Appendix for SeeEEG: Semantic-aware EEG-based Multi-Modal Retrieval-Augmented Generation for High-Fidelity Visual Brain Decoding

Jun-Mo Kim* Woohyeok Choi* Sang-Jun Park Keun-Soo Heo
Young-Han Son Ji-Hye Oh Dong-Hee Shin Tae-Eui Kam†
Korea University, Seoul, Korea
{wnsah1008, woohyeok_choi, wedm2401, gjrmstn1440,
yhson135, meeeo_, dongheeshin, kamte}@korea.ac.kr

A. Comparison of EEG Encoding Methods

In this study, we compare the performance of the Semantic Region-aware Transformer (SRT) with various EEG-based models. The competing models are widely used in EEG signal analysis and encompass diverse neural network-based approaches.

ShallowNet [9] and **DeepConvNet** [9] use convolutional architectures designed specifically for EEG tasks. ShallowNet captures basic spectral EEG features using simple convolutional layers, while DeepConvNet employs multiple convolutional and pooling layers to extract hierarchical temporal and spatial features from EEG data.

EEGNet [6] employs a compact convolutional neural network architecture with depthwise and separable convolutions, effectively capturing frequency-specific spatial representations and achieving generalizability across multiple EEG paradigms with fewer parameters.

ATCNet [1] integrates temporal convolutional layers with multi-head self-attention modules, effectively encoding low-level spatial-temporal features and subsequently highlighting essential temporal segments through attention mechanisms.

EEGConformer [11] combines convolutional modules with Transformer-based self-attention layers, effectively integrating local and global temporal features to improve the classification performance by capturing extensive temporal dependencies.

TSCeption [4] focuses on EEG-based emotion recognition, leveraging multi-scale temporal kernels and asymmetric spatial convolutional layers to effectively capture both temporal dynamics and the asymmetric spatial activations relevant to emotional processes.

NICE-EEG [12] utilizes a Temporal-Spatial Convolution (TSConv) architecture as its EEG encoder. The TSConv applies sequential temporal and spatial convolution layers to extract EEG representations, allowing the model to effectively capture both local temporal features and cross-channel spatial dependencies.

ATM-S [7] utilizes a Channel-wise Transformer encoder along with Temporal-Spatial convolution to model both spatial and temporal dependencies in EEG signals. The Channel-wise Transformer captures inter-channel relationships, while the Temporal-Spatial convolution enhances feature extraction by preserving EEG's temporal structures. This architecture is designed to improve EEG-based classification and decoding tasks.

B. Topological Analysis on EEG Encoder

To validate the neural plausibility of our EEG embeddings, we analyzed topographic maps of EEG regional representations across all subjects from both the EEG-to-Image and EEG-to-Text encoders. Specifically, the EEG-to-Image encoder initially shows posterior and occipital activation, which progressively shifts toward central and frontal regions, reflecting the hierarchical neural dynamics from early visual feature extraction to object recognition and perceptual decision-making [2, 3, 5]. In contrast, the EEG-to-Text encoder exhibits anterior-temporal and fronto-temporal activations, consistent with established neural signatures of semantic retrieval and language comprehension [8, 10]. This clear modality-specific differentiation offers empirical neuroscientific validation of our Semantic Region-aware Transformer (SRT) Encoder, underlining its effectiveness in extracting semantically meaningful EEG embeddings for accurate multimodal decoding.

C. Extended Evaluation of EEG-Based Image Generation and Limitation

We conducted additional comparative experiment with CognitionCapturer [13], a recent state-of-the-art framework for EEG-based visual brain decoding which reconstructs visual stimuli by leveraging three modalities: image, text, and depth. For a fair comparison, we averaged the performance across

