LEARNING EEG REPRESENTATIONS VIA TOPOLOGY-AWARE PRIVILEGED DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio and visual signals offer rich affective cues for emotion recognition, but they are identity-sensitive and raise privacy concerns at deployment. EEG is an appealing inference modality because it does not directly expose facial appearance or voice content, yet its low signal-to-noise ratio often limits recognition performance. We therefore adopt a cross-subject privileged learning setup, where audio and video are used only during training, enabling an EEG-only model at inference. However, existing cross-modal distillation objectives often overlook the region-level interaction structure inherent in EEG, even though affective processing is reflected in inter-regional dependencies. To address this, we propose Topology-Aware Relational Distillation (TRD): the teacher induces a region-dependency graph from audio-visually refined EEG region features, and the student aligns its induced graph by matching the global edge distribution rather than directly imitating features. Meanwhile, the modality gap remains a barrier because EEG representations can still be semantically misaligned with the teacher's multimodal decision space. To mitigate this gap, we further introduce Masked Decision-Space Alignment (MDA). MDA injects student EEG region features into a frozen teacher decision pathway with privileged modalities masked, and distills the resulting predictions so that the student can inherit multimodal semantics without requiring privileged modalities at inference. Under cross-subject EEG-only inference on EAV and MDMER, our method consistently improves the same student backbone trained without distillation, yielding up to +2.09 pp accuracy and +4.46 pp Macro-F1.

1 INTRODUCTION

Emotion recognition is a core problem in affective computing, supporting applications such as human-computer interaction, mental health assessment, and personalized intelligent systems (Muhammad et al., 2026; Li et al., 2025a). Recent progress is largely driven by multimodal learning, where audio and visual streams provide rich and directly observable affective cues and can be effectively combined with physiological signals, including EEG (Lee et al., 2024; Yang et al., 2024; Cai et al., 2023; Liu et al., 2024a). However, a growing body of work has highlighted that common audio-visual cues are identity-sensitive and therefore raise privacy concerns in practical deployment (Li et al., 2025a). In contrast, EEG does not directly expose facial appearance or voice content, making it an appealing sensing channel when identity-sensitive modalities cannot be used at inference time. To be practical under this privacy constraint, EEG-only emotion models should generalize to unseen subjects without subject-specific calibration. At the same time, EEG-based emotion recognition remains challenging due to the low signal-to-noise ratio, artifacts, and strong inter-subject variability, which often lead to weak discriminability and poor generalization across unseen subjects (Muhammad et al., 2026; Ding et al., 2025; Jang et al., 2025; Cai et al., 2023; Liu et al., 2024a). This creates a practical tension: privacy-constrained deployment favors EEG-only inference, while high recognition performance often benefits from audio-visual modalities.

A natural way to reconcile this tension is to treat audio and visual signals as privileged modalities: they are available during training to guide learning but are removed at inference time (Izmailov et al., 2017; Gou, 2021). Cross-modal knowledge distillation provides a principled mechanism for transferring knowledge from a multimodal teacher to a unimodal student under such modality asymmetry, and has been explored in affective and physiological settings (Liu et al., 2023; Wu et al., 2025).

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

However, conventional distillation objectives are typically formulated as soft-label matching or feature imitation (Gou, 2021; Hinton et al., 2015; Romero et al., 2015), which do not explicitly preserve the relational structure of EEG at the scalp-region level. Prior evidence suggests that emotion-related cognitive processing is reflected not only by local activations but also by inter-regional interactions. (Li et al., 2019; Wu et al., 2020; Liu et al., 2019). Meanwhile, in cross-modal distillation, directly forcing an EEG-only student to mimic the teacher's multimodal predictions can be suboptimal when modality gaps and supervision mismatch exist (Jang et al., 2025; Huo et al., 2024; Xue et al., 2023). These observations suggest that effective EEG distillation should transfer both (i) relational structure that is well aligned with EEG sensor-space topography and (ii) semantic compatibility with the teacher's decision space.

