Shrinking the Teacher: An Adaptive Teaching Paradigm for Asymmetric EEG-Vision Alignment

Lukun Wu[1], **Jie Li**[1*], **Ziqi Ren**[2], **Kaifan Zhang**[1], **Xinbo Gao**[1*]

[1]School of Electronic Engineering, Xidian University, Xi'an, China
[2]School of Life Science and Technology, Xidian University, Xi'an, China
lkwu@stu.xidian.edu.cn, {leejie, xbgao}@mail.xidian.edu.cn

Abstract

Decoding visual features from EEG signals is a central challenge in neuroscience, with cross-modal alignment as the dominant approach. We argue that the relationship between visual and brain modalities is fundamentally asymmetric, characterized by two critical gaps: a Fidelity Gap (stemming from EEG's inherent noise and signal degradation, vs. vision's high-fidelity features) and a Semantic Gap (arising from EEG's shallow conceptual representation, vs. vision's rich semantic depth). Previous methods often overlook this asymmetry, forcing alignment between the two modalities as if they were equal partners and thereby leading to poor generalization. To address this, we propose the adaptive teaching paradigm. This paradigm empowers the "teacher" modality (vision) to dynamically shrink and adjust its knowledge structure under task guidance, tailoring its semantically dense features to match the "student" modality (EEG)'s capacity. We implement this paradigm with the ShrinkAdapter, a simple yet effective module featuring a residual-free design and a bottleneck structure. Through extensive experiments, we validate the underlying rationale and effectiveness of our paradigm. Our method achieves a top-1 accuracy of 60.2% on the zero-shot brain-to-image retrieval task, surpassing previous state-of-the-art methods by a margin of 9.8%. Our work introduces a new perspective for asymmetric alignment: the teacher must shrink and adapt to bridge the vision-brain gap.

Code — https://github.com/LukunWuXDU/ATS
Extended version — https://arxiv.org/abs/2511.11422

1 Introduction

Visual neural decoding aims to interpret visual content from brain activity, serving as a bridge between human cognition and artificial intelligence while deepening our understanding of the human visual mechanism. Among various neuroimaging techniques, electroencephalography (EEG) has attracted significant attention due to its non-invasive nature, high temporal resolution, and portability, endowing it with greater potential for brain-computer interface (BCI) applications.

The dominant approach currently decodes visual content by aligning EEG signals with pretrained visual features. While most previous methods acknowledge the differences between the EEG and visual modalities, they still treat the alignment task as a symmetric problem, implicitly assuming that the two modalities have comparable fidelity and capacity. In contrast, we argue that such modality differences are inherently **asymmetric**.

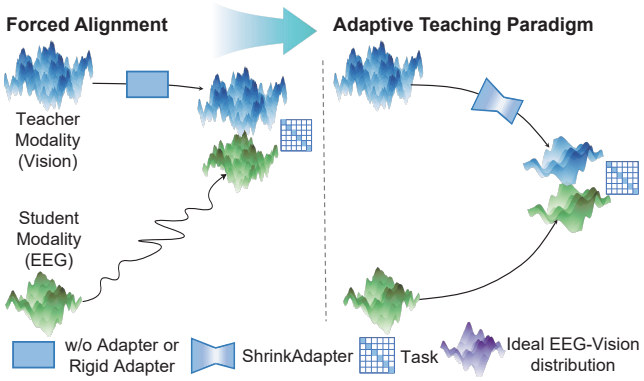

Figure 1: From Forced Alignment to Adaptive Teaching: A paradigm shift for asymmetric modality alignment.

We deconstruct this asymmetry into two core components. The first is a **Fidelity Gap**, arising from the physical limitations of EEG acquisition. As illustrated in Figure 2, spatially, the sparse distribution of electrodes and the inherent **volume conduction effect** (Michel and Murray 2012) lead to significant spatial blurring, as neural signals attenuate and diffuse while propagating through the head. This issue is compounded temporally by the **temporal aliasing** in Rapid Serial Visual Presentation (RSVP) paradigms (Grootswagers, Robinson, and Carlson 2019; Keysers et al. 2001), leading to significant cross-stimulus interference. These factors degrade the signal quality, resulting in a low-fidelity representation that starkly contrasts with the clean, detailed features from vision models. The second component is a more subtle **Semantic Gap**. It is questionable whether the human brain, within a fleeting $100 - 200$ ms exposure, can form a neural representation as semantically rich and nuanced as that of a large vision model trained on billions of images. The resulting brain signal, therefore, occupies a much smaller and less structured semantic subspace.

Given the profound asymmetry between the "teacher"

*Corresponding author.

