Dual-Stage Self-Attention for Contrastive EEG-Image Alignment

Lei Kang, Xuanshuo Fu, Xim Cerda-Company, Ernest Valveny, and Dimosthenis Karatzas

Computer Vision Center, Barcelona, Spain
Universitat Autònoma de Barcelona, Barcelona, Spain
{lkang, xuanshuo, xcerda, ernest, dimos}@cvc.uab.es

Abstract. Learning semantically organized representations from electroencephalography (EEG) is an important prerequisite for non-invasive brain-computer interfaces and other healthcare-oriented neural-signal applications, yet visual decoding from EEG remains difficult because scalp recordings are noisy, high-dimensional, and data-limited. Recent EEG-vision methods increasingly rely on pretrained visual spaces, but often leave open how EEG structure across electrodes and time should be modeled before cross-modal alignment. We propose a dual-stage self-attention framework for contrastive EEG-image alignment in which channel-wise self-attention first models inter-electrode dependencies while preserving temporal resolution, and temporal self-attention then aggregates evidence over time through a learnable `[CLS]` token to form a compact EEG embedding. The EEG embedding is aligned to a frozen pretrained SigLIP image space through a label-aware contrastive objective that treats same-class samples as positives. Under the EEGCVPR40 benchmark protocol used in this work, the proposed method outperforms CNN and generic transformer baselines with comparable parameter counts, and ablations show that avoiding premature channel aggregation is a key design choice. We interpret these results as benchmark-level evidence that delayed aggregation benefits EEG representation learning, while broader validation on randomized and clinically grounded datasets remains future work.

Keywords: EEG visual decoding · contrastive learning · multimodal alignment · transformer · brain-computer interface

1 Introduction

Electroencephalography (EEG) offers a compelling non-invasive interface for decoding visual information because it is relatively inexpensive, portable, and temporally precise. These properties make EEG especially attractive for practical brain-computer interfaces and for studying the fast dynamics of human visual recognition. At the same time, visual decoding from EEG remains challenging: scalp signals are noisy, spatially mixed, and highly variable across electrodes

and time, making it difficult to recover semantic information robustly from raw measurements [13, 17].

Recent studies nevertheless suggest that EEG does contain meaningful object-level information and that the use of pretrained visual feature spaces can improve the structure of learned neural representations [5]. Earlier work demonstrated the feasibility of visual object decoding from EEG and established public EEG-image benchmarks [17]. Subsequent multimodal studies showed that relating neural activity to visual representations can produce richer joint embeddings and more transferable models [11, 7]. More recent work has further advanced EEG-based visual decoding through knowledge distillation, contrastive learning, language guidance, and reconstruction-oriented objectives [4, 15, 6]. These developments indicate that semantic alignment to powerful pretrained visual encoders is a promising direction for EEG representation learning.

Despite this progress, two limitations remain visible in the literature. First, many approaches rely primarily on direct classification or generative reconstruction and therefore optimize the model for end-task prediction rather than for a semantically organized cross-modal embedding space. Second, even when transformer modules are used, EEG is often modeled in a way that does not explicitly disentangle where information is distributed across electrodes from when evidence accumulates across time. However, these two axes are central to the structure of EEG: electrodes provide complementary but correlated views of cortical activity, while temporal evolution carries the dynamics of visual processing. A representation learning framework that explicitly models both dimensions may therefore be better aligned with the nature of the signal itself.

To address this gap, we propose a dual-stage transformer architecture for EEG-image alignment. Our method first applies channel-wise self-attention to model inter-electrode dependencies without collapsing the temporal dimension. The resulting channel-refined representation is then passed to a temporal self-attention module equipped with a learnable `<CLS>` token, which aggregates temporally distributed evidence into a compact EEG embedding. Rather than training the model with a standard classification loss, we align EEG embeddings to frozen pretrained SigLIP image features using a class-supervised contrastive objective. This formulation encourages semantically related EEG trials to cluster around a structured visual feature space while preserving instance-level variability inside each class.