*Equal contribution
†Corresponding author

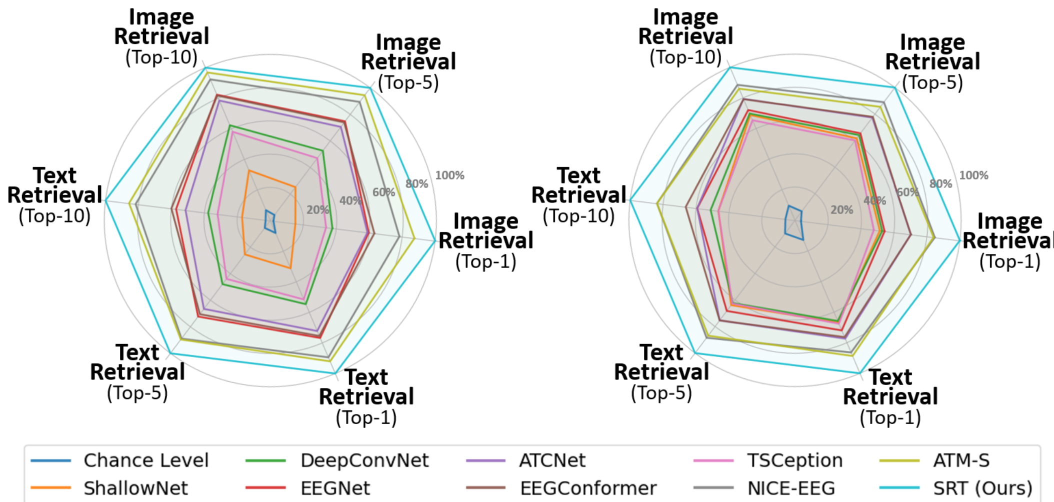

Figure A1. EEG-based retrieval performance comparison with competing methods. Left: Subject-dependent setting. Right: Subject-independent setting.

	Subject Dependent						Subject Independent					
	Image			Text			Image			Text		
Models	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
ShallowNet [9]	4.0	15.0	24.0	3.1	11.5	19.2	7.1	22.9	36.0	4.0	16.2	26.0
DeepConvNet [9]	9.9	31.3	45.5	6.6	21.7	33.6	7.3	23.8	36.5	4.4	15.7	25.6
EEGNet [6]	15.7	44.5	60.1	10.0	32.6	47.2	7.5	24.4	37.7	5.1	17.2	28.0
ATCNet [1]	15.5	42.0	57.3	9.0	30.0	44.4	9.7	28.7	41.5	5.1	19.0	30.1
EEGConformer [11]	16.6	44.1	59.7	10.5	31.8	46.4	9.7	28.9	41.4	5.7	19.1	29.7
TSCeption [4]	8.9	28.0	42.4	5.6	19.8	31.7	6.6	22.4	34.2	4.0	15.8	26.3
NICE-EEG [12]	20.6	53.2	67.4	14.3	40.1	55.0	11.7	33.0	46.3	7.2	22.3	33.6
ATM-S [7]	23.0	56.2	70.7	15.0	40.4	56.6	11.6	31.7	45.0	7.3	22.0	34.5
SRT (Ours)	**26.3**	**59.5**	**73.0**	**17.5**	**45.1**	**61.5**	**13.8**	**37.1**	**52.3**	**8.6**	**25.3**	**38.9**

Table A1. EEG-based image and text retrieval performance comparison with competing methods.

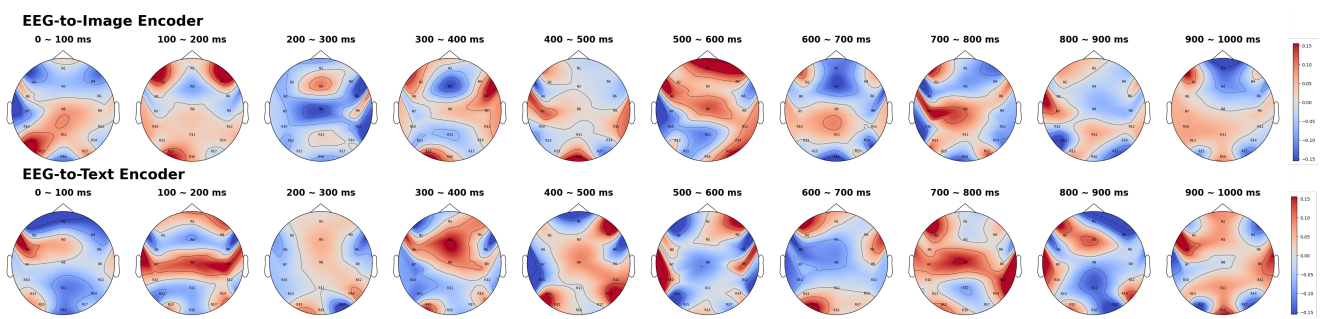

Figure A2. Visualization of topographical maps from the EEG-to-Image encoder (top row) and EEG-to-Text encoder (bottom row), displayed across successive 100 ms time windows (0–1000 ms). Red and blue regions denote relatively higher and lower activation, respectively, underscoring the distinct spatiotemporal activation patterns associated with each encoder.