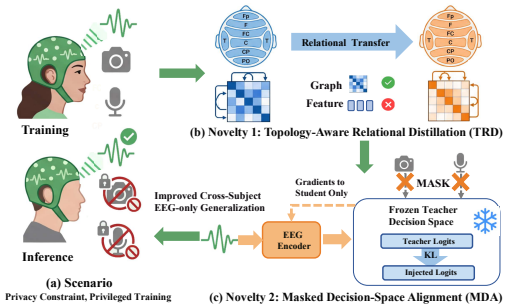


Figure 1: **Motivation and core novelties.** **(a)** Deployment under the privacy constraint requires EEG-only inference on unseen subjects, while audio–visual cues are available only during training as privileged modalities. **(b)** TRD transfers privileged knowledge by aligning the dependency-mass graph defined on a fixed scalp-region partition—thus distilling relations rather than brittle feature values. **(c)** MDA injects the student EEG representation into a frozen teacher decision pathway while masking privileged modalities, enforcing decision-space semantic alignment for EEG-only inference.

Motivated by this, we propose a topology-aware privileged distillation framework that transfers multimodal knowledge to an EEG-only student from two complementary perspectives, as shown in Fig. 1. First, we distill a topology-aware region dependency graph over predefined scalp regions, transferring privileged multimodal effects as inter-region dependency patterns, which is more robust than direct feature imitation. Second, to mitigate the modality gap between EEG-only representations and the teacher's multimodal decision space, we inject student EEG features into a frozen teacher pathway with non-EEG inputs masked, enforcing that EEG alone is semantically sufficient for the teacher's decision function.

The main contributions are threefold:

- We formulate EEG emotion recognition as a privileged cross-modal distillation problem, where auxiliary modalities are available only during training to improve generalization of an EEG-only model under subject shift.
- We develop Topology-Aware Relational Distillation (TRD) to capture and transfer cross-modal inter-region dependencies as structured relational priors, providing invariant supervision across subjects.
- We introduce Masked Decision-Space Alignment (MDA), which performs modality-suppressed alignment in the teacher's decision space, enabling the student to inherit multimodal semantics without access to privileged modalities at inference.

2 Related Works

Multimodal emotion recognition with EEG and external modalities. Multimodal emotion recognition has been extensively studied by combining EEG with external behavioral modalities such as audio and video, since speech and facial cues provide strong affective evidence and multimodal fusion often improves performance over unimodal systems (Muhammad et al., 2026; Lee et al., 2024; Yang et al., 2024; Wang et al., 2022). At the same time, recent work has emphasized that commonly used audio–visual signals may carry identity-sensitive information, raising privacy concerns and motivating privacy-aware or de-identified emotion understanding settings (Li et al., 2025a). Beyond improving multimodal fusion itself, a practical line of research considers modality-constrained deployment and transfers knowledge from multimodal teachers to unimodal students via cross-modal distillation (Liu et al., 2023; Wu et al., 2025). However, existing multimodal and cross-modal approaches largely focus on learning effective multimodal representations or transferring logits/features, while less attention has been paid to transferring multimodally induced EEG

relational structure into an EEG-only model. In contrast, our framework uses audio–visual modalities only during training to preserve privacy and distills multimodal knowledge into EEG through relations and semantics.

