International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 39, No. 15 (2025) 2552024 (14 pages)

DOI: 10.1142/S021800142552024X

Toward a Deep Understanding of Retinal Spike Trains via MLLMs and Diffusion Models

Shanshan Zhao[*,¶], Chenxi Qin[†,‡,‖], Jie Wu[§,**], Pengxiang Li[†,‡,††], Liqun Chen[†,‡,‡‡,‖‖], Wenwei Shao[†,‡,§§,‖‖] and Feng Zheng[*,¶¶,‖‖]

[*]*Department of Computer Science and Engineering*
Southern University of Science and Technology
Shenzhen, Guangdong 518055, P. R. China

[†]*Academy of Medical Engineering and Translational Medicine*
Medical College, Tianjin University, Tianjin 300072, P. R. China

[‡]*Haihe Laboratory of Brain-Computer Interaction and Human-Machine Integration Tianjin 300051, P. R. China*

[§]*Department of Computer Science and Engineering at Southern University of Science and Technology, Sustech P. R. China*

[¶]*zhaoss3@sustech.edu.cn*
[‖]*qin_chenxi@tju.edu.cn*
[**]*tracy_wufz@163.com*
[††]*lpx15290815854@163.com*
[‡‡]*chenliqunlab@tju.edu.cn*
[§§]*wenwei.shao@tju.edu.cn*
[¶¶]*f.zheng@ieee.org*

Received 12 May 2025
Accepted 13 July 2025
Published 22 October 2025

Visual information is first encoded into spike trains by retinal ganglion cells (RGCs), forming the foundation of biological vision. Decoding the semantic and perceptual content embedded in these spikes remains a key challenge in neuroscience. We introduce a novel framework that leverages recent advances in artificial intelligence to interpret spike-based visual signals. Our framework consists of spike-driven multimodal large language model (MLLM) and a spike-driven ControlNet to emulate two core functions of the visual cortex: semantic understanding and image reconstruction. By bridging biological and computational vision, our framework enables effective decoding of spike signals generated by RGCs. We further validate the model on the spike-driven MLLM on the MSCOCO test set. The experiment results demonstrate that our framework not only enhances interpretability and usability of spike-driven outputs but also holds promise as a tool to support neuroscience research and aid visually impaired individuals. Participant feedback highlights its potential to facilitate future advances in neural decoding and vision restoration technologies.

Keywords: Diffusion model; image reconstruction; multimodal large language model; semantic description.

[‖‖] Corresponding authors.

1. Introduction

The retina is pivotal in visual perception, converting light signals into neural signals that are relayed to the brain. Within the retinal structure, retinal ganglion cells (RGCs) are crucial, processing signals from photoreceptors and transmitting them as action potentials, commonly known as retinal spikes train, to the brain[15] (all mentions of "spikes" below refer to retinal spike trains). These spikes are fundamental to the brain's ability to reconstruct visual scenes.[30] The transmission of these spikes brings both visual and partial language information to the brain. Therefore, exploring the mapping relationship between spikes and visual as well as language information is crucial for neuroscientists in analyzing visual information.

A key challenge is deciphering the visual information or meaning contained within spikes. Researchers have dedicated efforts to decode physical world information perceived by organisms from biological signals, with notable focus on extracting language and visual information from brainwaves.[14,25,37] Additionally, given the significance of the visual system, some studies have concentrated on decoding visual information from retinal spikes.[5,29,32] However, these studies often involve image reconstruction within predefined tasks, and substantial differences from biological vision systems remain. With the advancement of general AI technologies, especially language-driven multimodal large language models (MLLMs)[1,27] and diffusion models,[18] increasingly robust capabilities have emerged in text and image generation tasks. Some research indicates a high degree of alignment between the hierarchical structures and functionalities of artificial neural networks and biological vision networks,[15,33,43] suggesting the meaningful pursuit of using various networks to simulate the functions of different brain regions or neural populations. Our work is dedicated to the design of a spike-driven visual cognition system. It integrates MLLMs and diffusion models, aiming to emulate the complex functionalities of biological vision cognitive systems. By decoding intricate information contained within neuronal firing patterns, the system achieves recognition and understanding of visual scenes. This helps visual neuroscientists utilize artificial intelligence as a tool to understand the visual and semantic information within spikes.