Conceptually, the proposed design is simple but important. The channel-wise stage asks the model to first organize the spatial structure of the EEG recording by learning which electrodes interact most strongly for a given input. The temporal stage then integrates this refined representation over time to form a summary embedding. In this way, the model does not force a single attention module to solve two distinct problems at once. Instead, it decomposes EEG representation learning into two sequential operations that are both interpretable and well matched to the data modality.

Our work makes three contributions:

- We introduce a dual-stage EEG transformer that explicitly separates channel-wise relational modeling from temporal semantic aggregation, providing a principled architecture for EEG representation learning in EEG-image alignment.
- We formulate EEG visual decoding as class-supervised contrastive alignment to a frozen pretrained SigLIP visual space, enabling semantically structured EEG embeddings without relying on task-specific image reconstruction or direct closed-set classification.
- Through ablation and baseline comparison, we show that preserving channel tokens before temporal aggregation is a key design principle: the best-performing configuration avoids channel-stage aggregation and uses a temporal `<CLS>` token for final summarization.

2 Related Work

2.1 EEG-based visual decoding and multimodal alignment

Early work on EEG-based visual decoding established that EEG recordings contain information about viewed object categories and can be used to predict image-related labels [17]. Later studies shifted from direct classification toward multimodal learning, where EEG signals were paired with image features to construct joint latent spaces [11, 7]. This shift was important because it reframed the problem from predicting a fixed label set to learning semantically structured neural representations. More recent methods extended this direction through knowledge distillation, diffusion-based reconstruction, CLIP-style alignment, and visually or linguistically guided decoding [4, 2, 14, 12]. Relative to this line of work, our novelty is not generic contrastive alignment itself, but the architectural factorization of EEG encoding into a channel-relational stage followed by a temporal aggregation stage.

2.2 Transformer modeling for EEG

Transformer architectures have become increasingly important in EEG analysis because self-attention can capture long-range dependencies more flexibly than standard convolutional or recurrent modules. Hybrid architectures such as EEG Conformer combine shallow convolutional feature extraction with transformer encoders to learn local and global EEG structure [16]. Other recent spatio-temporal transformer variants similarly aim to improve EEG decoding by modeling channel interactions and temporal evolution more explicitly. Our method is aligned with this trend, but differs in using a *sequential* decomposition in which channel-wise refinement precedes all global temporal summarization.

2.3 Contrastive learning and target visual spaces

Contrastive learning has emerged as a powerful framework for organizing semantically meaningful embedding spaces across modalities. In vision-language

modeling, pretrained encoders such as SigLIP provide rich visual representations learned at scale [19]. In supervised representation learning, label-aware contrastive objectives encourage same-class samples to cluster while preserving finer-grained structure [8]. Our work adopts this perspective for EEG-image alignment: rather than claiming a novel contrastive loss, we use an established label-aware objective together with a frozen visual space and focus the main methodological contribution on the EEG encoder design.

3 Method

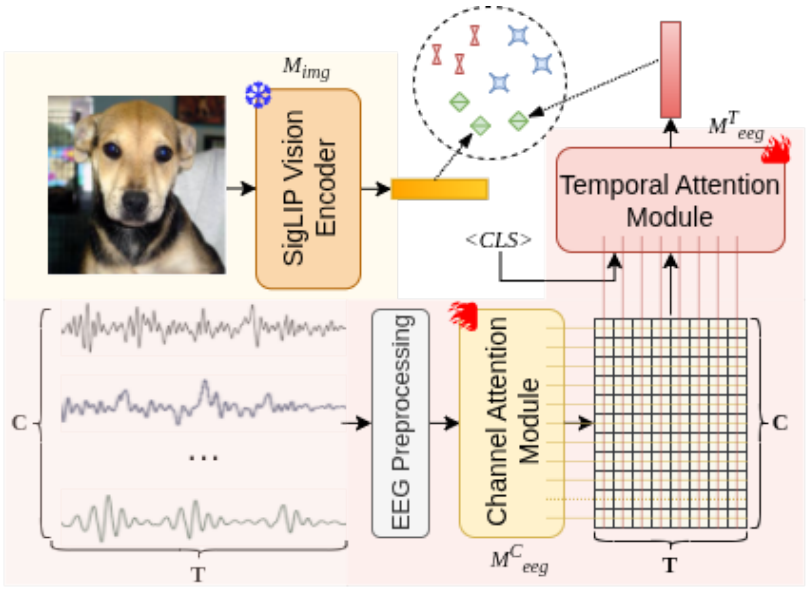

Fig. 1. Proposed dual-stage channel-wise and temporal self-attention architecture for EEG–image alignment.