Knowledge distillation, graph distillation, and topology-aware EEG modeling. Knowledge distillation has been widely studied for transferring knowledge from a teacher to a student, including response-based (soft-label) distillation (Hinton et al., 2015), feature-based distillation with intermediate hints (Romero et al., 2015), and attention-based feature matching (Zagoruyko & Komodakis, 2017; Ji et al., 2021). Beyond these formulations, relation-based methods argue that knowledge is better represented by structural relations among samples or features, e.g., via distances/angles (Park et al., 2019), pairwise similarity preservation (Tung & Mori, 2019), contrastive objectives (Tian et al., 2019), or probabilistic relation matching (Passalis et al., 2021). A broad taxonomy further includes graph-based distillation, where graphs serve as carriers of relational knowledge or guide knowledge propagation (Gou, 2021). In the EEG domain, topology-aware modeling and functional connectivity analysis provide strong motivation for relational objectives, as emotion-related processing is associated with inter-regional interactions rather than isolated local activations (Li et al., 2019; Liu et al., 2019; Wu et al., 2020). Recent EEG models also incorporate neurophysiological priors and graph structures to improve generalization (Ding et al., 2025; 2024). Meanwhile, cross-modal distillation faces additional challenges such as modality gap and supervision mismatch. Teacher-head based semantic supervision has been proposed to provide decision-space guidance. (Jang et al., 2025; Wang et al., 2024; Ni et al., 2022). Our method differs in that it unifies (i) Topology-Aware Relational Distillation (TRD) over brain regions and (ii) Masked Decision-Space Alignment (MDA) under modality suppression in the cross-modal distillation framework.

3 METHODOLOGY

Motivation and overview. We study emotion recognition under the privacy constraint where identity-sensitive audio–visual signals are available only during training but disallowed at deployment. Accordingly, we aim to train an EEG-only model that generalizes to unseen subjects. This setting raises two challenges: (*i*) strong cross-subject variability makes direct feature imitation brittle, and (*ii*) the EEG-only student may be semantically misaligned with the teacher's multimodal decision space. To address these challenges, we transfer privileged knowledge through two complementary routes: (1) **Topology-Aware Relational Distillation (TRD)**, which transfers multimodally induced inter-region dependency patterns that matches the induced global edge distribution, and (2) **Masked Decision-Space Alignment (MDA)**, which enforces semantic compatibility with the teacher while masking non-EEG inputs. During training, the multimodal teacher uses privileged modalities to *refine* (rather than replace) EEG region features via a gated residual update, producing refined EEG representations that remain EEG-grounded. The student then learns from the teacher via TRD(Sec. 3.2) and MDA (Sec. 3.3). An overview is shown in Fig. 2.

3.1 PROBLEM SETUP

Let $\mathbf{X} \in \mathbb{R}^{C\times T}$ denote an EEG segment with C channels and T time points, and let $\mathbf{a} \in \mathbb{R}^{D_a}$ and $\mathbf{v} \in \mathbb{R}^{D_v}$ denote the corresponding audio and visual features. The emotion label is $y \in \{1, \dots, K\}$. During training, all modalities $(\mathbf{X}, \mathbf{a}, \mathbf{v})$ are available, whereas during inference only $\mathbf{X}$ is accessible. We consider a cross-subject protocol where test subjects are unseen during training.

We use a modality-asymmetric teacher–student framework: the teacher $\mathcal{T}$ is trained with $(\mathbf{X}, \mathbf{a}, \mathbf{v})$, while the student $\mathcal{S}$ only observes $\mathbf{X}$. Following the spirit of local–global region-wise EEG modeling (Ding et al., 2024), we partition EEG sensors into N predefined scalp topographic regions $\{\mathcal{R}_i\}_{i=1}^{N}$ (i.e., electrode groups in sensor space) based on electrode labels in the extended 10–20/10–10 system (Jasper, 1958; Oostenveld & Praamstra, 2001). The teacher and student share the same electrode-to-region mapping and a fixed region order. After EEG encoding and region aggregation via lightweight attention pooling, the teacher and student produce region-level representations:

$$\mathbf{R}^t = [\mathbf{r}_1^t; \dots; \mathbf{r}_N^t], \qquad \mathbf{R}^s = [\mathbf{r}_1^s; \dots; \mathbf{r}_N^s] \in \mathbb{R}^{N\times D}, \tag{1}$$

where D is the region feature dimension.