To build a spike-driven vision cognition system, we utilized the RGC simulation tool Pranas[7] to generate spike responses to natural images across multiple datasets. Inspired by the contrastive learning paradigm of CLIP,[34] we pretrained a spike encoder on spike-image pairs to align retinal spike signals with natural image features in a shared high-dimensional space. The pretrained spike encoder was then integrated into ControlNet,[45] using spike embeddings as conditional inputs for image reconstruction. To enable semantic interpretation, we repurposed the simulated spike dataset to construct an instruction-tuning corpus and fine-tuned LLaVA[27] using LoRA,[20] improving the model's ability to process spike-formatted inputs. Finally, we combined the fine-tuned large language model and modified diffusion

model that biologically inspired vision cognition framework. Below we list our key contributions:

- We present the spike understanding model that can help us deeply understand the spike signals.
- Our proposed framework contains a spike-driven MLLM and spike-driven ControlNet, capable of decoding spike signals into both semantic descriptions and images.
- Experimental results demonstrate that our framework can assist neuroscientists and professionals in other domains who require efficient analysis and interpretation of spike data.

2. Related Work

Encoding of Visual Scenes by RGCs. Light is transduced by photoreceptors and sequentially processed by horizontal, bipolar, and amacrine cells before RGCs integrate the signal and convey spikes to the brain.[16,22] Early retinal prostheses used direct electrical stimulation[21] and have since advanced clinically,[11,17] yet their benefit is largely confined to degenerative retinal disorders and not to optic-nerve or cortical damage.[3,6] To emulate natural RGC signaling, recent studies pair multi-electrode-array recordings with deep networks, modeling retinal responses to natural scenes in animals and humans.[19,30] Building on this line of work, we draw inspiration from CLIP to align image features and spike trains in a shared embedding space, seeking a high-fidelity model of how the retina encodes natural visual scenes.

Decoding Visual Scenes from Retinal Spikes. Reconstructing natural images from neural activity remains challenging. Existing decoders-spanning fMRI, LGN or retinal spikes, and V1 calcium signals still yield low fidelity natural scene reconstructions.[9,41,44] Diffusion models have boosted fMRI/EEG reconstructions,[36] but their use on spikes is hindered by scarce, spatially inconsistent ex vivo settings and the absence of standard spike-feature pipelines. Simulated RGC datasets indicate that large sample sizes can recover visual content reliably,[24,44] and rodent findings may transfer to primates,[10] underscoring the potential and the data bottleneck of spike-based visual decoding.

Artificial Intelligence Technology. Diffusion generative models now dominate image synthesis and, when conditioned on biological signals, can reconstruct internal visual content. EEG-guided DreamDiffusion produces high-fidelity images directly from brain waves.[4] MindEye and BrainDiVE translate fMRI patterns into viewed images.[25,28] These successes highlight diffusion models' suitability for bio-signal-to-image decoding, though spikes remain unexplored. Contrastive learning is an effective method for cross-modal representation learning, optimizing by maximizing the cosine similarity of positive pairs while minimizing that of negative pairs. Previous work has demonstrated the feasibility and efficacy of contrastive learning

with neural signal data.[12,35] For instance, CLIP[34] exemplifies a multimodal contrastive model by mapping images and text captions into a shared embedding space. Although the precise mapping and encoding mechanisms between visual information and neural spikes remain unclear, leveraging the framework of the CLIP, contrastive learning can efficiently utilize existing data to align spike with image within a high-level space. We use the CLIP loss[34] as our contrastive objective. This loss is bidirectional and helps improve both image and spike retrieval. LLMs have likewise enabled brain-to-text decoding. DeWave maps EEG to sentences via a quantized-vector encoder plus.BART[14] UniCoRN and related work treat fMRI as a "foreign language" decoded by BART or GPT-style models.[8,39,42] To fill the gap for retinal spikes, we curate a spike-instruction dataset and fine-tune LLaVA,[27] demonstrating the first spike-to-semantic pipeline.

3. Method

Our framework comprises two primary components: the Spike-Driven ControlNet and the Spike-Driven MLLM. The Spike-Driven ControlNet serves as an image reconstruction pipeline that extracts visual features from spike signals and reconstructs them into images. The Spike-Driven MLLM is a multimodal large language model that converts spike-encoded information into semantic description.

3.1. *Spike-Driven ControlNet*

In Fig. 1, we can see that the proposed Spike-Driven ControlNet consists of two sequential stages: visual information extraction and image reconstruction. Given a dataset $\mathbb{D}(S, I)$, where $S \in \mathbb{R}^{N\times C\times T}$ denotes the spike data, with N representing the number of samples, C is the number of neuronal channels, and T is the length of the temporal window, $I \in \mathbb{R}^{N\times C\times H\times W}$ represents the associated image data, where H and W denote the spatial dimensions of the image.