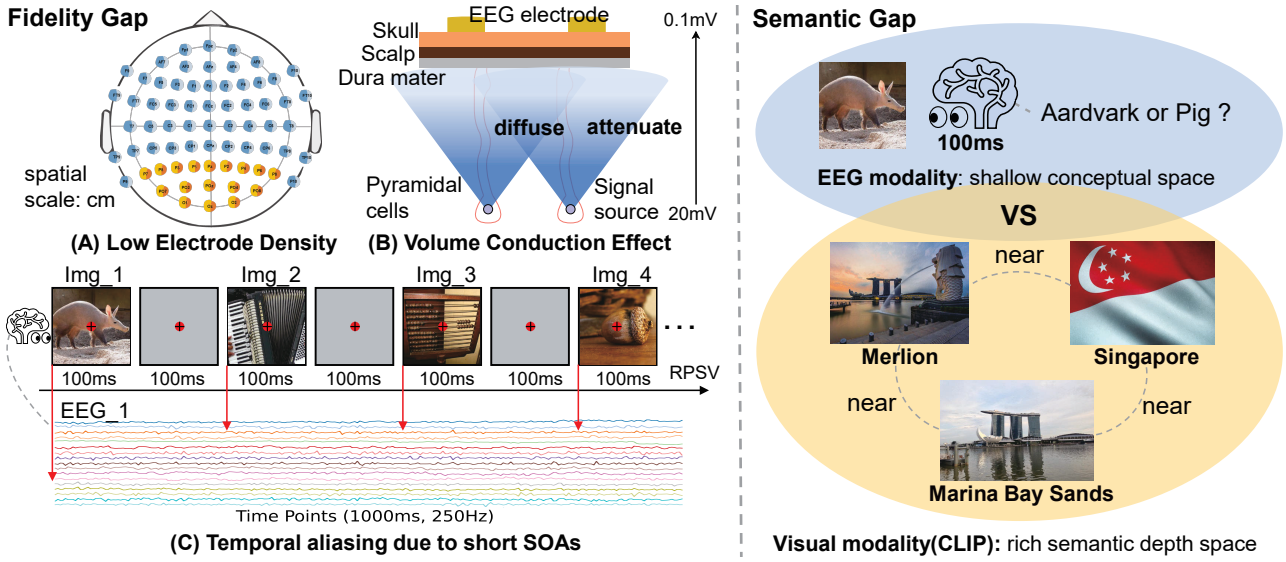

Figure 2: The physiological basis for our motivation: Deconstructing the asymmetric modality gap between vision and EEG into Fidelity Gap and Semantic Gap.

modality (the visual features from a pretrained vision model) and the "student" modality (the EEG signal), forcing the student to directly learn from the fixed teacher—a process we term **Forced Alignment**—is an ill-posed strategy that often leads to overfitting to noise. This motivates a conceptual shift to our **Adaptive Teaching Paradigm**, as shown in Figure 1. Inspired by the pedagogical principle of teaching according to aptitude, we argue the teacher modality must dynamically shrink and adjust its knowledge structure under **task guidance**, tailoring its semantically dense features to match the student modality (EEG)'s capacity, thereby achieving a more robust and generalizable alignment.

To realize this paradigm, we introduce the **Adaptive Teaching System (ATS)**, a framework designed to better bridge the vision-brain gap, thereby improving neural decoding performance. For the teacher, we develop the **ShrinkAdapter**, a simple yet efficient module that provides the conditions for the teacher modality to unconstrainedly shrink and adjust its knowledge structure. For the student, we design the **Shared Temporal Attention Encoder (STAE)** to effectively extract salient features from temporally noisy EEG data.

Our contributions can be summarized as follows:

1. To our knowledge, we are the first to deconstruct the vision-brain modality gap into an asymmetric problem, comprising a Fidelity Gap and a Semantic Gap.
2. We introduce a framework called Adaptive Teaching System (ATS), which implements our Adaptive Teaching Paradigm through the ShrinkAdapter. Additionally, the Shared Temporal Attention Encoder (STAE) enhances feature extraction from EEG signals.
3. Through extensive experiments, we validate the underlying rationale and effectiveness of our paradigm. Our work achieves SOTA performance and provides new insights into asymmetric alignment tasks in neuroscience.

2 Related Work

2.1 Visual Neural Decoding

Visual neural decoding aims to interpret brain activity to retrieve, recognize, or reconstruct visual stimuli. While early efforts often utilized fMRI for its high spatial resolution (Ren et al. 2021; Takagi and Nishimoto 2023; Scotti et al. 2024), EEG has recently gained significant attention due to its high temporal resolution, low cost, and portability (Spampinato et al. 2017; Jiao et al. 2019). The advent of large-scale datasets like THINGS-EEG (Gifford et al. 2022), coupled with pioneering frameworks like BraVL (Du et al. 2023) and NICE (Song et al. 2024), laid the groundwork for the current research landscape. Consequently, zero-shot EEG-to-image retrieval based on self-supervised contrastive learning has been cemented as the dominant paradigm, making the pursuit of a more effective and robust vision-brain alignment the next core challenge for the field.

2.2 Multi-Modal Alignment in EEG Retrieval

The prevailing approach for EEG retrieval is contrastive learning, which aligns EEG features with pretrained visual embeddings in a shared latent space. Previous works have generally improved upon this paradigm from two main directions. The first is **information enhancement**. Based on the premise that human cognition is inherently multi-modal, these methods enrich the learning process with auxiliary data to better interpret the rich information within EEG signals. For instance, BraVL and NICE++ (Song et al. 2025) incorporate an additional text modality, while CognitionCapturer (Zhang et al. 2025) even introduces depth map information. The second direction is **constraint reinforcement**. These methods acknowledge the existence of the modality differences and design more sophisticated alignment strategies to mitigate it. For example, MB2C (Wei et al. 2024) introduces a cycle-consistency loss, and VE-SDN (Chen et al.