3.1 Problem formulation

Let an EEG trial be denoted by $X \in \mathbb{R}^{C\times T}$, where $C = 128$ is the number of EEG channels and $T = 440$ denotes the number of temporal samples retained from the post-stimulus interval. Each EEG trial is paired with an image I and a semantic label $y \in \{1, \ldots, K\}$, where $K = 40$ for EEGCVPR40 [17]. Our goal is to learn an EEG encoder that maps X into a compact embedding space aligned with pretrained image features extracted from the associated visual stimulus.

Formally, we seek an encoder M_{eeg} and projection head $g(\cdot)$ such that

$$z^{eeg} = g(M_{eeg}(X)) \in \mathbb{R}^d, \tag{1}$$

and such that z^{eeg} is close to the image embedding

$$z^{img} = M_{img}(I) \in \mathbb{R}^d, \tag{2}$$

when the two samples belong to the same semantic category. Instead of optimizing only for closed-set classification, we learn this representation through class-supervised contrastive alignment.

3.2 Dataset and preprocessing

Experiments are conducted on EEGCVPR40 [17], which contains EEG recordings collected while subjects viewed images from 40 object categories. Each trial consists of a post-stimulus EEG response paired with the image that elicited it and the corresponding class label. Following our setup, each EEG trial is preprocessed and represented as a tensor of shape 128×440, corresponding to 128 channels over a 440 ms analysis window. This representation is used as the input to the proposed EEG encoder.

3.3 Frozen image encoder

To provide a semantically rich target space for alignment, we use a frozen pretrained SigLIP image encoder M_{img}. The image branch is not fine-tuned on EEGCVPR40. Instead, it serves as a stable source of visual embeddings with strong semantic organization learned from large-scale image–text data. Given an image I_i, the encoder outputs an image feature

$$z_i^{img} = \mathrm{norm}(M_{img}(I_i)), \tag{3}$$

where $\mathrm{norm}(\cdot)$ denotes ℓ_2 normalization. Freezing the image encoder has two advantages in our setting: it reduces the risk of overfitting on a relatively small EEG benchmark, and it forces the EEG branch to adapt to an already structured semantic space rather than co-adapting jointly with the visual encoder.

3.4 Dual-stage EEG encoder

The core of the proposed method is a dual-stage transformer that separately models channel-wise and temporal structure.

Channel-wise self-attention. EEG recordings are first processed by a channel-wise self-attention module M_{eeg}^C designed to capture inter-electrode dependencies. The input tensor $X \in \mathbb{R}^{C\times T}$ is treated as a structured multichannel signal in which each channel provides a complementary view of the same underlying perceptual event. Channel-wise self-attention learns pairwise relationships among electrodes and produces a channel-refined representation

$$H^C = M_{eeg}^C(X), \qquad H^C \in \mathbb{R}^{C\times T}. \tag{4}$$

A central design choice in our method is that this stage does *not* aggregate the channel dimension in the final model. Instead, all channel-refined features are preserved and passed to the temporal stage. This allows the temporal module to operate on a signal that has already been reorganized according to inter-electrode dependencies, without sacrificing channel-specific information too early.