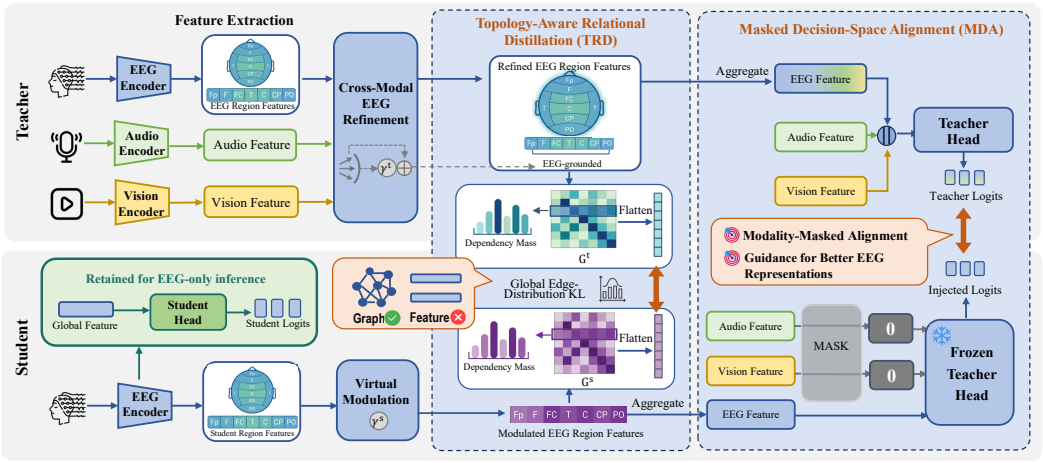


Figure 2: Overview of the proposed topology-aware privileged distillation framework. The multimodal teacher refines EEG region features with privileged audio–visual cues and constructs a region dependency graph over a topology-aware scalp-region partition. The EEG-only student learns from the teacher via (i) TRD that transfers multimodally induced inter-region dependency patterns by aligning the global edge distribution of the graphs rather than directly imitating features, and (ii) MDA that mitigates the modality gap by injecting student virtual-modulated region features into a frozen teacher decision pathway with non-EEG inputs masked. During inference, only the student EEG branch is used.

Teacher refined EEG representation. During training, the teacher further uses privileged modalities to *refine* the EEG regions:

$$\tilde{\mathbf{R}}^t = f_{\text{ref}}(\mathbf{R}^t, \mathbf{a}, \mathbf{v}) \in \mathbb{R}^{N\times D}, \tag{2}$$

where $f_{\text{ref}}(\cdot)$ is a teacher-side refinement function. We intentionally keep f_{ref} lightweight and EEG-grounded, so that privileged modalities serve as training-time guidance rather than dominant predictors. We will use $\tilde{\mathbf{R}}^t$ to construct the teacher graphs (Sec. 3.2) and to produce the teacher fusion targets for decision-space alignment (Sec. 3.3).

3.2 Topology-Aware Relational Distillation (TRD)

To better transfer privileged multimodal effects across subjects, we distill inter-region relations via the graph distribution on the predefined scalp-region topology, rather than directly enforcing strict feature matching.

Graph construction. To define a dependency-graph construction operator $\Phi(\mathbf{Z};\tau)$ for a region feature matrix $\mathbf{Z} = [\mathbf{z}_1;\ldots;\mathbf{z}_N] \in \mathbb{R}^{N\times D}$, we first normalized each region feature by

$$\hat{\mathbf{z}}_i = \frac{\mathrm{LN}(\mathbf{z}_i)}{\|\mathrm{LN}(\mathbf{z}_i)\|_2}, \qquad i = 1,\ldots,N, \tag{3}$$

where $\mathrm{LN}(\cdot)$ denotes layer normalization. The pairwise similarity is then computed by

$$s_{ij} = \hat{\mathbf{z}}_i^\top \hat{\mathbf{z}}_j, \tag{4}$$

and the region dependency graph is obtained via row-wise softmax:

$$\Phi(\mathbf{Z};\tau)_{ij} = \frac{\exp(s_{ij}/\tau)}{\sum_{k=1}^{N}\exp(s_{ik}/\tau)}, \qquad i,j = 1,\ldots,N, \tag{5}$$

where $\tau > 0$ is a temperature parameter. Each row of $\Phi(\mathbf{Z};\tau) \in \mathbb{R}^{N\times N}$ forms a probability distribution over regions; thus, the graph is a row-stochastic (generally asymmetric) dependency matrix, with diagonal entries reflecting self-dependency mass.

Teacher–student graphs. The teacher graph is constructed from the refined EEG representation,

$$\mathbf{G}^t = \Phi(\tilde{\mathbf{R}}^t;\tau_t). \tag{6}$$

and during teacher training, we further apply an auxiliary graph-verification objective to encourage $\mathbf{G}^t$ to be discriminative (see Appendix B.1 for more details). Each entry G_{ij} can be interpreted as a normalized region-to-region dependency mass induced by the similarity of learned region representations, and each row $G_{i,:}$ summarizes how region i distributes its dependency mass over all regions. We use this dependency matrix as a distillation target rather than for value aggregation.

For the student, directly constructing a graph from $\mathbf{R}^s$ would ignore the multimodal effects embedded in $\tilde{\mathbf{R}}^t$. We therefore introduce a virtual modulation mechanism. Specifically, we first compute a global EEG context

$$\mathbf{c}^s = \frac{1}{N}\sum_{i=1}^{N} \mathbf{r}_i^s, \tag{7}$$

and then modulate each student region by

$$\mathbf{r}_i^{s\prime} = \mathbf{r}_i^s + \gamma_i^s \, f_{\mathrm{mod}}([\mathbf{r}_i^s; \mathbf{c}^s]), \qquad i = 1, \ldots, N, \tag{8}$$

where $f_{\mathrm{mod}}(\cdot)$ is a learnable modulation function, $[\cdot;\cdot]$ denotes concatenation, and $\gamma_i^s \in (0, 1)$ is a learnable gate. Denoting the modulated student representation by

$$\mathbf{R}^{s\prime} = [\mathbf{r}_1^{s\prime}; \ldots; \mathbf{r}_N^{s\prime}], \tag{9}$$

the student graph is defined analogously as

$$\mathbf{G}^s = \Phi(\mathbf{R}^{s\prime}; \tau_s). \tag{10}$$

Edge-distribution distillation. $\mathbf{G}$ is row-stochastic by construction. For $\mathbf{G} \in \mathbb{R}^{N\times N}$, we first vectorize it as $\mathrm{vec}(\mathbf{G}) \in \mathbb{R}^{N^2}$ and define an induced edge distribution

$$\mathbf{p}(\mathbf{G}; T_g) = \mathrm{Softmax}\left(\frac{\log(\mathrm{vec}(\mathbf{G}) + \epsilon)}{T_g}\right), \tag{11}$$

where ϵ is a small constant and $T_g > 0$ is the distillation temperature. We then align teacher and student graphs by

$$\mathcal{L}_{\mathrm{graph}} = D_{\mathrm{KL}}\big(\mathbf{p}(\mathbf{G}^t; T_g) \,\|\, \mathbf{p}(\mathbf{G}^s; T_g)\big) \cdot T_g^2. \tag{12}$$

3.3 Masked Decision-Space Alignment (MDA)

Relational transfer alone does not guarantee semantic compatibility with the teacher's multimodal decision space. We therefore inject the student's EEG representation into a frozen teacher pathway while masking non-EEG inputs, enabling the student to inherit multimodal semantics without access to privileged modalities at inference.