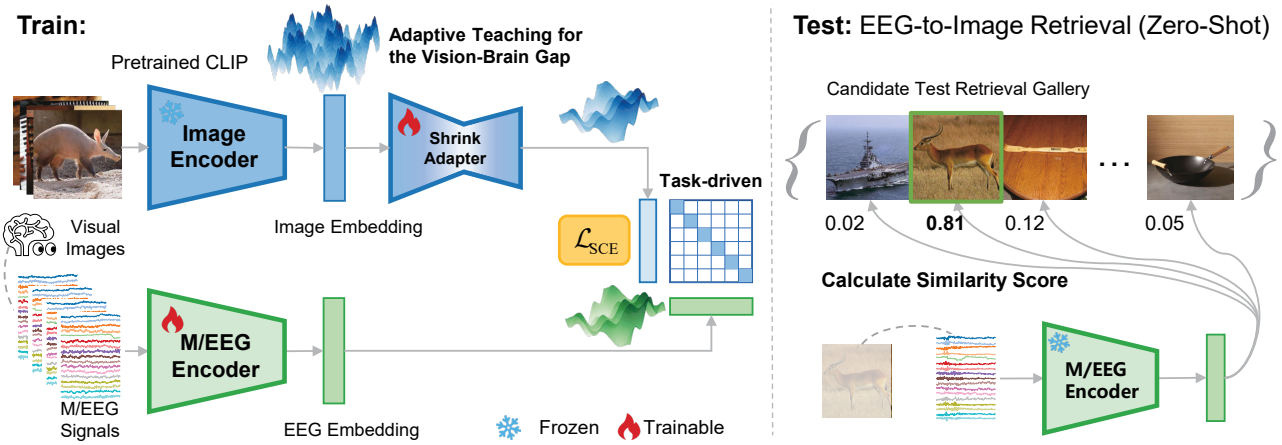

Figure 3: Overview of the Adaptive Teaching System for zero-shot EEG-to-Image retrieval. Training: A ShrinkAdapter enables the visual "teacher" to adapt its features for alignment with the EEG "student" via InfoNCE loss. Testing: The trained student encoder performs zero-shot image retrieval from a candidate set.

2024) constructs a joint semantic space to improve alignment.

While these approaches have advanced the field significantly, they largely ignore the inherent asymmetry between the visual and EEG modalities, leading to suboptimal generalization performance. A notable exception is the UBP (Wu et al. 2025), which, for the first time, approached the problem from the biological characteristics of EEG signals. It introduced a dynamically adjusted blur prior to bridge what it termed the "System Gap" and "Random Gap". This hand-crafted prior proved highly effective. However, being manually designed, it is neither comprehensive nor flexible enough to capture the full complexity of the modality gap. This limitation motivates us to re-examine this gap and propose the **Adaptive Teaching Paradigm** as a more fundamental and flexible solution.

2.3 Feature Adaptation for Pre-trained Models

A projection layer is typically used when adapting large pretrained models for downstream tasks, primarily for light feature fine-tuning. (Kumar et al. 2022) Yet, when applied to the vision-EEG field, this approach is often adopted without critical examination, leaving its true potential and strategic importance critically overlooked. Consequently, prior works either omit this projection entirely (e.g., UBP), placing the full learning burden on the "student," or treat it as a generic stabilizer (e.g., NICE, MB2C), failing to harness its full potential to address the asymmetric modality gap.

3 Methodology

3.1 Problem Formulation

Our ultimate goal is to decode the visual features of stimuli directly from brain signals, thereby exploring the informational capacity of EEG signals and advancing our understanding of the human visual mechanism. The zero-shot retrieval task serves as an effective paradigm for this objective. Following the standard setup (Song et al. 2024), we are given a training dataset $D_{train} = \{(x_v, x_b, y)\}$, where x_v is an image, x_b is the corresponding brain signal (EEG/MEG), and y is the class label of x_v from a set of training classes, Y_{seen}. The model is trained only on D_{train}.

As illustrated in Figure 3, during the test phase, the model is evaluated on a test set $D_{test} = \{(x_v^u, x_b^u, y^u)\}$, where the labels y^u belong to a set of testing classes Y_{unseen}, ($Y_{seen} \cap Y_{unseen} = \emptyset$). For a given test brain signal x_b^u, the objective is to retrieve its corresponding image x_v^u from a candidate set X_v^u containing all images from Y_{unseen}. This is achieved by identifying the image feature with the highest similarity score to the brain feature in the learned shared latent space.

3.2 The Adaptive Teaching System (ATS)

To address the asymmetric modality gap, we introduce the Adaptive Teaching System (ATS), illustrated in Figure 3. The framework comprises two main branches that map inputs into a shared latent space.

The **visual branch** (the "teacher") first uses a powerful pretrained vision encoder f_V (e.g., from CLIP) (Radford et al. 2021) to extract a high-dimensional feature vector $h_v = f_V(x_v)$. Crucially, we then introduce a trainable ShrinkAdapter, denoted as f_A, which dynamically adapts this feature into a representation $z_v = f_A(h_v)$ that is more accessible for the student modality.

The **brain branch** (the "student") employs a trainable encoder f_B, which learns to map a brain signal x_b into an embedding $z_b = f_B(x_b)$ in the shared latent space.

The alignment between the adapted teacher and student is driven by a symmetric contrastive loss, whose objective is to pull positive pairs closer together while pushing negative pairs apart. Following the standard in-batch negative sampling setup, within a mini-batch of size N, a corresponding image-brain pair constitutes a positive pair, while all other non-corresponding pairings serve as negative pairs. We employ the Symmetric Cross-Entropy (SCE) loss, an objective derived from the InfoNCE loss (Wang et al. 2021; van den Oord, Li, and Vinyals 2019), to jointly optimize the

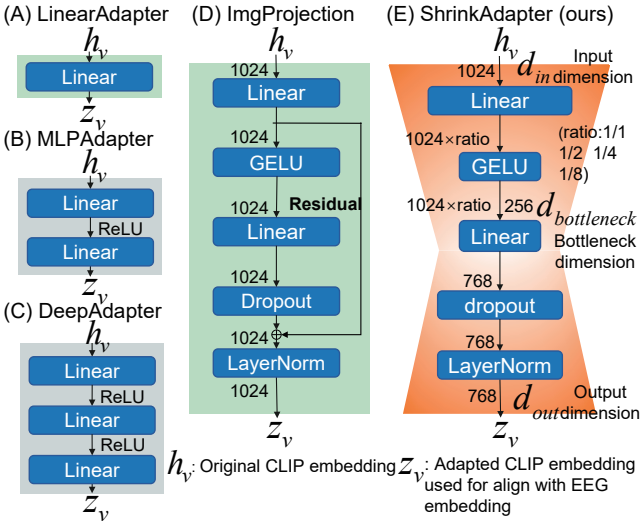

Figure 4: Structure of different adapters.