Temporal self-attention. The second stage applies temporal self-attention through a transformer module M_{eeg}^{T}. We first convert the channel-refined representation into a sequence of temporal tokens. For each time index t, the corresponding channel vector $H_{:,t}^{C} \in \mathbb{R}^{C}$ is projected into the transformer hidden space:

$$u_t = W_t H_{:,t}^{C} + b_t, \qquad u_t \in \mathbb{R}^{d}. \tag{5}$$

These tokens are arranged into the sequence

$$U = [u_1, u_2, \ldots, u_T] \in \mathbb{R}^{T \times d}. \tag{6}$$

A learnable `<CLS>` token u_{cls} is prepended,

$$U^{0} = [u_{\text{cls}}; U] + E, \tag{7}$$

where E denotes learnable positional encodings. The temporal transformer then computes

$$H^{T} = M_{eeg}^{T}(U^{0}), \tag{8}$$

and the final EEG representation is taken from the output associated with the `<CLS>` token:

$$h^{eeg} = H_{0}^{T}. \tag{9}$$

This design allows the model to aggregate evidence across time through attention rather than simple averaging. Importantly, temporal aggregation occurs *after* channel relations have already been modeled, which is a key distinction from single-stage architectures.

Projection head. The temporal representation is projected into the same embedding space as the image features:

$$z_i^{eeg} = \text{norm}(W_p h_i^{eeg} + b_p), \tag{10}$$

where W_p and b_p are learnable parameters and $\text{norm}(\cdot)$ denotes ℓ_2 normalization.

3.5 Class-supervised contrastive alignment

Given a mini-batch of paired EEG trials and images, we align EEG embeddings with image embeddings using a class-supervised InfoNCE objective. Let z_i^{eeg} be the EEG embedding of sample i, and let $\{z_j^{img}\}_{j=1}^{N}$ be the image embeddings in the batch. For anchor i, the positive set is defined as

$$P(i) = \{j \mid y_j = y_i\}, \tag{11}$$

that is, all image features in the batch sharing the same semantic label as the EEG trial. The similarity between EEG and image embeddings is measured by cosine similarity:

$$s(i,j) = \frac{(z_i^{eeg})^\top z_j^{img}}{\|z_i^{eeg}\|_2 \|z_j^{img}\|_2}. \tag{12}$$

The loss for anchor i is

$$\mathcal{L}_i = -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(s(i,p)/\tau)}{\sum_{a=1}^{N} \exp(s(i,a)/\tau)}, \tag{13}$$

where τ is a temperature parameter. The total loss over a batch is

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_i. \tag{14}$$

This objective differs from standard pairwise contrastive training in two ways. First, it takes advantage of semantic labels by treating same-class examples as positives rather than only the exact paired image. Second, it encourages the EEG branch to organize itself around a semantically meaningful visual space rather than to merely separate the training classes by a discriminative classifier. This is particularly useful in EEG decoding, where sample efficiency and representation structure are both important.

3.6 Training and evaluation

During training, the image encoder is kept frozen in order to preserve the semantic structure of the pretrained visual embedding space. In contrast, all EEG-side modules are optimized end-to-end, allowing the model to learn a mapping from EEG signals to the corresponding visual-semantic representations. This design ensures that the EEG encoder is trained to align neural responses with a stable visual target space rather than adapting both modalities simultaneously.

We evaluate the learned EEG representations on both the training and test splits using top-k classification accuracy and mean Average Precision (mAP). These metrics measure how well the EEG embeddings retrieve or rank the correct visual class among a set of candidate classes.

Given an EEG query embedding $\mathbf{z}_i^e$ and a set of candidate visual embeddings $\{\mathbf{z}_j^v\}_{j=1}^N$, we compute a similarity score, for example using cosine similarity, as shown in Formula 12.

The candidates are then ranked in descending order according to s_{ij}. Let y_i denote the ground-truth class label of the i-th EEG sample, and let $\mathcal{R}_i^{(k)}$ be the set of class labels appearing in the top-k ranked candidates for that sample. The top-k accuracy is defined as

$$\text{Acc@}k = \frac{1}{M} \sum_{i=1}^{M} \mathbb{I}\left[y_i \in \mathcal{R}_i^{(k)}\right], \tag{15}$$

where M is the number of EEG samples and $\mathbb{I}[\cdot]$ is the indicator function. In this work, we report Acc@1 and Acc@5:

$$\text{Acc@1} = \frac{1}{M} \sum_{i=1}^{M} \mathbb{I}\left[y_i = \hat{y}_i^{(1)}\right], \tag{16}$$

$$\text{Acc@5} = \frac{1}{M} \sum_{i=1}^{M} \mathbb{I}\left[y_i \in \{\hat{y}_i^{(1)}, \hat{y}_i^{(2)}, \ldots, \hat{y}_i^{(5)}\}\right], \tag{17}$$

where $\hat{y}_i^{(r)}$ denotes the class label of the candidate ranked at position r.