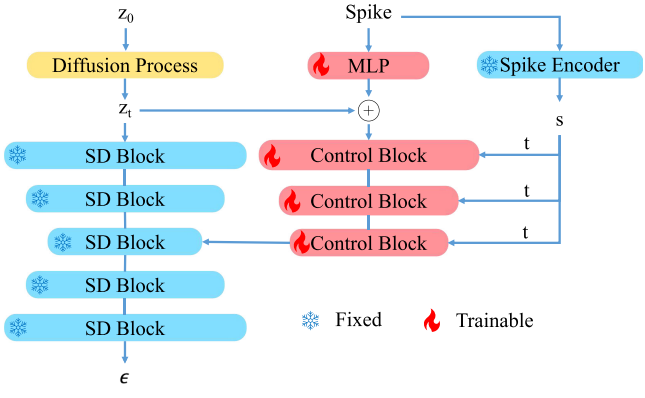


Fig. 1. Spike-Driven ControlNet architecture.

To extract meaningful features from spike sequences, we adopt the pretrained ViT-L/14 model from EVA-CLIP[38] as our image encoder which projects input images into a feature space denoted as $Z_I \in \mathbb{R}^{N\times D}$. For the spike-based feature extraction, we design a four layer MLP to extract features from the spike input, followed by a linear projection layer that maps spike into the shared feature space $Z_S \in \mathbb{R}^{N\times D}$. Furthermore, we leverage the contrastive learning objective of CLIP to align the two modalities.

In the reconstruction stage, we utilize ControlNet as our conditional guided diffusion model. The spike-derived features, aligned to the unified image feature space, are injected into the diffusion model as conditional signals to guide image generation. As shown in Fig. 1, we innovatively designed a spike-sensitive control layer. The control layer receives two inputs: the original spike S and the spike embedding S_x from the Spike Encoder. The diffusion process can be represented as

$$y_c = \mathcal{F}(x;\Theta) + \mathcal{Z}(\mathcal{F}(x, s(x);\Theta_c)) + \mathcal{Z}(c;\Theta_m,\Theta_z), \tag{1}$$

where Θ represents the model parameters of Stable Diffusion, Θ_c denotes the model parameters of ControlNet, Θ_m represents the pretrained parameters of spike encoder, Θ_{z2} is another set of parameters for zero convolution layers, and Θ_z denotes zero convolution.

The final optimization objective of the Spike-Driven ControlNet is

$$\mathcal{L} = \mathbb{E}_{z_0,t,s,\epsilon\sim\mathcal{N}(0,1)}[||\epsilon - \epsilon_\theta(z_t, t, s)||_2^2]. \tag{2}$$

3.2. *Spike-Driven MLLM*

Multimodal large language models (MLLMs) are trained by the image-language pairs that cannot interpret neural spike data. To bridge this gap, we propose a spike-driven MLLM based on LLaVA that can accurately respond to spike-based queries in Fig. 2.

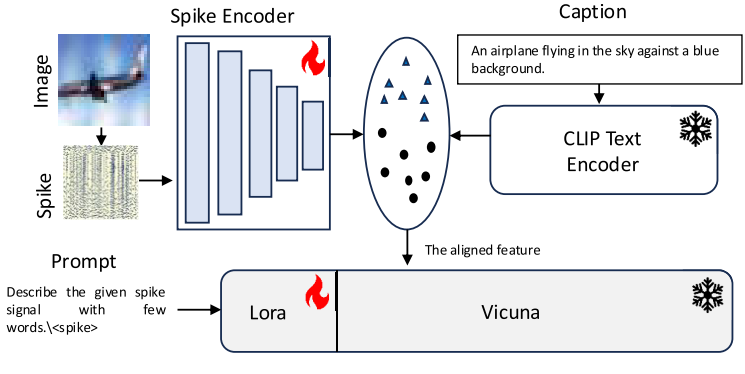


Fig. 2. Spike-Driven MLLM architecture.

3.2.1. *Spike encoder*

We denote a training sample as $\langle \text{Image } I, \text{Spike } S, \text{Prompt } P, \text{Caption } C \rangle$, where each image I is paired with a retinal spike train S and a natural-language caption C. The spike signal is first preprocessed by temporal binning: for a recording window of length T we count spikes in m equal bins, resulting in an input vector $\mathbf{s}_0 \in \mathbb{R}^{d_{\text{in}}}$ with $d_{\text{in}} = m$.