ShrinkAdapter f_A and the brain encoder f_B:

$$\mathcal{L}_{\text{SCE}} = -\frac{1}{2N}\sum_{i=1}^{N}\left[\log\frac{\exp(z_{v,i}^{\top}z_{b,i}/\tau)}{\sum_{k=1}^{N}\exp(z_{v,i}^{\top}z_{b,k}/\tau)} + \log\frac{\exp(z_{b,i}^{\top}z_{v,i}/\tau)}{\sum_{k=1}^{N}\exp(z_{b,i}^{\top}z_{v,k}/\tau)}\right], \quad (1)$$

where $(z_{v,i}, z_{b,i})$ represents a positive pair of L2-normalized embeddings, $(\cdot)^{\top}(\cdot)$ denotes cosine similarity, and τ is a learnable temperature parameter. This objective not only encourages the student to align with the teacher but, critically, compels the teacher (via the trainable ShrinkAdapter) to adapt its representation z_v to be more accessible to the student, thus realizing our adaptive teaching paradigm. The algorithmic flow is detailed in Appendix A.4.

3.3 ShrinkAdapter

Based on our analysis of the asymmetric modality gap, we argue that the pre-trained "teacher" modality contains knowledge that is complex and redundant relative to the "student's" capacity. Therefore, the core of adaptive teaching is to enable the teacher to adapt and shrink its knowledge to match the student's potential. We realize this through our ShrinkAdapter, a module guided by two core principles: adaptive freedom via a residual-free design, and information shrinking via a bottleneck.

First, to provide the teacher with the freedom to fully adapt, we deliberately **eliminate residual connections**. A residual connection (He et al. 2016), while commonly used in prior work to preserve the original feature distribution (Figure 4(D)), is fundamentally in conflict with our design philosophy. By removing this constraint, the adapter is free to learn an entirely new representation, driven by the task.

Second, to shrink the teacher's redundant information, we employ a **bottleneck architecture**. As illustrated in Figure 4(E), this structure forces the visual features through a low-dimensional bottleneck, thereby filtering out irrelevant information. The architecture is defined as:

$$z_v = f_A(h_v) = W_{up}\text{GELU}(W_{down}h_v), \quad (2)$$

where W_{down} and W_{up} are linear layers, and we define the bottleneck compression ratio as $d_{\text{bottleneck}}/d_{\text{in}}$, with the bottleneck dimension satisfying $d_{\text{bottleneck}} \ll d_{\text{in}}$.

The design of the ShrinkAdapter is theoretically motivated by the **Information Bottleneck (IB) principle** (Tishby, Pereira, and Bialek 2000). The IB principle seeks to learn a maximally compressed representation z_v of an input h_v that preserves the most relevant information about a target z_b. This trade-off is formalized by minimizing the Lagrangian:

$$\mathcal{L}_{\text{IB}} = I(h_v; z_v) - \beta I(z_v; z_b), \quad (3)$$

where $I(\cdot;\cdot)$ is the mutual information and β is a Lagrange multiplier. The ShrinkAdapter implements this principle through its core mechanisms. The **bottleneck architecture** is the primary mechanism for minimizing the compression term, $I(h_v; z_v)$, and this is enabled by the **residual-free design**, which removes a major obstacle to this goal. Meanwhile, our **alignment loss** ($\mathcal{L}_{\text{SCE}}$) serves as a proxy to maximize $I(z_v; z_b)$, the term representing the task-relevant information for alignment. Through this synergistic design, our ATS effectively realizes the IB principle.

3.4 Shared Temporal Attention Encoder (STAE)

To enhance the "student's" ability to learn from the adapted "teacher," we propose the **Shared Temporal Attention Encoder (STAE)**. Its objective is to mitigate the temporal aliasing effects inherent in the RSVP paradigm, as illustrated in Figure 2(C), by learning to apply greater attention to the most informative temporal segments of the EEG signal.

Our STAE builds upon the EEGProject (Wu et al. 2025), a simple yet effective encoder. To reduce the risk of overfitting, STAE learns a single, parameter-efficient, **shared** temporal attention vector $\alpha \in \mathbb{R}^T$—instead of a complex 2D spatio-temporal map—which is then applied uniformly across all channels to reweight the input brain signal $x_b \in \mathbb{R}^{C\times T}$:

$$x_b' = x_b \odot \text{softmax}(\alpha), \quad (4)$$

where C is the number of channels, T is the number of time steps, and $\odot$ denotes element-wise multiplication with broadcasting. This principled design allows the model to suppress noise from aliasing and amplify critical temporal features. Further architectural details of STAE are provided in Appendix A.3.

4 Experiments and Analysis

4.1 Experimental Setup and Main Results

Datasets and Preprocessing. Our experiments are conducted on two large-scale public datasets: **THINGS-EEG** (Gifford et al. 2022) and **THINGS-MEG** (Hebart et al. 2023), with the details shown in Table 1. The EEG dataset contains data from ten participants who underwent a time-efficient RSVP. The training set includes 1654 concepts $\times$ 10 images $\times$ 4 repetitions. The test set includes

Type	Sub	Chan	Train*	Test*	SOA(ms)
EEG	10	63	1654 | 10 | 4	200 | 1 | 80	200 (100)
MEG	4	271	1854 | 12 | 1	200 | 1 | 12	1500±200 (500)

* concepts (classes) | conditions (images) | repetitions (times).

Table 1: Configuration details of the THINGS-EEG and THINGS-MEG datasets (Sub: Subjects, Chan: Channels).

200 concepts $\times$ 1 image $\times$ 80 repetitions. The stimulus onset asynchronies (SOAs) is 200 ms, including a stimulus of 100 ms followed by a 100 ms blank screen. For preprocessing, we epoched EEG data into trials ranging from 0 to 1000 ms after stimulus onset. All EEG repetitions of each image were averaged to ensure a high signal-noise-ratio (SNR). The MEG dataset, including its configuration and preprocessing, follows a similar protocol (Song et al. 2025; Wu et al. 2025).