To further evaluate the quality of the full ranking, we report mean Average Precision. For each query i, the Average Precision is computed as

$$\text{AP}_i = \frac{1}{|\mathcal{G}_i|} \sum_{r=1}^{N} P_i(r)\, \text{rel}_i(r), \tag{18}$$

where $\mathcal{G}_i$ is the set of relevant candidates for query i, $\text{rel}_i(r) \in \{0, 1\}$ indicates whether the candidate ranked at position r is relevant, and $P_i(r)$ is the precision at rank r:

$$P_i(r) = \frac{1}{r} \sum_{t=1}^{r} \text{rel}_i(t). \tag{19}$$

The final mAP is obtained by averaging the Average Precision values over all queries:

$$\text{mAP} = \frac{1}{M} \sum_{i=1}^{M} \text{AP}_i. \tag{20}$$

Together, Acc@1 measures whether the highest-ranked candidate corresponds to the correct class, Acc@5 measures whether the correct class appears within the five most similar candidates, and mAP captures the overall ranking quality across all candidate classes. Higher values for these metrics indicate stronger alignment between EEG embeddings and the semantic organization of the visual target space.

4 Experiments and Results

4.1 Implementation Details

In this work, we train the proposed model with a batch size of 512 and an initial learning rate of 1×10^{-4}. The visual backbone is initialized using the pre-trained SigLIP weights `google/siglip-base-patch16-224` from Hugging Face. The embedding dimension is set to 768 for both EEG and visual representations.

Optimization is performed using the AdamW optimizer with a weight decay of 1×10^{-4}. We adopt a cosine learning rate scheduler with a warm-up period of 10 epochs, and the minimum learning rate is set to 1×10^{-6}. All experiments are conducted on an NVIDIA A40 GPU.

4.2 Experimental setup

We compare the proposed method to CNN- and transformer-based baselines with comparable depth and parameter count. This comparison is important because it isolates the architectural contribution of dual-stage structured attention rather than simply attributing gains to model scale. The proposed model contains 99.35M parameters, compared with 93.05M for the CNN baseline and 96.24M for the transformer baseline. Ablation experiments are conducted for 200 training epochs to study the effect of aggregation choices in the channel-wise and temporal modules.

4.3 Ablation on aggregation strategy

Table 1 evaluates six variants formed by combining different aggregation strategies for the channel-wise module (`Chan`) and temporal module (`Tim`). Three main observations emerge.

Table 1. Ablation study of channel-wise (Chan) and temporal (Tim) attention modules in 200 training epochs. '−' indicates no aggregation, 'CLS' denotes aggregation via a learnable `<CLS>` token, and 'MEAN' denotes mean pooling.

Chan	Tim	Train Set			Test Set		
		Acc@1↑	Acc@5↑	mAP↑	Acc@1↑	Acc@5↑	mAP↑
CLS	CLS	0.645	0.811	0.613	0.586	0.744	0.561
CLS	MEAN	0.854	0.930	0.796	0.871	0.931	0.792
MEAN	CLS	0.797	0.936	0.744	0.717	0.889	0.673
MEAN	MEAN	0.996	0.998	0.958	0.954	0.965	0.906
−	CLS	0.994	0.998	0.957	**0.993**	<u>0.995</u>	**0.947**
−	MEAN	0.968	0.997	0.890	<u>0.991</u>	**0.999**	<u>0.874</u>

First, *premature aggregation in the channel stage is harmful.* The weakest configuration is `CLS and CLS`, where both modules compress information through a learned token. This result suggests that collapsing the channel dimension too early discards useful electrode-specific structure before temporal reasoning can exploit it.