The *Spike Encoder* is a multilayer perceptron that progressively projects $\mathbf{s}_0$ onto a low-dimensional semantic space. Let hidden dimensions be $\{h_1, h_2, \ldots, h_{n-1}\}$ and output dimension d_{out}. For layer $i (i = 1, \ldots, n-1)$ we represented as

$$\mathbf{h}_i = \sigma(\text{BN}(\mathbf{W}_i \mathbf{h}_{i-1} + \mathbf{b}_i)), \quad \mathbf{h}_0 = \mathbf{s}_0, \tag{3}$$

where $\mathbf{W}_i \in \mathbb{R}^{h_i \times h_{i-1}}$, $\mathbf{b}_i \in \mathbb{R}^{h_i}$, $\text{BN}(\cdot)$ is batch normalization and $\sigma(\cdot)$ is ReLU. A dropout layer (rate $p = 0.2$) follows each activation to mitigate over-fitting. The final layer omits the nonlinearity:

$$\mathbf{H}_s = \mathbf{W}_n \mathbf{h}_{n-1} + \mathbf{b}_n, \quad \mathbf{H}_s \in \mathbb{R}^{d_{\text{out}}}. \tag{4}$$

This gradual dimension reduction allows early layers to capture fine-grained temporal patterns while encouraging deeper layers to learn high-level semantics.

To align spike representations with language space, we encode the paired caption C using the *frozen* CLIP text encoder $T_\theta(\cdot)$:

$$\mathbf{H}_c = T_\theta(C), \quad \theta \text{ is fixed.} \tag{5}$$

During training, we minimize a temperature-scaled InfoNCE loss:

$$\mathcal{L}_{\text{contrastive}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathbf{H}_s^{(i)}, \mathbf{H}_c^{(i)})/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\mathbf{H}_s^{(i)}, \mathbf{H}_c^{(j)})/\tau)}, \tag{6}$$

where $\text{sim}(\cdot, \cdot)$ is cosine similarity, τ is a learnable temperature, and N is the batch size. The objective maximises similarity for matched spike-caption pairs while minimising it for mismatched pairs, thereby forcing $\mathbf{H}_s$ to reside in the same semantic manifold as CLIP language embeddings.

3.2.2. *Spike decoder*

Given a spike embedding $\mathbf{H}_s$ from the Spike Encoder and an instruction prompt X_P (e.g. "*Could you describe the spike signal?*"), the Spike Decoder autoregressively generates a natural-language sequence $Y^* = (y_1^*, \ldots, y_L^*)$ with length L measured in BPE tokens. We adopt Vicuna-13B as the backbone LLM and attach a LoRA adapter to all attention and feed-forward layers. The conditional likelihood is computed as

$$p_\theta(Y^* | X_S, X_P) = \prod_{i=1}^{L} p_\theta(y_i^* | \mathbf{H}_s, X_P, Y_{<i}^*), \tag{7}$$

where θ denotes trainable LoRA parameters, $\mathbf{H}_s$ encodes the spike train X_S, X_P is the instruction prompt and $Y_{<i}^*$ are previously generated tokens.

4. Experiments

To evaluate the effectiveness of our proposed method, we conducted comprehensive experiments on the test set.

4.1. *Datasets*

The MSCOCO dataset is a widely adopted benchmark in the vision community, extensively utilized for tasks such as object detection, image captioning, and instance segmentation. It includes 82 783 images for training and 40 504 for validation in its original configuration. For consistency and fair comparison with previous work, we adopt the widely-used Karpathy split.[23] This split reorganizes the dataset into 113 287 images for training, 5000 for validation, and 5000 for testing, with each image associated with five human-annotated captions. All evaluations in this study are conducted on the test set defined by the Karpathy split. We train Spike-Driven ControlNet on the CelebA dataset, which contains annotated face images, and evaluate its zero-shot classification performance on CIFAR-10.

We further evaluated our approach on CIFAR-10. Specifically, we sampled 5000 images to form a dedicated test set and used Pranas[7] to sequentially generate spike signals for each image. In parallel, we employed GPT-4o mini[1] to produce one caption per test image, which we treat as the ground-truth reference for subsequent evaluation.

For these datasets, we construct a set of instructions and generate corresponding spike signals for each image using Pranas. These instructions are used to prompt the model to decode the spike signals into image captions. The specific instruction formats are summarized in Table 1.

4.2. *Evaluation metrics*

To comprehensively evaluate the quality of the generated captions, we adopt a suite of widely used metrics from the MSCOCO captioning benchmark, including BLEU,[31] CIDEr,[40] SPICE,[2] METEOR,[13] and ROUGE.[26] These metrics assess different

Table 1. Instruction prompts used for spike signal captioning on MSCOCO and CIFAR10.