Frozen teacher decision pathway. Let f_{agg}^t denote the teacher's region-to-global aggregation module and f_{cls}^t denote the teacher's fusion classifier. The teacher's multimodal prediction is

$$\hat{\mathbf{y}}^t = f_{\mathrm{cls}}^t\Big(f_{\mathrm{agg}}^t(\tilde{\mathbf{R}}^t) \,\oplus\, e_a(\mathbf{a}) \,\oplus\, e_v(\mathbf{v})\Big), \tag{13}$$

where $e_a(\cdot)$ and $e_v(\cdot)$ are the teacher's internal global embeddings for audio and vision, and $\oplus$ denotes concatenation.

Modality-suppressed injection. We inject the student's modulated EEG features $\mathbf{R}^{s\prime}$ into the frozen teacher pathway while masking privileged modalities:

$$\hat{\mathbf{y}}^{\mathrm{inj}} = f_{\mathrm{cls}}^t\big(f_{\mathrm{agg}}^t(\mathbf{R}^{s\prime}) \,\oplus\, \mathbf{0} \,\oplus\, \mathbf{0}\big). \tag{14}$$

During teacher training, $e_a(\mathbf{a})$ and $e_v(\mathbf{v})$ are replaced by $\mathbf{0}$ with a probability, so the masked input in Eq. equation 14 is a missing-modality pattern that f_{cls}^t has already seen, which reduces the missing-modality shift at the frozen classifier (Appendix C.1).

MDA objective. We match the injected prediction to the teacher's multimodal prediction by

$$\mathcal{L}_{\mathrm{inj}} = D_{\mathrm{KL}}\left(\mathrm{Softmax}\left(\frac{\hat{\mathbf{y}}^t}{T_{\mathrm{inj}}}\right) \,\|\, \mathrm{Softmax}\left(\frac{\hat{\mathbf{y}}^{\mathrm{inj}}}{T_{\mathrm{inj}}}\right)\right) \cdot T_{\mathrm{inj}}^2, \tag{15}$$

where T_{inj} is the distillation temperature. Since both f^t_{agg} and f^t_{cls} are frozen, minimizing $\mathcal{L}_{\text{inj}}$ necessarily reshapes the student's representation $\mathbf{R}^{s\prime}$ and propagates supervision to the virtual modulation module and the upstream EEG encoder. This design makes the MDA objective act as a compressed semantic constraint from the teacher decision space.

The final training objective is

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_g \mathcal{L}_{\text{graph}} + \lambda_i \mathcal{L}_{\text{inj}}, \tag{16}$$

where $\mathcal{L}_{\text{cls}}$ denotes the supervised classification loss, $\mathcal{L}_{\text{graph}}$ denotes the TRD loss defined in Eq. equation 12, $\mathcal{L}_{\text{inj}}$ denotes the MDA loss defined in Eq. equation 15, and $\lambda_g, \lambda_i \geq 0$ are balancing coefficients.

4 Experiments

4.1 Experimental Setup

Datasets and protocol. We evaluate on EAV (Lee et al., 2024) (42 subjects) and MDMER (Yang et al., 2024) (73 subjects) under a cross-subject protocol and only EEG is used at inference. EAV is a 5-class task with 30-channel EEG segments of 500 time points and pre-extracted 768-D audio/visual features. For MDMER, we use the processed binary setting (positive vs. negative) by excluding mixed-emotion trials due to their higher ambiguity; each sample contains 18-channel EEG segments of 300 time points, and the privileged modality is the official 768-D video feature. We use fixed subject-level splits for all methods and ablations: EAV uses an 6:2:2 train/val/test ratio, while MDMER uses 8:1:1; the ratios differ only to account for the different subject counts (a larger dataset allows a smaller validation/test fraction without sacrificing evaluation stability).

