 | 28.89 | 22.22 | 3.95 | 32.52 |
| | | Ours | 30.56(↓1.3) | 17.99(0.0) | 9.87(↑0.8) | 5.49(↑0.6) | 10.39(↑0.01) | 28.80(↓0.1) | 24.89(↑2.7) | 4.78(↑0.8) | 33.23(↑0.7) |

Equation 10 compares the directions of estimated conditional posterior means at shared target-derived anchors. We use visual posterior matching to denote this directional surrogate, rather than matching complete conditional distributions. The reward does not constrain feature norms or higher-order moments, and agreement at the sampled anchors does not establish distributional equivalence.

Figure 1*(b)*illustrates this construction. We use $x$-prediction as the default reward computation; Table 3 includes an ODE-integration VPM reference.

After policy initialization, we freeze $\epsilon_b$, $\epsilon'_v$, and $u_\varphi$, initialize $\pi_\theta$ from $\pi_{\mathrm{ref}}$, and optimize the caption policy with KL-regularized GRPO (Guo et al., 2025),

$$\mathcal{L}_{\mathrm{rl}} = -\mathbb{E}_{(b,v,c_{\mathrm{ref}})\sim\mathcal{D},\, c_{\mathrm{rol}}\sim\pi_\theta}\Big[R_{\mathrm{vpm}}(\hat{\boldsymbol{z}}_\varphi^{\mathrm{rol}}, \hat{\boldsymbol{z}}_\varphi^{\mathrm{ref}}) - \beta\, D_{\mathrm{KL}}(\pi_\theta(\cdot \mid h_b) \,\|\, \pi_{\mathrm{ref}}(\cdot \mid h_b))\Big]. \tag{11}$$

We set the KL coefficient to $\beta = 10^{-3}$.

For each response, GRPO samples $G = 8$ rollout captions, standardizes their rewards within the group, and applies the standard clipped policy objective.

# EXPERIMENTS

## EXPERIMENTAL DETAILS

**Datasets** For fMRI, we use responses from subjects 1, 2, 5, and 7 of the Natural Scenes Dataset (NSD) (Allen et al., 2022), collected while participants viewed natural images. Each subject contributes 24,980 training samples and 982 test images; for evaluation, we average the three responses




Figure 3: **Qualitative comparison of brain image captions.** We compare MindEye2, UMBRAE, and our VPM-optimized policy on representative visual stimuli. Green text denotes visually correct elements, while orange text denotes visually incorrect elements.

to each test image. For EEG, we use subject 8 from THINGS-EEG (Gifford et al., 2022; Grootswagers et al., 2022), with 1,654 training and 200 test object concepts. We generate its captions with Qwen/Qwen3-VL-8B-Instruct (Bai et al., 2025a). Together, NSD and THINGS-EEG provide rich visual stimuli and widely used benchmarks for evaluating brain decoding across complementary fMRI and EEG measurements. For both modalities, we train and evaluate a separate model within each subject; no data or parameters are shared across subjects. More details are in the supplementary material.

**Models and feature extractors** We adopt Qwen2.5-VL-3B-Instruct (Bai et al., 2025b) as the default vision-language-model backbone and additionally use HuggingFaceTB/SmolVLM2-2.2B-Instruct (Marafioti et al., 2025). Before policy optimization, we supervised fine-tune each VLM on the image–caption pairs in the corresponding dataset using rank-8 LoRA. Unless otherwise stated, image and text features are extracted by the image encoder and corresponding text encoder of OpenCLIP ViT-H/14 (Radford et al., 2021; Schuhmann et al., 2022), respectively. In encoder ablations, H/14 denotes OpenCLIP ViT-H/14, MPNet denotes MPNet-base, and DreamSim denotes the DreamSim SynCLR ViT-B/16 encoder (Fu et al., 2023).