ID	Instructions
1	Could you describe the spike signal?
2	Can you provide me with a description of the spike signal?
3	Write a description for the spike signal
4	Describe the content of the spike signal
5	Provide a caption for the spike signal
6	Please provide a short description of the spike signal
7	Provide a description of the spike signal
8	Provide a description of what is presented in the spike signal
9	What is the spike signal like? Could you give me a description?
10	Use a few words to illustrate what is happening in the spike signal

aspects of caption quality from both syntactic and semantic perspectives. We evaluate the reconstruction performance using three standard metrics: PSNR, SSIM, and LPIPS. PSNR and SSIM assess pixel-level fidelity and structural similarity, respectively, with higher values indicating better quality. LPIPS measures perceptual similarity based on deep features, where lower scores denote closer alignment with human perception. Together, these metrics provide a balanced evaluation of both objective accuracy and perceptual quality. We employed retrieval metrics to measure the performance of the spike encoder. specifically showing the Recall performance at different Top-K values, including R@1, R@5, and R@10.

4.3. *Experiment result of the Spike-Driven ControlNet*

To assess the transferability of our Spike-Driven ControlNet, we evaluate its performance in two downstream tasks: cross-modal retrieval and zero-shot classification, as summarized in Table 2. The retrieval results on the Celeb dataset achieve 3.3% at R@1, 10.00% at R@5, and 15.97% at R@10, indicating that the encoder effectively captures discriminative features for spike-image alignment. In addition, the zero-shot classification results on CIFAR-10 show a top-1 accuracy of 10.01%, further demonstrating the generalization capability of the spike-derived features without task-specific fine-tuning. Figure 3 shows the reconstruction result, the first line is the groundtruth, the second line is the reconstruction result of our Spike-Driven ControlNet, we can see that the reconstructed result is similar with the groundtruth. In Fig. 4, the first column shows the ground truth, each image correspond to a spike. After we encode the spike by the proposed spike encoder and

Table 2. Retrieval and zero-shot classification results.

Retrieval			
Datasets	R@1	R@5	R@10
Celeb	3.3	10.00	15.97
Zero-shot classification			
Cifar10	10.01	22.90	29.93

Fig. 3. The image reconstruction result of the Spike-Driven ControlNet.

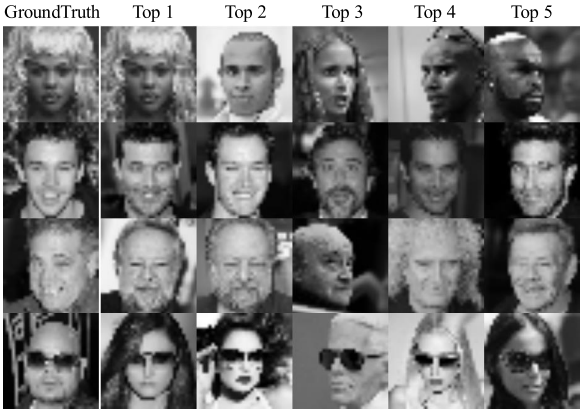


Fig. 4. The top-5 retrieval results according to the spike encoded features.

perform top-5 retrieval on the test set, the retrieval result shows that the extracted result are similar with the ground truth. These results validate the feasibility of applying our method to a wide range of real-world scenarios, highlighting the encoder's potential in both representation learning and practical applications.

4.4. *Experiment result of the Spike-Driven MLLM*

The experiment results of MSCOCO and CIFAR10 are shown in Table 3. Since our model is fine-tuned on the MSCOCO training set, the evaluation on the MSCOCO test set reflects performance under supervised fine-tuning (SFT). In contrast, our model has never been trained on CIFAR10, making the evaluation on the CIFAR10 test set a zero-shot scenario. This experimental design allows us to assess the generalization ability of the model across different datasets.

As shown in the results, our model achieves a CIDEr score of 5.1 on the CIFAR10 test set in a zero-shot setting, indicating that it can understand and generate semantically meaningful captions even for previously unseen spike inputs. Under SFT on MSCOCO, the model achieves a CIDEr score of 15.4, demonstrating its ability to accurately translate spike signals into textual descriptions when trained on in-domain data. The performance gap of 10.3 CIDEr between MSCOCO and CIFAR10 highlights the effectiveness of SFT in enhancing the model's captioning performance on familiar distributions.

Table 3. Experiment results tested on different test sets.

Dataset	BLEU-1	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
			Zero-shot			
CIFAR10	5.9	0.6	4.4	9.9	5.1	3.0
			Supervised fine-tuning			
MSCOCO	**14.6**	**1.1**	**6.3**	**13.7**	**15.4**	**7.1**

Table 4. Ablation study on input conditions for Spike-Driven ControlNet.