Implementation details. We instantiate the sensor-space region partition with $N{=}7$ regions for EAV and $N{=}5$ regions for MDMER, using fixed electrode-to-region mappings shared by the teacher and student (Appendix A.5). The student is an EEGNet-style encoder on both datasets; the teacher is EmT-style on EAV and EEGNet-style on MDMER. Models are trained with AdamW (learning rate $5{\times}10^{-4}$, weight decay 10^{-2}) using cosine learning-rate decay with linear warm-up and gradient clipping (max-norm 1.0); batch sizes are 32 (EAV) and 64 (MDMER). We apply lightweight EEG augmentations (Gaussian noise, temporal masking, and channel dropout). All experiments are implemented in PyTorch and run on a single NVIDIA GeForce RTX 5090 GPU.

4.2 Main Results

Comparison with EEG-only baselines. We compare the final student with representative EEG-only baselines under the same cross-subject protocol. As shown in Table 1, our full method achieves the best EEG-only inference performance on both datasets, improving over the strongest baseline by +3.09 pp Acc / +2.68 pp Macro-F1 on EAV and +1.93 pp Acc / +1.29 pp Macro-F1 on MDMER. To control for the student backbone and optimization, we further report *Ours (Cls-only)*, which is trained with EEG supervision only (like prior EEG-only baselines). Compared with it, our full method yields consistent gains. These results confirm that privileged modalities, when transferred through our distillation objectives, improve cross-subject generalization for EEG-only inference. Implementation details of EEG-only baselines are in Appendix A.3.

Comparison with conventional distillation baselines. We further compare with conventional distillation baselines under our teacher–student backbones and protocol. Table 2 shows that conventional KD objectives yield mixed effects: some improve a subset of metrics, but several fail to outperform or even fall below *Cls-only*, indicating that they do not reliably transfer privileged knowledge under modality gap and cross-subject shift. In contrast, our method consistently outperforms both *Cls-only* and all conventional KD variants on both datasets, confirming that the advantage stems from transferring privileged knowledge through EEG-structured relational patterns and modality-suppressed semantic alignment, rather than enforcing rigid, unstructured imitation. Distillation targets for each KD baseline are in Appendix A.4.

Table 1: EEG-only inference performance on EAV and MDMER under the cross-subject setting. *Ours (full)* uses privileged audio–visual modalities only during training. Best results are in **bold**.

Method	Year	EAV		MDMER	
		Acc. (%)	Macro-F1 (%)	Acc. (%)	Macro-F1 (%)
EEGNet (Lawhern et al., 2018)	2018	38.87 ± 1.36	35.53 ± 2.05	53.20 ± 1.14	45.71 ± 7.26
EEGFormer (Chen et al., 2024)	2024	30.87 ± 0.24	27.27± 0.80	51.04 ± 2.63	40.84 ± 10.24
LaBraM (Jiang et al., 2024)	2024	34.85 ± 0.94	33.47 ± 0.87	52.78 ± 0.43	48.25 ± 3.61
iTransformer (Liu et al., 2024b)	2024	29.32 ± 0.64	28.87 ± 0.35	50.86 ± 0.90	49.89 ± 2.28
EMT (Ding et al., 2025)	2025	37.38 ± 1.75	34.66 ± 3.66	51.35 ± 0.47	50.37 ± 1.28
STRFL (Hu et al., 2025)	2025	38.42 ± 1.97	38.86 ± 1.82	53.11 ± 1.43	52.38 ± 2.66
EEG-SCMM (Liu et al., 2025)	2025	38.11 ± 1.01	38.09 ± 1.26	52.13 ± 0.55	49.54 ± 3.75
Ours (Cls-only)	–	40.81 ± 1.73	38.57 ± 3.54	53.04 ± 0.56	49.21 ± 4.64
Ours (full)	–	**41.96 ± 1.41**	**41.54 ± 1.64**	**55.13 ± 1.40**	**53.67 ± 1.31**

Table 2: Comparison with conventional distillation objectives under the same teacher–student backbones, cross-subject protocol, and EEG-only inference. Best results are in **bold**.