### Quantitative Evaluation

**Comparison with baselines** We compare VPM with SDRecon, OneLLM, BrainCap, MindEye2, NeuroVLA, UMBRAE, VINDEX, MindLLM, and DMM (Takagi & Nishimoto, 2023; Han et al., 2024; Ferrante et al., 2023; Scotti et al., 2024; Shen et al., 2024; Xia et al., 2024; Xia & Oztireli, 2025; Qiu et al., 2025; Ye et al., 2026) on the NSD subject-1 split in Table 1. The initialized policy is competitive with the strongest prior systems, and VPM establishes the best result in every reported column. Its aligned gains in higher-order BLEU, CIDEr, SPICE, and SentenceScore show that the improvement extends beyond lexical overlap to reference-specific composition and sentence-level semantics.

**Policy optimization** Table 2 evaluates VPM with both VLM backbones on four-subject fMRI averages and THINGS-EEG subject 8. On fMRI, VPM improves higher-order caption quality for both backbones, with Qwen2.5-VL showing the broadest gains across the reported measures. The EEG results follow the same pattern on CIDEr, SPICE, and SentenceScore, while lower-order lexical

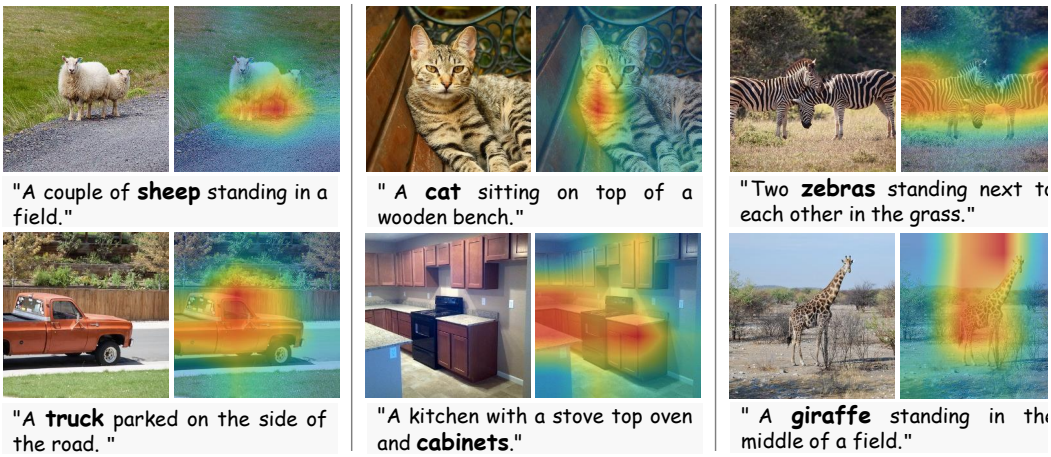



Figure 4: **Noun-conditioned brain–image grounding.** Each pair shows a subject-1 stimulus, its caption generated by our policy, and the gradient-weighted symmetric correspondence between brain and visual tokens for the bolded noun. Warm colors indicate stronger correspondence; cool colors indicate weaker correspondence.

Table 3: **Comparison of post-training objectives.** Base denotes the Qwen2.5-VL policy after policy initialization. Each optimized method starts from this same policy. Results are four-subject averages.

| Method | BLEU-4 | ROUGE-L | CIDEr | SPICE |
|---|---|---|---|---|
| Base | 21.17 | 46.71 | 66.66 | 12.97 |
| DPO | 21.99 | 47.11 | 67.95 | 12.95 |
| CLIP-RL | 22.94 | 47.37 | 70.32 | 13.39 |
| VPM (ODE integration) | 23.76 | 48.13 | 71.62 | **13.74** |
| VPM ($x$-prediction) | **24.33** | **48.36** | **72.26** | 13.73 |

metrics vary more across backbones. These trends show that VPM transfers across model families and neural modalities while most consistently improving reference-specific semantic content.