Second, *retaining the full channel-refined representation is highly effective.* The best overall test performance is obtained by `Chan− and Tim CLS`, which reaches 0.993 Acc@1, 0.995 Acc@5, and 0.947 mAP. This indicates that the channel-wise module is most useful as a relational refinement stage rather than as a global pooling stage. In other words, its role is to reorganize the signal across electrodes, not to summarize it into a single vector.

Third, *the temporal stage is the right place for final aggregation.* When channel aggregation is removed, using a temporal `<CLS>` token gives the strongest

Acc@1 and mAP, while mean pooling gives slightly higher Acc@5. This pattern is intuitive: a learnable temporal token can focus on the most discriminative temporal evidence, whereas mean pooling produces a broader but less selective summary. Since our goal is discriminative semantic alignment, the `Chan− and Tim CLS` configuration is adopted as the final model.

The comparison between the last two variants isolates the effect of temporal aggregation when channel-wise aggregation is completely removed. Both variants preserve the full channel-refined representation after the channel attention stage, but differ in whether temporal evidence is summarized by a learnable `<CLS>` token or by mean pooling. The *Chan−* and *Tim* `<CLS>` variant achieves the best test Acc@1 and mAP, reaching 0.993 and 0.947, respectively, whereas *Chan−* and *Tim MEAN* obtains a slightly higher Acc@5 but a substantially lower mAP. This suggests that mean pooling provides a broader and smoother summary that may retain the correct class within the top-ranked candidate set, but lacks the selectivity required for precise semantic ranking. In contrast, the temporal `<CLS>` token acts as an adaptive query over time, allowing the model to emphasize discriminative temporal evidence while suppressing less informative or noisy intervals. The large mAP gap indicates that the benefit of `<CLS>` aggregation is not merely a marginal top-1 improvement, but rather a stronger global alignment between EEG embeddings and the frozen visual semantic space. These results support the central architectural principle of the proposed model, channel-wise attention should refine inter-electrode dependencies without premature compression, while final semantic aggregation should be deferred to the temporal stage and performed by a learnable, selective mechanism.

Overall, the ablation study supports a simple but important design principle: preserve fine-grained channel information for as long as possible, and perform global summarization only after temporal dependencies have been modeled.

4.4 Comparison with baselines

Table 2 compares our final model with CNN and transformer baselines of comparable parameter count. The proposed method substantially outperforms both alternatives on all reported test metrics. In particular, our method achieves 0.993 Acc@1 and 0.947 mAP on the test set, compared with 0.609 and 0.475 for the CNN baseline, and 0.801 and 0.667 for the transformer baseline.

This margin is noteworthy for two reasons. First, it shows that simply using a transformer is not sufficient; the way attention is structured matters. The transformer baseline is clearly stronger than the CNN baseline, which confirms the value of attention-based modeling for EEG. However, the proposed dual-stage design still yields a large additional gain, indicating that explicitly separating channel-wise and temporal reasoning provides a more suitable inductive bias for EEG–image alignment.

Second, the gains are obtained without dramatically increasing parameter count. The performance improvement therefore appears to stem from better representation learning rather than brute-force scaling. The proposed model uses the visual foundation model only as a frozen target space and learns the full

Table 2. Comparison with CNN and transformer baselines of comparable depth and parameter count.

Method	Params	Train Set			Test Set		
	(M)	Acc@1↑	Acc@5↑	mAP↑	Acc@1↑	Acc@5↑	mAP↑
CNN	93.05	0.844	0.984	0.827	0.609	0.770	0.475
Trans.	96.24	0.830	0.948	0.711	0.801	0.905	0.667
Ours	99.35	0.994	0.998	0.957	**0.993**	**0.995**	**0.947**

improvement on the EEG side through structured attention and contrastive supervision.