Spike	Spike Embedding	SSIM↑	PSNR↑	LPIPS↓
√	×	0.386 ± 0.0726	$\mathbf{13.17 \pm 1.62}$	0.393 ± 0.0559
×	√	0.373 ± 0.0836	11.17 ± 1.91	0.392 ± 0.0634
√	√	$\mathbf{0.454 \pm 0.0868}$	13.14 ± 2.0919	$\mathbf{0.337 \pm 0.0604}$

4.5. *Ablation study*

To determine the most effective conditioning strategy, we conduct a comparative study of different injection schemes. Specifically, we evaluate two primary configurations: (1) using only the spike embedding (i.e. the output of the Spike Encoder) as the conditional input, and (2) jointly using both the raw spike signal and its encoded embedding. We hypothesize that the latter configuration, which preserves both low-level and high-level spike information, would result in superior reconstruction quality.

To validate this hypothesis, we perform extensive ablation studies in Table 4. When using only the raw spike or the spike embedding as the conditional input, the reconstruction quality remains limited across all metrics. In contrast, the combination of both the raw spike and its encoded embedding yields the best performance in terms of structural similarity (SSIM) and perceptual quality (LPIPS). This indicates that preserving both low-level and high-level spike information is beneficial for enhancing reconstruction accuracy. Notably, the configuration using only raw spikes achieves the highest PSNR. However, it underperforms in SSIM and LPIPS compared to the joint setting. This suggests that relying solely on low-level features does not necessarily guarantee better overall reconstruction quality, particularly in terms of structural and perceptual fidelity.

5. Conclusion

We propose an overall framework designed to deeply understand retinal spike through advanced artificial intelligence methods. Our framework unites multimoal large language models with diffusion-based generation to translate retinal ganglion-cell spike trains into both semantic description and images. By exploiting CLIP's multimodal alignment, spike embeddings reside in a shared semantic-visual space, letting a model trained only on CelebA achieve robust zero-shot classification on CIFAR-10. A ControlNet-conditioned Stable Diffusion module reconstructs images from spikes, while LoRA fine-tuning equips LLaVA with spike comprehension. Our framework offers neuroscientists a fresh avenue for decoding visual information, supplies a theoretical basis for fully spike-driven visual prostheses.

Acknowledgments

S. Zhao and C. Qin have contributed equally to this work. This work was supported by the National Key Research and Development Program of China (Grant No. 2022YFF1202903).

ORCID

Shanshan Zhao https://orcid.org/0009-0004-9888-9983
Feng Zheng https://orcid.org/0000-0002-1701-9141

References

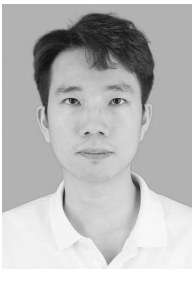
1. J. Achiam *et al.*, Gpt-4 technical report, preprint (2023), arXiv:2303.08774.
2. P. Anderson, B. Fernando, M. Johnson and S. Gould, Spice: Semantic propositional image caption evaluation, in *Computer Vision–ECCV 2016: 14th European Conf. Amsterdam, The Netherlands, 11–14 October 2016, Proceedings, Part V 14* (Springer, 2016), pp. 382–398.
3. L. N. Ayton *et al.*, An update on retinal prostheses, *Clin. Neurophysiol.* **131**(6) (2020) 1383–1398.
4. Y. Bai, X. Wang, Y.-P. Cao, Y. Ge, C. Yuan and Y. Shan, Dreamdiffusion: Generating high-quality images from brain EEG signals, preprint (2023), arXiv:2306.16934.
5. T. Benster, D. Babino, J. Thickstun, M. Hunt, X. Liu, Z. Harchaoui, S. Oh and R. N. Van Gelder, Reconstruction of visual images from mouse retinal ganglion cell spiking activity using convolutional neural networks, bioRxiv (2022), https://doi.org/10.1101/2022.06.10.482188.
6. E. Bloch, Y. Luo and L. da Cruz, Advances in retinal prosthesis systems, *Ther. Adv. Ophthalmol.* **11** (2019) 2515841418817501.
7. B. Cessac, P. Kornprobst, S. Kraria, H. Nasser, D. Pamplona, G. Portelli and T. Viéville, Pranas: A new platform for retinal analysis and simulation, *Front. Neuroinform.* **11** (2017) 49.
8. X. Chen, C. Du, C. Liu, Y. Wang and H. He, Open-vocabulary auditory neural decoding using FMRI-prompted LLM, preprint (2024), arXiv:2405.07840.
9. Z. Chen, J. Qing and J. H. Zhou, Cinematic mindscapes: High-quality video reconstruction from brain activity, *Adv. Neural Inf. Process. Syst.* **36** (2024) 24841–2485.
10. A. Corna, P. Ramesh, F. Jetter, M.-J. Lee, J. H. Macke and G. Zeck, Discrimination of simple objects decoded from the output of retinal ganglion cells upon sinusoidal electrical stimulation, *J. Neural Eng.* **18**(4) (2021) 046086.
11. L. da Cruz *et al.*, Five-year safety and performance results from the argus ii retinal prosthesis system clinical trial, *Ophthalmology* **123**(10) (2016) 2248–2254.
12. A. Défossez, C. Caucheteux, J. Rapin, O. Kabeli and J.-R. King, Decoding speech from non-invasive brain recordings, https://arxiv.org/abs/2208.12266 (2022).
13. M. Denkowski and A. Lavie, Meteor universal: Language specific translation evaluation for any target language, in *Proc. Ninth Workshop on Statistical Machine Translation* (ACL, 2014), pp. 376–380.
14. Y. Duan, C. Chau, Z. Wang, Y.-K. Wang and C.-T. Lin, Dewave: Discrete encoding of EEG waves for EEG to text translation, *Adv. Neural Inf. Process. Syst.* **36** (2024) 9907–9918.
15. L. Dyballa, A. Rudzite, M. S. Hoseini, M. Thapa, M. P. Stryker, G. D. Field and S. W. Zucker, Population encoding of stimulus features along the visual hierarchy, *Proc. Natl. Acad. Sci.* **121**(4) (2024) e2317773121.
16. T. Euler, S. Haverkamp, T. Schubert and T. Baden, Retinal bipolar cells: Elementary building blocks of vision, *Nat. Rev. Neurosci.* **15**(8) (2014) 507–519.
17. A. C. Ho *et al.*, Long-term results from an epiretinal prosthesis to restore sight to the blind, *Ophthalmology* **122**(8) (2012) 1547–1554.