In our work, THINGS-EEG served as the primary dataset for deriving main results and analyses, while THINGS-MEG was employed to validate our findings and test the generalizability of our approach on a different neuroimaging modality.

Vision Encoders. Our research employs the visual branches of models from OpenCLIP (Ilharco et al. 2021) and DINOv2 (Oquab et al. 2024) as the vision encoder. We conducted extensive experiments across 10 different models (7 from CLIP and 3 from DINOv2). For the detailed comparative analysis within this paper, we focus on two representative encoders from CLIP: RN50 and ViT-L/14. These models were selected to represent different underlying architectures (CNN vs. Transformer) and pre-training data scales (400M vs. 2B pairs), respectively. Unless otherwise specified, RN50 is employed as the default model.

Brain Encoders. For the M/EEG branch, we adopt the Shared Temporal Attention Encoder (STAE) as proposed in Section 3.4. To further evaluate the generalizability of our method, we additionally conducted experiments with alternative architectures, including ShallowNet (Schirrmeister et al. 2017), EEGNet (Lawhern et al. 2018), TSConv (Song et al. 2024) and EEGProject (Wu et al. 2025).

Baselines. We compare our proposed ATS framework against a wide range of recent methods, including BraVL (Du et al. 2023), NICE (Song et al. 2024), ATM-S (Li et al. 2024), CognitionCapturer (Zhang et al. 2025), VE-SDN (Chen et al. 2024) and the previous best-performing method UBP (Wu et al. 2025). The positioning of most methods within the broader research landscape is discussed in Section 2.

Further details on data preprocessing, hyperparameters, hardware configurations and introduction to comparative methods are provided in Appendix A.

Main Results. As shown in Table 2, ATS establishes state-of-the-art performance on the primary THINGS-EEG benchmark, achieving a Top-1 accuracy of **60.2%** and surpassing the previous SOTA (UBP) by a significant margin of **9.8%**. The effectiveness and generalizability of our ATS framework is further validated on the THINGS-MEG dataset, where it achieves a significant **3.0%** Top-1 accuracy improvement. The complete EEG and MEG results for both intra-subject and inter-subject settings are provided in Appendix B.1.

Our method also substantially improves retrieval accuracy when we use the ViT-L/14 model as the vision encoder. However, the overall performance is lower than that achieved with RN50 by a margin of **10.0%**. This finding suggests that an overly powerful "teacher" model may inadvertently widen the asymmetric modality gap, making the adaptive teaching more challenging and thus leading to a significant decrease in performance. This trend is consistently observed across all 10 tested encoders on THINGS-EEG, as detailed in Appendix B.2.

4.2 Validating the Adaptive Teaching Paradigm

Building on the significant overall performance gains of ATS, we now systematically dissect and validate our core contribution: the adaptive teaching paradigm and its implementation via the ShrinkAdapter.

As shown in Figure 5, our proposed **ShrinkAdapter** significantly outperforms both the w/o Adapter baseline and other conventional adapter designs. This overall superior performance highlights that the design of the ShrinkAdapter itself is crucial.

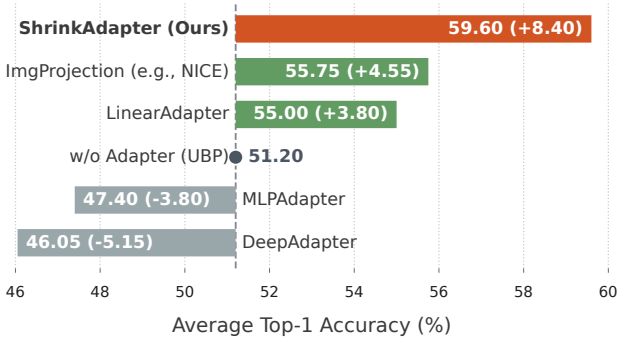

Figure 5: Performance comparison of Adapter architectures on the THINGS-EEG Dataset.

The Necessity of Freedom of Adaptation. Based on our theory of adaptive teaching, we argue that any design choice intended to preserve the teacher's original knowledge is in fact harmful. For an effective alignment, the teacher must be granted the freedom to fully adapt its knowledge structure. We validate this principle from two perspectives: architectural constraints and loss-based constraints.

First, we investigate architectural constraints by ablating the residual connection, a common operation in projection layers (ImgProjection) used to enforce feature preservation. As shown in Table 3, removing the residual connection leads to a consistent and significant improvement across all ShrinkAdapter configurations.

Next, we examine the effect of an explicit loss-based constraint. Inspired by the multimodal similarity-keeping strategy proposed in (Chen and Wei 2024), we introduce a se-