Table 3 compares the proposed dual-stage self-attention framework with recent state-of-the-art EEG-vision methods. Under the EEGCVPR40 benchmark protocol used in this work, our method achieves the strongest reported performance, with 0.993 Acc@1 and 0.995 Acc@5. These results indicate that explicitly modeling EEG structure in two stages before cross-modal alignment is beneficial: channel-wise self-attention captures inter-electrode dependencies while preserving temporal detail, and temporal self-attention then consolidates the resulting sequence into a compact embedding for contrastive alignment with the frozen SigLIP image space. Compared with prior methods, the proposed design yields more robust retrieval performance rather than relying only on a generic backbone or direct projection into the visual embedding space. The high mAP of 0.947 further suggests that the learned EEG embedding space is not only accurate at top-rank retrieval, but also consistently well organized across the ranked candidate set.

Table 3. Comparison with the recent state-of-the-art methods.

Method	Year	Acc@1↑	Acc@5↑	mAP↑
MB2C [18]	2024	0.937	0.992	–
SYNAPSE [9]	2025	0.602	0.952	–
Uni-Neur2Img [1]	2025	0.930	–	–
AVDE [3]*	2026	0.300	0.582	–
Ours	2026	**0.993**	**0.995**	**0.947**

* Results are reported under a zero-shot setting and are therefore not directly comparable with those of the other methods.

4.5 Discussion

The results suggest that the success of the proposed method comes from the interaction of three components rather than from any single ingredient alone. The

frozen SigLIP encoder provides a semantically structured visual manifold. The class-supervised contrastive loss encourages EEG trials from the same category to align with that manifold in a label-aware way. The dual-stage EEG encoder supplies an architecture capable of preserving and transforming EEG structure before alignment occurs.

The ablation further reveals an interpretable mechanism behind the model's behavior. Channel-wise attention is most effective when it acts as a refinement operator over electrodes, not as an early pooling mechanism. Temporal attention then benefits from receiving this richer intermediate representation and from using a dedicated learnable token to integrate distributed evidence across time. This division of labor gives the model a clearer internal organization than architectures that attempt to jointly collapse all dimensions at once.

At the same time, the present findings should be interpreted as benchmark-level evidence rather than a complete statement about general EEG visual decoding in the wild. EEGCVPR40 is a relatively small and historically important benchmark, but block-design EEG experiments have also motivated broader methodological discussion in the literature. Accordingly, the strength of the reported results is best understood as evidence that the proposed architectural bias is highly effective under the current setup, while broader validation across larger, more diverse, or more challenging protocols remains an important next step [10].

5 Conclusion

We presented a dual-stage self-attention framework for contrastive EEG–image alignment that explicitly separates channel-wise relational modeling from temporal semantic aggregation. The method aligns EEG embeddings to frozen pretrained SigLIP image features using a class-supervised contrastive objective, thereby encouraging semantically meaningful EEG representations without relying on direct closed-set classification or image generation as the primary learning signal. Across ablation and baseline comparisons, the results consistently show that preserving channel-refined features before temporal aggregation is a decisive architectural choice. The final model substantially outperforms comparable CNN and transformer baselines, demonstrating that *how* EEG structure is modeled matters as much as *which* backbone family is used.

Beyond the specific benchmark, the proposed framework offers a clean and extensible perspective on EEG representation learning. The architecture reflects the intrinsic structure of EEG by first modeling inter-electrode dependencies and then integrating evidence across time, while the contrastive training objective anchors the learned EEG features in a powerful semantic visual space. This combination provides a robust learning principle for future EEG-based applications in visual decoding, retrieval, semantic understanding, and multimodal brain representation learning.

6 Acknowledgements

This work has been partially supported by the Beatriu de Pinós del Departament de Recerca i Universitats de la Generalitat de Catalunya (2022 BP 00256); the predoctoral program AGAUR-FI ajuts (2026 FI-3 00470) Joan Oró, which are backed by the Secretariat of Universities and Research of the Department of Research and Universities of the Generalitat of Catalonia, as well as the European Social Plus Fund; the Grant PID2021-128178OB-I00, PID2023-146426NB-100, PID2023-151083NA-I00, PID2024-162555OB-I00 funded by MICIU/AEI/10.13039/501100011033, ERDF, EU "A way of making Europe" and by the Generalitat de Catalunya CERCA Program; the European Social Plus Fund, European Lighthouse on Safe and Secure AI (ELSA) from the European Union's Horizon Europe programme under grant agreement No 101070617.