18. J. Ho, A. Jain and P. Abbeel, Denoising diffusion probabilistic models, in *Proc. 34th Int. Conf. Neural Information Processing Systems*, eds. H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan and H. Lin, Vol. 33 (Curran Associates, Inc., 2020), pp. 6840–6851.
19. B. D. Hoshal *et al.*, Stimulus invariant aspects of the retinal code drive discriminability of natural scenes, bioRxiv **121**(52) (2023).
20. E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang and W. Chen, Lora: Low-rank adaptation of large language models, preprint (2021), arXiv:2106.09685.
21. M. S. Humayun *et al.*, Interim results from the international trial of second sight's visual prosthesis, *Ophthalmology* **119**(4) (2012) 779–788.
22. D. Karamanlis, H. M. Schreyer and T. Gollisch, Retinal encoding of natural scenes, *Annu. Rev. Vision Sci.* **8**(1) (2022) 171–193.
23. A. Karpathy and L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (IEEE, 2015) pp. 3128–3137.
24. W. Li, A. N. J. Raj, T. Tjahjadi and Z. Zhuang, Fusion of anns as decoder of retinal spike trains for scene reconstruction, *Appl. Intell.* **52**(13) (2022) 15164–15176.
25. D. Li, C. Wei, S. Li, J. Zou and Q. Liu, Visual decoding and reconstruction via eeg embeddings with guided diffusion, preprint (2024), arXiv:2403.07721.
26. C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in *Text Summarization Branches Out* (ACL, 2004), pp. 74–81.
27. H. Liu, C. Li, Q. Wu and Y. J. Lee, Visual instruction tuning, in *Proc. 37th Int. Conf. Neural Information Processing Systems* (Curran Associates Inc., Red Hook, USA, 2023), pp. 34892–34916.
28. A. Luo, M. Henderson, L. Wehbe and M. Tarr, Brain diffusion for visual exploration: Cortical discovery using large scale generative models, *Adv. Neural Inf. Process. Syst.* **36** (2024) 75740–75781.
29. K. Ma, A. N. J. Raj, V. Rajangam, T. Tjahjadi, M. Liu and Z. Zhuang, Retinal spike train decoder using vector quantization for visual scene reconstruction, *Complex Intell. Syst.* **10** (2024) 3445–3458.
30. N. Maheswaranathan *et al.*, Interpreting the retinal neural code for natural scenes: From computations to neurons, *Neuron* **111**(17) (2023) 2742–2755.
31. K. Papineni, S. Roukos, T. Ward and W.-J. Zhu, Bleu: A method for automatic evaluation of machine translation, in *Proc. 40th Annual Meeting of the Association for Computational Linguistics* (ACL, 2002), pp. 311–318.
32. N. Parthasarathy, E. Batty, W. Falcon, T. Rutten, M. Rajpal, E. J. Chichilnisky and L. Paninski, Neural networks for efficient bayesian decoding of natural images from retinal neurons, *Adv. Neural Inf. Process. Syst.* **30** (2017).
33. J. S. Prince, G. Fajardo, G. A. Alvarez and T. Konkle, Manipulating dropout reveals an optimal balance of efficiency and robustness in biological and machine visual systems, in *Proc. Twelfth Int. Conf. Learning Representations*, Vienna, Austria, 7–11 May 2024.
34. A. Radford *et al.*, Learning transferable visual models from natural language supervision, in *Proc. Int. Conf. Machine Learning* (PMLR, 2021), pp. 8748–8763.
35. S. Schneider, J. H. Lee and M. W. Mathis, Learnable latent embeddings for joint behavioural and neural analysis, *Nature* **617**(7960) (2023) 360–368.
36. P. Scotti *et al.*, Reconstructing the mind's eye: FMRI-to-image with contrastive learning and diffusion priors, *Adv. Neural Inf. Process. Syst.* **36** (2024) 24705–24728.
37. P. Singh, D. Dalal, G. Vashishtha, K. Miyapuram and S. Raman, Learning robust deep visual representations from EEG brain recordings, in *Proc. IEEE/CVF Winter Conf. Applications of Computer Vision (WACV)* (IEEE, 2024), pp. 7553–7562.