Method	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		Subject 6		Subject 7		Subject 8		Subject 9		Subject 10		Avg	
	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
BraVL	6.1	17.9	4.9	14.9	5.6	17.4	5.0	15.1	4.0	13.4	6.0	18.2	6.5	20.4	8.8	23.7	4.3	14.0	7.0	19.7	5.8	17.5
NICE	13.2	39.5	13.5	40.3	14.5	42.7	20.6	52.7	10.1	31.5	16.5	44.0	17.0	42.1	22.9	56.1	15.4	41.6	17.4	45.8	16.1	43.6
NICE-S	13.3	40.2	12.1	36.1	15.3	39.6	15.9	49.0	9.8	34.4	14.2	42.4	17.9	43.6	18.2	50.2	14.4	38.7	16.0	42.8	14.7	41.7
NICE-G	15.2	40.1	13.9	40.1	15.7	42.7	17.6	48.9	9.0	29.7	16.3	44.4	14.9	43.1	20.3	52.1	14.1	39.7	19.6	46.7	15.6	42.8
MB2C	23.6	56.3	22.6	50.5	26.3	60.1	34.8	67.0	21.3	53.0	31.0	62.3	25.0	54.8	39.0	69.3	27.5	59.3	33.1	70.8	28.4	60.3
ATM-S	25.6	60.4	22.0	54.5	24.0	62.4	31.4	60.9	12.9	43.0	21.4	51.1	30.5	61.5	38.8	72.0	34.4	51.5	29.1	63.5	28.5	60.4
CogCap	31.4	79.6	31.4	77.8	38.1	85.6	40.3	85.8	24.4	66.3	34.8	78.7	34.6	80.9	48.1	88.6	37.4	79.3	35.5	79.2	35.6	80.2
VE-SDN	32.6	63.7	34.4	69.9	38.7	73.5	39.8	72.0	29.4	58.6	34.5	68.8	34.5	68.3	49.3	79.8	39.0	69.6	39.8	75.3	37.2	69.9
UBP	40.5	71.0	49.5	82.5	49.5	82.0	49.5	76.0	45.0	73.0	56.5	83.0	48.5	80.0	57.0	86.0	44.0	76.0	64.0	87.5	50.4	79.7
ATS	**53.0**	**79.0**	**62.0**	**87.5**	**61.5**	**89.0**	**57.0**	**86.5**	**55.0**	**84.0**	**68.0**	**90.5**	**53.0**	**84.0**	**66.5**	**91.0**	**58.5**	**86.0**	**67.5**	**89.0**	**60.2**	**86.7**

Table 2: Top-1 and Top-5 accuracy (%) for 200-way zero-shot retrieval on THINGS-EEG.

mantic distribution consistency loss to our main objective. The loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{SCE}} + \lambda \cdot \left[1 - \text{sim}_{\cos}(\mathbf{M}_{h_v}, \mathbf{M}_{z_v})\right], \quad (5)$$

where $\mathbf{M}_{h_v} = \text{sim}_{\cos}(h_v, h_v)$ is the self-similarity matrix of the original CLIP features, representing their intrinsic semantic distribution. $\mathbf{M}_{z_v}$ is the equivalent for the adapted features. The final cosine similarity term thus measures the consistency of these semantic distributions before and after adaptation, while λ controls the strength of this constraint. Figure 6(B) shows a clear trend: as λ increases (strengthening semantic preservation), Top-1 accuracy drops steadily.

Taken together, these experiments—both architectural and loss-based—strongly support our main claim: for effective asymmetric alignment, the teacher modality must have the freedom to adjust its knowledge structure fundamentally. This validates the core of our Adaptive Teaching Paradigm.

Ratio	Avg. Top-1 Acc (%)			Avg. Top-5 Acc (%)		
	w/ Res	w/o Res	Improv.	w/ Res	w/o Res	Improv.
1:1	54.35	**57.80**	+3.45	83.85	**85.90**	+2.05
1:2	54.45	**58.65**	+4.20	83.45	**86.25**	+2.80
1:4	54.05	**59.60**	+5.55	83.25	**87.55**	+4.30
1:8	53.60	**56.05**	+2.45	82.75	**86.70**	+3.95

Table 3: Ablation study of residual connection in the ShrinkAdapter. The first column denotes the bottleneck compression ratio (e.g., 1:4).

The Necessity of Filtering Redundant Information. Based on our analysis of the asymmetric modality gap, we believe the visual modality is information-redundant relative to the EEG modality. To address this, we introduce a bottleneck structure inspired by the IB principle. We further hypothesize that alignment performance is sensitive to the choice of latent space dimension. To validate these hypotheses, we evaluate two key hyperparameters of our ShrinkAdapter: the bottleneck compression ratio and its output dimension, which determines the size of the latent space.

The results in Figure 6(A) provide strong empirical support for our hypotheses. We observe a clear optimal configuration at a **1/4 bottleneck ratio** and a **768-dimensional latent space**, which achieves the best performance. Notably, excessive compression (e.g., a 1/8 ratio) leads to a sharp drop in accuracy due to over-filtering of essential information, while omitting compression entirely (1/1 ratio) also underperforms the optimal setting. Furthermore, simply increasing the latent space dimension to 2048 does not yield additional gains, indicating that increased capacity does not necessarily translate to better performance. These findings highlight the importance of a well-calibrated bottleneck and confirm the thoroughness of our hyperparameter search.

In summary, the experimental results validate the rationale of the adaptive teaching paradigm and demonstrate the effectiveness of our proposed design.

4.3 Analysis of the Shared Temporal Attention Encoder (STAE)

We validate the effectiveness of STAE by comparing it with established EEG encoders. As shown in Table 4, our proposed encoder consistently outperforms both classic and recent methods.

EEG Encoder	Avg. Top-1	Avg. Top-5
EEGNet	25.65	57.70
ShallowNet	31.30	65.25
TSConv (NICE)	44.85	76.75
EEGProject (UBP)	56.75	84.30
STAE (ours)	**60.20**	**86.65**

Table 4: Comparison of different EEG encoder architectures.

To provide insight into the mechanism behind this success, we visualized the learned attention weights in Figure 6(C). The visualization reveals that STAE automatically learns to focus its processing on the $50-400$ ms window after stimulus onset. This learned temporal focus is both significant and neuroscientifically plausible. The start of this window at $\sim$50 ms aligns with the latency for the initial feedforward sweep of visual information to travel from the retina to the primary visual cortex (Thorpe, Fize, and Marlot 1996). Furthermore, the sustained attention across this ex-