References

1. Bai, X., Yu, R., Xiu, J., Zhou, P., Xia, J., Ji, P.: Uni-neur2img: Unified neural signal-guided image generation, editing, and stylization via diffusion transformers. arXiv preprint arXiv:2512.18635 (2025)
2. Camaret Ndir, T., Schirrmeister, R.T., Ball, T.: Eeg-clip: Learning eeg representations from natural language descriptions. Frontiers in Robotics and AI **12**, 1625731 (2025)
3. Dai, S., Xiao, H., Yu, S., Ye, Q.: Autoregressive visual decoding from eeg signals. In: The Fourteenth International Conference on Learning Representations (2026)
4. Ferrante, M., Boccato, T., Bargione, S., Toschi, N.: Decoding visual brain representations from electroencephalography through knowledge distillation and latent diffusion models. Computers in Biology and Medicine **178**, 108701 (2024). https://doi.org/10.1016/j.compbiomed.2024.108701
5. Gomez-Andres, A., Cerda-Company, X., Cucurell, D., Cunillera, T., Rodríguez-Fornells, A.: Decoding agency attribution using single trial error-related brain potentials. Psychophysiology **61**(1), e14434 (2024)
6. Guenther, S., Kosmyna, N., Maes, P.: Image classification and reconstruction from low-density eeg. Scientific Reports **14**, 16436 (2024). https://doi.org/10.1038/s41598-024-66228-1
7. Horikawa, T., Kamitani, Y.: Generic decoding of seen and imagined objects using hierarchical visual features. Nature Communications **8**, 15037 (2017). https://doi.org/10.1038/ncomms15037
8. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in neural information processing systems **33**, 18661–18673 (2020)
9. Lee, J., Kang, H.: Synapse: Synergizing an adapter and finetuning for high-fidelity eeg synthesis from a clip-aligned encoder. arXiv preprint arXiv:2511.17547 (2025)
10. Li, R., Johansen, J.S., Ahmed, H., Ilyevsky, T.V., Wilbur, R.B., Bharadwaj, H.M., Siskind, J.M.: The perils and pitfalls of block design for eeg classification experiments. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(1), 316–333 (2020)
11. Palazzo, S., Spampinato, C., Kavasidis, I., Giordano, D., Schmidt, J., Shah, M.: Decoding brain representations by multimodal learning of neural activity and visual

features. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(11), 3833–3849 (2021). https://doi.org/10.1109/TPAMI.2020.2995909
12. Rajabi, N., Ribeiro, A.H., Vasco, M., Taleb, F., Björkman, M., Kragic, D.: Human-aligned image models improve visual decoding from the brain. In: Forty-second International Conference on Machine Learning (2025)
13. Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T.H., Faubert, J.: Deep learning-based electroencephalography analysis: a systematic review. Journal of Neural Engineering **16**(5), 051001 (2019). https://doi.org/10.1088/1741-2552/ab260c
14. Shi, E., Hu, H., Yuan, Q., Zhao, K., Yu, S., Zhang, S.: Brainalign: Eeg-vision alignment via frequency-aware temporal encoder and differentiable cluster assigner. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 98–108. Springer (2025)
15. Song, Y., Wang, Y., He, H., Gao, X.: Recognizing natural images from eeg with language-guided contrastive learning. IEEE Transactions on Neural Networks and Learning Systems (2025)
16. Song, Y., Zheng, Q., Liu, B., Gao, X.: Eeg conformer: Convolutional transformer for eeg decoding and visualization. IEEE Transactions on Neural Systems and Rehabilitation Engineering **31**, 710–719 (2022)
17. Spampinato, C., Palazzo, S., Kavasidis, I., Giordano, D., Souly, N., Shah, M.: Deep learning human mind for automated visual classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4503–4511 (2017). https://doi.org/10.1109/CVPR.2017.479
18. Wei, Y., Cao, L., Li, H., Dong, Y.: Mb2c: Multimodal bidirectional cycle consistency for learning robust visual neural representations. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 8992–9000 (2024)
19. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)