all subjects. As shown in Table A2, the experimental results not only demonstrate that our proposed method outperforms competing method on most evaluation metrics, but also show the generalizability of our proposed method.

Our SeeEEG substantially outperformed the state-of-the-art methods [7, 13] on high-level semantic metrics (AlexNet, Inception, CLIP, and SwaV), while exhibiting relatively lower performance on low-level metrics (PixCorr and SSIM),

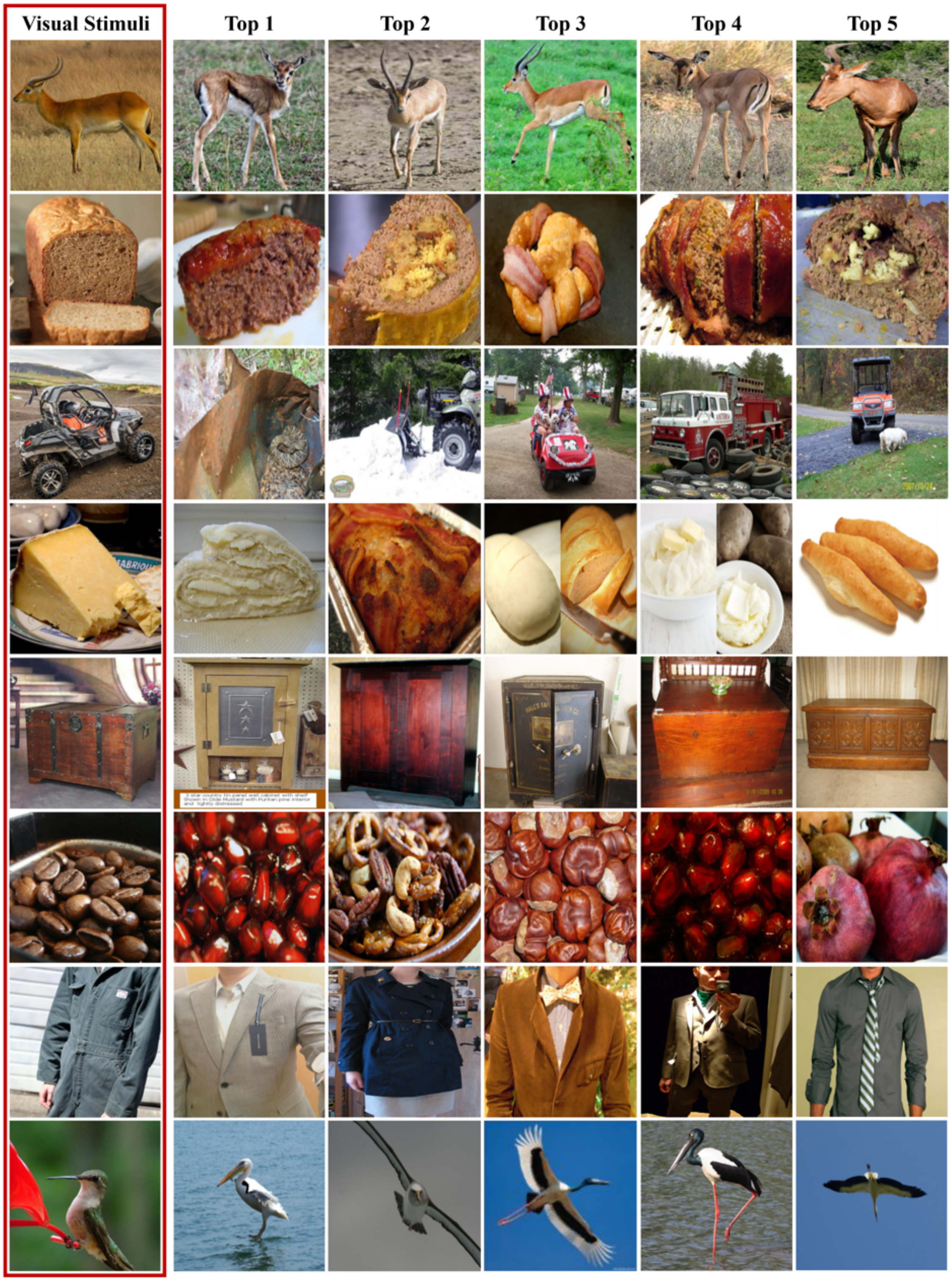

Figure A3. Example retrieval results from ILSVRC dataset using image-aligned EEG embeddings. Each row begins with the visual stimulus, followed by the top five most similar images retrieved from the the dataset.

as shown in Tables 3 and A2. This discrepancy can be attributed to the fact that image retrieval from the external

Table A2. The quantitative comparison of generation performance with the state-of-the-art method. All results are averaged across all subjects.

Methods	PixCorr↑	SSIM↑	AlexNet(2)↑	AlexNet(5)↑	Inception↑	CLIP↑	SwAV↓
[13]	**0.15**	**0.347**	0.754	0.623	0.669	0.715	0.590
Ours	0.11	0.335	**0.782**	**0.861**	**0.724**	**0.795**	**0.573**

reference database is primarily driven by semantic similarity, which does not necessarily ensure pixel-level or structural resemblance between the ground truth and the generated images. To mitigate this limitation, we plan to develop a retrieval algorithm that incorporates structural similarity and to explore brain encoding methods that more effectively leverage low-level visual information encoded in the brain.

References

[1] Hamdi Altaheri, Ghulam Muhammad, and Mansour Alsulaiman. Physics-informed attention temporal convolutional network for eeg-based motor imagery classification. *IEEE transactions on industrial informatics*, 19(2):2249–2258, 2022. 1, 2

[2] Thomas Carlson, David A Tovar, Arjen Alink, and Nikolaus Kriegeskorte. Representational dynamics of object vision: the first 1000 ms. *Journal of vision*, 13(10):1–1, 2013. 1