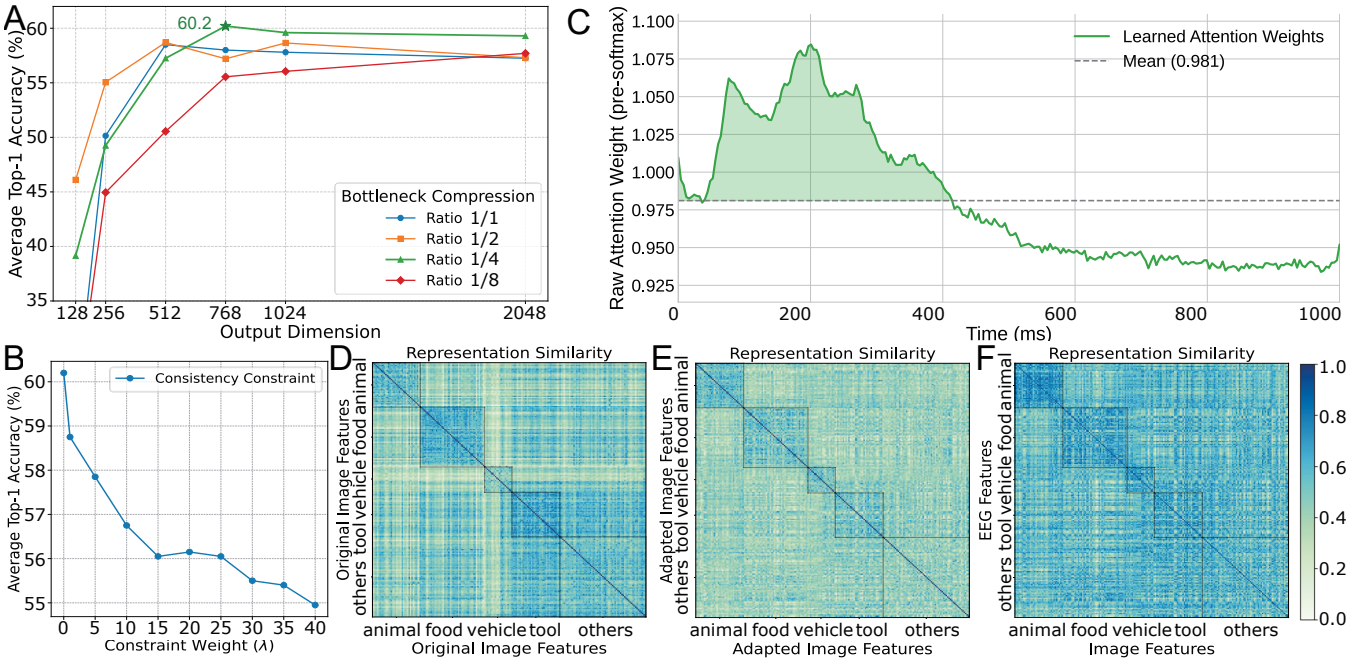

Figure 6: (A) Optimal ShrinkAdapter structure search. (B) Negative impact of consistency constraints. (C) Learned temporal attention weights. (D-F) Representation Similarity Matrices (RSMs) showing: (D) self-similarity of original visual features, (E) self-similarity of adapted visual features, and (F) the cross-modal similarity between EEG and adapted visual features.

tended period is highly consistent with the temporal dynamics of visual object recognition reported in previous studies (Gifford et al. 2022). Ultimately, these qualitative results reveal the mechanism underlying our performance gains and demonstrate the neuroscientific plausibility of our model through consistency with established findings.

4.4 Qualitative Analysis of the ATS

To dissect the mechanisms underlying adaptive teaching and successful retrieval, we employ Representational Similarity Analysis (RSA) (Cichy and Oliva 2020) on the test set, yielding the representational similarity matrices (RSMs) by averaging all subjects. Additionally, we grouped the 200 test concepts into five superordinate categories: animal, food, vehicle, tool, and others, following (Song et al. 2024).

As shown in Figure 6(D), the original image feature h_v, acting as the "teacher," exhibits strong intra-category aggregation (bright diagonal blocks), confirming its highly semantic nature. Notably, it also contains nuanced inter-category similarity, which we consider redundant semantic knowledge that exceeds the "student" modality's capacity. After being refined by our adaptive teaching paradigm, as shown in Figure 6(E), this redundant information is effectively filtered out, while the core conceptual semantics are preserved. This result visually confirms that our adaptive teaching paradigm operates as intended.

A central question remains: is the discrimination we obtain from EEG driven by high-level semantic knowledge or low-level perceptual features (e.g., color, texture, spatial layout)? As shown in Figure 6(F), the bright diagonal blocks in the cross-modal RSM confirm that the features extracted from EEG are semantically meaningful at the category level. Furthermore, the RSM reveals notable off-diagonal inter-class similarity. An analysis of the top-5 retrieval results suggests this is because the model also leverages shared low-level features—such as color, texture, and orientation—for retrieval (Retrieval Case Analysis detailed in Appendix B.3). The bright main diagonal in the cross-modal RSM explains our high 60.2% Top-1 retrieval accuracy. In summary, our adaptive teaching enables the model to extract a hybrid representation from EEG, encoding both conceptual semantics and fundamental perceptual properties.

5 Conclusion

In this work, we reframed the vision-brain alignment problem by identifying its fundamental asymmetry, which we deconstructed into Fidelity Gap and Semantic Gap. To bridge this vision-brain gap, we proposed the Adaptive Teaching Paradigm, a conceptual shift that empowers the visual "teacher" to dynamically shrink and adapt its knowledge to match the capacity of the EEG "student". We implemented this with the ShrinkAdapter, a simple yet effective module whose residual-free, bottleneck design is guided by the Information Bottleneck principle. We also proposed the Shared Temporal Attention Encoder to mitigate temporal noise. Through extensive experiments, we validated the underlying rationale and effectiveness of our paradigm and core components. Our approach not only achieved state-of-the-art performance by a remarkable margin but also provides valuable insights for asymmetric alignment tasks.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants No. U25A20531 and U22A2096. We gratefully acknowledge the High Performance Computing Center of Xidian University for providing computational resources for this work.

References

Chen, C.-S.; and Wei, C.-S. 2024. Mind's Eye: Image Recognition by EEG via Multimodal Similarity-Keeping Contrastive Learning. arXiv:2406.16910.