38. Q. Sun, Y. Fang, L. Wu, X. Wang and Y. Cao, Eva-clip: Improved training techniques for clip at scale, preprint (2023), arXiv:2303.15389.
39. J. Tang, A. LeBel, S. Jain and A. G. Huth, Semantic reconstruction of continuous language from non-invasive brain recordings, *Nat. Neurosci.* **26**(5) (2023) 858–866.
40. R. Vedantam, C. L. Zitnick and D. Parikh, Cider: Consensus-based image description evaluation, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (IEEE, 2015), pp. 4566–4575.
41. C. Wang, H. Yan, W. Huang, J. Li, Y. Wang, Y.-S. Fan, W. Sheng, T. Liu, R. Li and H. Chen, Reconstructing rapid natural vision with fmri-conditional video generative adversarial network, *Cereb. Cortex* **32**(20) (2022) 4502–4511.
42. N. Xi, S. Zhao, H. Wang, C. Liu, B. Qin and T. Liu, Unicorn: Unified cognitive signal reconstruction bridging cognitive signals and human language, in *Annual Meeting of the Association for Computational Linguistics*, Toronto, Canada, 2023, pp. 13277–13291.
43. H. Yang, J. Gee and J. Shi, Brain decodes deep nets, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* (IEEE, 2024), pp. 23030–23040.
44. Z. Yu, T. Bu, Y. Zhang, S. Jia, T. Huang and J. K. Liu, Robust decoding of rich dynamical visual scenes with retinal spikes, *IEEE Trans. Neural Netw. Learn. Syst.* **36**(2) (2024) 3396–3409.
45. L. Zhang, A. Rao and M. Agrawala, Adding conditional control to text-to-image diffusion models, in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)* (IEEE, 2023), pp. 3836–3847.

Shanshan Zhao is a postdoctoral researcher in the Department of Computer Science and Engineering at the Southern University of Science and Technology. Her main research directions include image understanding, video understanding, and multimodal large models.

Chenxi Qin is a Master's student with the School of Medicine, Tianjin University, China. His research interest focuses on organ-on-a-chip brain-machine interfaces.



Jie Wu is a research assistant in the Department of Computer Science and Engineering at Southern University of Science and Technology (Sustech). His research interests include vision-language models and AI4Science.

Pengxiang Li, Ph.D. candidate at the Tianjin University Academy of Medical Engineering. Main research directions: artificial intelligence, brain-computer interfaces.

Liqun Chen, Ph.D. Supervisor, Professor, Senior Engineer, Chief Scientist of the National Key R&D Program, Vice Dean of the Tianjin University School of Clinical Medicine, Selected for the National Postdoctoral Innovative Talents Support Program, Selected for the China Association for Science and Technology Leadership Program, Tianjin "Project + Team" Innovative Talent, and Tianjin Outstanding Young Scientific and Technological Worker. Main research directions include: cross-disciplinary research on on-chip visual brain-computer interfaces and neuroregulatory information interaction; functionalization, intelligence, and safety assessment of brain organoids; environmental health research based on organoid and brain-computer interface technologies; molecular mechanisms of stress factor regulation in neuro-metabolism-related diseases in precision medicine.

Wenwei Shao is an Associate Professor with the School of Medicine, Tianjin University, China. His research interests focus on organ-on-a-chip brain-machine interfaces, brain organoids, and gene therapy.

Feng Zheng is Associate Professor with the Department of Computer Science and Engineering, Southern University of Science and Technology (SUSTech),China. His research interests include machine learning, computer vision, and human-computer interaction.