[3] Radoslaw Martin Cichy, Dimitrios Pantazis, and Aude Oliva. Resolving human object recognition in space and time. *Nature neuroscience*, 17(3):455–462, 2014. 1

[4] Yi Ding, Neethu Robinson, Su Zhang, Qiuhao Zeng, and Cuntai Guan. Tsception: Capturing temporal dynamics and spatial asymmetry from eeg for emotion recognition. *IEEE Transactions on Affective Computing*, 14(3):2238–2250, 2022. 1, 2

[5] Tijl Grootswagers, Susan G Wardle, and Thomas A Carlson. Decoding dynamic brain patterns from evoked responses: A tutorial on multivariate pattern analysis applied to time series neuroimaging data. *Journal of cognitive neuroscience*, 29(4): 677–697, 2017. 1

[6] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5): 056013, 2018. 1, 2

[7] Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, and Quanying Liu. Visual decoding and reconstruction via EEG embeddings with guided diffusion. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1, 2

[8] Simone Palazzo, Concetto Spampinato, Isaak Kavasidis, Daniela Giordano, Joseph Schmidt, and Mubarak Shah. Decoding brain representations by multimodal learning of neural activity and visual features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3833–3849, 2020. 1

[9] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38(11):5391–5420, 2017. 1, 2

[10] Irina Simanova, Marcel Van Gerven, Robert Oostenveld, and Peter Hagoort. Identifying object categories from event-related eeg: toward decoding of conceptual representations. *PloS one*, 5(12):e14465, 2010. 1

[11] Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2022. 1, 2

[12] Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from EEG for object recognition. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2

[13] Kaifan Zhang, Lihuo He, Xin Jiang, Wen Lu, Di Wang, and Xinbo Gao. Cognitioncapturer: Decoding visual stimuli from human eeg signal with multimodal information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14486–14493, 2025. 1, 2, 4