Table 3 compares objectives from the same Qwen2.5-VL base policy. DPO (Rafailov et al., 2023) remains close to the initialized policy, whereas CLIP-RL (Cho et al., 2022) provides a moderate improvement. Both VPM variants give a broader gain than these alternatives. The $x$-prediction variant is strongest on BLEU-4, ROUGE-L, and CIDEr, while full sampling has a marginal advantage on SPICE. We therefore use $x$-prediction by default because it evaluates the learned field directly, without 20-step integration at each anchor.

### Qualitative Evaluation

We also conduct visual comparisons with MindEye2 and UMBRAE. As shown in Figure 3, the baseline captions often replace the salient subject with a plausible but mismatched category, such as a bird for a fire hydrant or giraffes for zebras. Our captions retain the main objects and relations in these examples, including the hydrant, cow, rider on a horse, train on tracks, and pair of zebras, although they can still introduce unsupported details such as a fence.

We further examine noun-level grounding for subject 1. Figure 4 projects the gradient importance of each bolded noun from brain tokens onto the image-token grid through symmetric correspondence. The localized responses follow the sheep, cat, truck, cabinets, zebras, and giraffe across distinct scenes, indicating that the brain representation retains spatially resolved semantic information used by the policy.

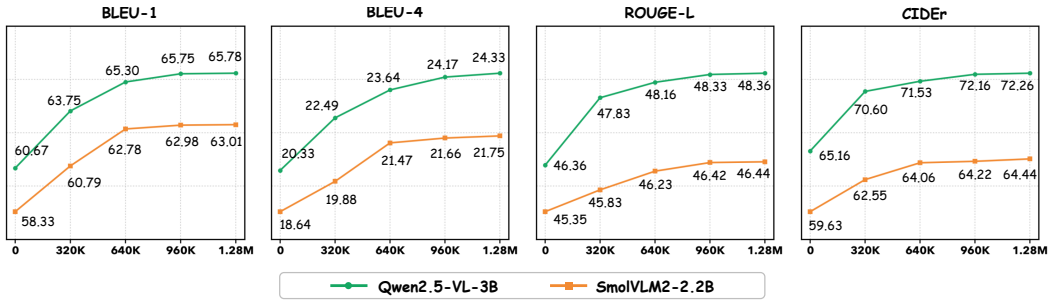



Figure 5: **Scaling ImageNet pretraining data for the flow critic.** Performance of VPM-optimized Qwen2.5-VL and SmolVLM2 policies as the pretraining corpus grows. Each curve begins with no pretraining (0) and ends with the original 1.28M-pair pretraining setting. Values are four-subject NSD averages.

Table 4: **Encoder ablation with $x$-prediction.** We vary the image and text encoders while holding the reward computation fixed. Results are four-subject averages.

| Image encoder | Text encoder | BLEU-4 | ROUGE-L | CIDEr | SPICE |
|---|---|---|---|---|---|
| H/14 | H/14 | 24.33 | 48.36 | 72.26 | 13.73 |
| | MPNet | 22.00 | 47.52 | 69.44 | 14.18 |
| DreamSim | H/14 | 22.24 | 47.72 | 70.09 | 13.84 |
| | MPNet | 21.64 | 47.38 | 68.63 | 14.17 |

### Ablative Study

Unless otherwise specified, ablations use Qwen2.5-VL-3B with VPM and report means over NSD subjects 1, 2, 5, and 7.

**Flow-critic pretraining** Figure 5 traces flow-critic pretraining from no pretraining to 1.28M ImageNet pairs. Both backbones improve consistently as the corpus grows. The curves rise most sharply in the early stage and become more gradual after 640K pairs, particularly for SmolVLM2. This shared trend across all four measures shows that the critic benefits from broader visual pretraining, while the later data primarily refines an already useful representation.