Chen, H.; He, L.; Liu, Y.; and Yang, L. 2024. Visual Neural Decoding via Improved Visual-EEG Semantic Consistency. arXiv:2408.06788.

Cichy, R. M.; and Oliva, A. 2020. AM/EEG-fMRI fusion primer: resolving human brain responses in space and time. *Neuron*, 107(5): 772–781.

Du, C.; Fu, K.; Li, J.; and He, H. 2023. Decoding Visual Neural Representations by Multimodal Learning of Brain-Visual-Linguistic Features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Gifford, A. T.; Dwivedi, K.; Roig, G.; and Cichy, R. M. 2022. A large and rich EEG dataset for modeling human visual object recognition. *NeuroImage*, 264: 119754.

Grootswagers, T.; Robinson, A. K.; and Carlson, T. A. 2019. The representational dynamics of visual objects in rapid serial visual processing streams. *NeuroImage*, 188: 668–679.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hebart, M. N.; Contier, O.; Teichmann, L.; Rockter, A. H.; Zheng, C. Y.; Kidder, A.; Corriveau, A.; Vaziri-Pashkam, M.; and Baker, C. I. 2023. THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, 12: e82580.

Ilharco, G.; Wortsman, M.; Wightman, R.; Gordon, C.; Carlini, N.; Taori, R.; Dave, A.; Shankar, V.; Namkoong, H.; Miller, J.; Hajishirzi, H.; Farhadi, A.; and Schmidt, L. 2021. OpenCLIP. https://doi.org/10.5281/zenodo.5143773. Accessed: 2025-07-30.

Jiao, Z.; You, H.; Yang, F.; Li, X.; Zhang, H.; and Shen, D. 2019. Decoding EEG by Visual-guided Deep Neural Networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 1387–1393.

Keysers, C.; Xiao, D.-K.; Földiák, P.; and Perrett, D. I. 2001. The speed of sight. *Journal of cognitive neuroscience*, 13(1): 90–101.

Kumar, A.; Raghunathan, A.; Jones, R.; Ma, T.; and Liang, P. 2022. Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution. arXiv:2202.10054.

Lawhern, V. J.; Solon, A. J.; Waytowich, N. R.; Gordon, S. M.; Hung, C. P.; and Lance, B. J. 2018. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of neural engineering*, 15(5): 056013.

Li, D.; Wei, C.; Li, S.; Zou, J.; and Liu, Q. 2024. Visual Decoding and Reconstruction via EEG Embeddings with Guided Diffusion. In *Advances in Neural Information Processing Systems*, volume 37, 102822–102864. Curran Associates, Inc.

Michel, C. M.; and Murray, M. M. 2012. Towards the utilization of EEG as a brain imaging tool. *Neuroimage*, 61(2): 371–385.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Assran, M.; Ballas, N.; Galuba, W.; Howes, R.; Huang, P.-Y.; Li, S.-W.; Misra, I.; Rabbat, M.; Sharma, V.; Synnaeve, G.; Xu, H.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2024. DINOv2: Learning Robust Visual Features without Supervision. arXiv:2304.07193.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ren, Z.; Li, J.; Xue, X.; Li, X.; Yang, F.; Jiao, Z.; and Gao, X. 2021. Reconstructing seen image from brain activity by visually-guided cognitive representation and adversarial learning. *NeuroImage*, 228: 117602.

Schirrmeister, R. T.; Springenberg, J. T.; Fiederer, L. D. J.; Glasstetter, M.; Eggensperger, K.; Tangermann, M.; Hutter, F.; Burgard, W.; and Ball, T. 2017. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human brain mapping*, 38(11): 5391–5420.

Scotti, P.; Banerjee, A.; Goode, J.; Shabalin, S.; Nguyen, A.; Dempster, A.; Verlinde, N.; Yundler, E.; Weisberg, D.; Norman, K.; et al. 2024. Reconstructing the mind's eye: fMRI-to-image with contrastive learning and diffusion priors. *Advances in Neural Information Processing Systems*, 36.

Song, Y.; Liu, B.; Li, X.; Shi, N.; Wang, Y.; and Gao, X. 2024. Decoding Natural Images from EEG for Object Recognition. In *International Conference on Learning Representations*.

Song, Y.; Wang, Y.; He, H.; and Gao, X. 2025. Recognizing Natural Images From EEG With Language-Guided Contrastive Learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Spampinato, C.; Palazzo, S.; Kavasidis, I.; Giordano, D.; Souly, N.; and Shah, M. 2017. Deep learning human mind for automated visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6809–6817.

Takagi, Y.; and Nishimoto, S. 2023. 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*, 14453–14463.

Thorpe, S.; Fize, D.; and Marlot, C. 1996. Speed of processing in the human visual system. *nature*, 381(6582): 520–522.

Tishby, N.; Pereira, F. C.; and Bialek, W. 2000. The information bottleneck method. arXiv:physics/0004057.

van den Oord, A.; Li, Y.; and Vinyals, O. 2019. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748.

Wang, Z.; Li, Y.-L.; Guo, Y.; and Wang, S. 2021. Combating noise: semi-supervised learning by region uncertainty quantification. *Advances in Neural Information Processing Systems*, 34: 9534–9545.

Wei, Y.; Cao, L.; Li, H.; and Dong, Y. 2024. MB2C: Multimodal Bidirectional Cycle Consistency for Learning Robust Visual Neural Representations. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 8992–9000.

Wu, H.; Li, Q.; Zhang, C.; He, Z.; and Ying, X. 2025. Bridging the Vision-Brain Gap with an Uncertainty-Aware Blur Prior. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2246–2257.

Zhang, K.; He, L.; Jiang, X.; Lu, W.; Wang, D.; and Gao, X. 2025. Cognitioncapturer: Decoding visual stimuli from human eeg signal with multimodal information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 14486–14493.