**Image and text encoders** The matched OpenCLIP ViT-H/14 image–text space gives the strongest overall performance in Table 4. A shared contrastive geometry lets the flow field connect linguistic conditions to visual directions without learning an additional cross-encoder calibration from limited subject data. DreamSim emphasizes perceptual similarity, while MPNet provides a stronger language-only semantic space; in this ablation, these alternatives slightly improve SPICE in several settings but yield lower BLEU-4, ROUGE-L, and CIDEr than the matched H/14 configuration.

## Conclusion

In this paper, we propose Visual Posterior Matching (VPM) for brain image captioning. VPM keeps policy optimization tied to the visual evidence paired with each brain response, improving stimulus-specific captions while permitting valid linguistic variation. This is achieved by a conditional flow critic that produces caption-conditioned clean-feature estimates at shared noisy anchors. Comparing the estimates from rollout and reference captions provides the policy reward, while freezing the brain-side interface confines optimization to the language readout. Experiments on NSD and THINGS-EEG demonstrate consistent improvements across Qwen2.5-VL and SmolVLM2. The quantitative comparisons, scaling study, and qualitative examples further show that the objective improves caption quality while retaining salient objects and relations from the stimulus. These results establish VPM as an effective sequence-level objective for brain-conditioned caption generation.

## REFERENCES


Michael Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

Hyojin Bahng, Caroline Chan, Fredo Durand, and Phillip Isola. Cycle consistency as reward: Learning image-text alignment without human preferences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22934–22946, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Roman Beliy, Guy Gaziv, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fmri. *Advances in Neural Information Processing Systems*, 32, 2019.

Jiaxuan Chen, Yu Qi, Yueming Wang, and Gang Pan. Bridging the gap between brain and machine in interpreting visual semantics: Towards self-adaptive brain-to-text decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21938–21948, 2025a.

Jiaxuan Chen, Yu Qi, Yueming Wang, and Gang Pan. Mindgpt: Interpreting what you see with non-invasive brain recordings. *IEEE Transactions on Image Processing*, 2025b.

Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan-Ting Hsu, Jianlong Fu, and Min Sun. Show, adapt and tell: Adversarial training of cross-domain image captioner. In *Proceedings of the IEEE international conference on computer vision*, pp. 521–530, 2017.

Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. Fine-grained image captioning with clip reward. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 517–527, 2022.

Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=OnD9zGAGT0k.

Changde Du, Kaicheng Fu, Jinpeng Li, and Huiguang He. Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10760–10777, 2023.

Maksim Dzabraev, Alexander Kunitsyn, and Andrei Ivaniuta. Vlrm: Vision-language models act as reward models for image captioning. *arXiv preprint arXiv:2404.01911*, 2024.

Matteo Ferrante, Furkan Ozcelik, Tommaso Boccato, Rufin VanRullen, and Nicola Toschi. Brain captioning: Decoding human brain activity into images and text. *arXiv preprint arXiv:2305.11560*, 2023.

Stephanie Fu, Netanel Yakir Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=DEiNSfh1k7.

Manu Gaur, Darshan Singh, and Makarand Tapaswi. No detail left behind: Revisiting self-retrieval for fine-grained image captioning. *arXiv preprint arXiv:2409.03025*, 2024.

Alessandro T Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M Cichy. A large and rich eeg dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022.

Tijl Grootswagers, Ivy Zhou, Amanda K Robinson, Martin N Hebart, and Thomas A Carlson. Human eeg recordings for 1,854 concepts presented in rapid serial visual presentation streams. *Scientific Data*, 9(1):3, 2022.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26584–26595, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Haonan Jia, Shichao Dong, Xin Dong, Zenghui Sun, Jin Wang, Jinsong Lan, Xiaoyong Zhu, Bo Zheng, and Kaifu Zhang. Cross-modal identity mapping: Minimizing information loss in modality conversion via reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 766–777, 2026.

Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008.

Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, and Quanying Liu. Visual decoding and reconstruction via eeg embeddings with guided diffusion. *Advances in Neural Information Processing Systems*, 37:102822–102864, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Yizhuo Lu, Changde Du, Qiongyi Zhou, Dianpeng Wang, and Huiguang He. Minddiffuser: Controlled image reconstruction from human brain activity with semantic and structural diffusion. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 5899–5908, 2023.

Weijian Mai and Zhijun Zhang. Unibrain: Unify image reconstruction and captioning all in one diffusion model from human brain activity. *arXiv preprint arXiv:2308.07428*, 2023.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

Yoichi Miyawaki, Hajime Uchida, Okito Yamashita, Masa-aki Sato, Yusuke Morito, Hiroki C Tanabe, Norihiro Sadato, and Yukiyasu Kamitani. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5):915–929, 2008.

Weikang Qiu, Zheng Huang, Haoyu Hu, Aosong Feng, Yujun Yan, and Rex Ying. Mindllm: A subject-agnostic and versatile model for fmri-to-text decoding. *arXiv preprint arXiv:2502.15786*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HPuSIXJaa9.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*, 2015.

Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. Deep reinforcement learning-based image captioning with embedding reward. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 290–298, 2017.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7008–7024, 2017.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Paul Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth Norman, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36:24705–24728, 2023.

Paul Steven Scotti, Mihir Tripathy, Cesar Torrico, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A. Norman, and Tanishq Mathew Abraham. Mindeye2: Shared-subject models enable fMRI-to-image with 1 hour of data. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=65XKBGH5PO.

Guobin Shen, Dongcheng Zhao, Xiang He, Linghao Feng, Yiting Dong, Jihang Wang, Qian Zhang, and Yi Zeng. Neuro-vision to language: Enhancing brain recording-based visual reconstruction and language interaction. *Advances in Neural Information Processing Systems*, 37:98083–98110, 2024.

Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLoS computational biology*, 15(1):e1006633, 2019.

Jiahe Shi, Yali Li, and Shengjin Wang. Partial off-policy learning: Balance accuracy and diversity for human-oriented image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2187–2196, 2021.

Yuyang Shi, Valentin De Bortoli, George Deligiannidis, and Arnaud Doucet. Conditional simulation using diffusion schrödinger bridges. In *Uncertainty in Artificial Intelligence*, pp. 1792–1802. PMLR, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14453–14463, 2023.

Zhijiang Tang, Linhua Wang, Jiaxin Qi, Weihao Jiang, Peng Hou, Anxiang Zeng, and Jianqiang Huang. Cccaption: Dual-reward reinforcement learning for complete and correct image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22153–22163, 2026.

Haitao Wu, Qing Li, Changqing Zhang, Zhen He, and Xiaomin Ying. Bridging the vision-brain gap with an uncertainty-aware blur prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2246–2257, 2025.

Weihao Xia and Cengiz Oztireli. Exploring the visual feature space for multimodal neural decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4370–4379, 2025.

Weihao Xia, Raoul de Charette, Cengiz Öztireli, and Jing-Hao Xue. Umbrae: Unified multimodal brain decoding. In *ECCV 2024-European Conference on Computer Vision*, 2024.

Chunyu Ye, Yunhao Zhang, Jingyuan Sun, Chong Li, Yang Zhao, and Shaonan Wang. Decoding the multimodal mind: Generalizable brain-to-text translation via multimodal alignment and adaptive routing. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 22532–22546, 2026.

Li Zhang, Flood Sung, Feng Liu, Tao Xiang, Shaogang Gong, Yongxin Yang, and Timothy M Hospedales. Actor-critic sequence training for image captioning. *arXiv preprint arXiv:1706.09601*, 2017.

Jiawen Zheng, Haonan Jia, MING LI, Yuhui Zheng, Yufeng Zeng, Yang Gao, and Chen Liang. Learning brain representation with hierarchical visual embeddings. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=IEq71qS8B7.