Method	Year	EAV		MDMER	
		Acc. (%)	Macro-F1 (%)	Acc. (%)	Macro-F1 (%)
Cls-only	–	40.81 ± 1.73	38.57 ± 3.54	53.04 ± 0.56	49.21 ± 4.64
Soft-label KD (Hinton et al., 2015)	2015	40.09 ± 0.48	36.85 ± 0.55	52.02 ± 2.49	48.88 ± 3.95
Global-feat KD (Romero et al., 2015)	2015	40.13 ± 1.51	37.24 ± 2.92	51.65 ± 0.76	51.16 ± 0.90
RKD (Park et al., 2019)	2019	39.76 ± 1.89	37.64 ± 2.74	52.12 ± 1.60	50.31 ± 2.57
SPKD (Tung & Mori, 2019)	2019	39.56 ± 1.34	36.96 ± 1.74	53.16 ± 0.99	51.72 ± 1.42
CRD (Tian et al., 2019)	2019	38.24 ± 2.72	35.52 ± 2.55	50.41 ± 1.87	48.72 ± 1.03
CDGKD (Wu et al., 2025)	2025	39.85 ± 1.43	36.99 ± 3.02	53.64 ± 1.60	48.30 ± 7.05
AMBOKD (Li et al., 2025b)	2026	40.30 ± 1.35	37.68 ± 3.01	52.01 ± 0.61	43.94 ± 7.51
Ours	–	**41.96 ± 1.41**	**41.54 ± 1.64**	**55.13 ± 1.40**	**53.67 ± 1.31**

4.3 Ablation Study

Effect of TRD and MDA. To evaluate the contribution of each component, we conduct ablation experiments by removing TRD and MDA from the full model. All variants use the same student architecture and training protocol, differing only in the enabled distillation losses. Table 3a shows that adding either TRD or MDA improves both Accuracy and Macro-F1 over the classification-only baseline on EAV and MDMER. TRD delivers the larger standalone gains, and combining TRD with MDA achieves the best overall performance, suggesting the two are complementary.

What to distill from privileged fusion: first-order refined features vs. second-order refined relations. A key design choice of our framework is what form of privileged EEG knowledge to transfer. The teacher produces a multimodally refined region representation $\tilde{\mathbf{R}}^t$ via cross-modal refinement, while the EEG-only student produces a modulated region representation $\mathbf{R}^{s\prime}$ via virtual modulation (Eq. equation 8). Both $\tilde{\mathbf{R}}^t$ and $\mathbf{R}^{s\prime}$ are the immediate precursors for constructing graphs (Eq. equation 6–equation 10). We therefore compare two supervision forms: (*i*) **First-order feature distillation**, which directly aligns $\mathbf{R}^{s\prime}$ to $\tilde{\mathbf{R}}^t$ using a cosine objective; and (*ii*) **Second-order graph distillation** (ours), which aligns the induced graph distributions $\mathbf{G}^s = \Phi(\mathbf{R}^{s\prime}; \tau_s)$ and $\mathbf{G}^t = \Phi(\tilde{\mathbf{R}}^t; \tau_t)$. Table 3b shows that distilling refined graph relations consistently outperforms directly distilling refined EEG features on both datasets. This supports our motivation that, under cross-subject variability, a distributional relational target provides a more transferable target than feature matching, even when both originate from the same privileged teacher refinement.

Hyperparameter sensitivity. Fig. 3 varies one hyperparameter of the full model at a time, with the others held at the reported default. For λ_g, a large weight works best: smaller values fall toward Cls-only, and raising λ_g past the default lowers Macro-F1 only mildly, so topology transfer should be strong while remaining balanced with classification. For λ_i, the curve is non-monotonic and highest at λ_i=0.1; away from this value, the gain is weaker and less consistent. T_g remains effective around its default, whereas T_{inj} has a sharper peak there. Accuracy follows the same trends (Appendix B.2).

Table 3: Ablations on EAV and MDMER. Each cell reports Acc./Macro-F1 (%).

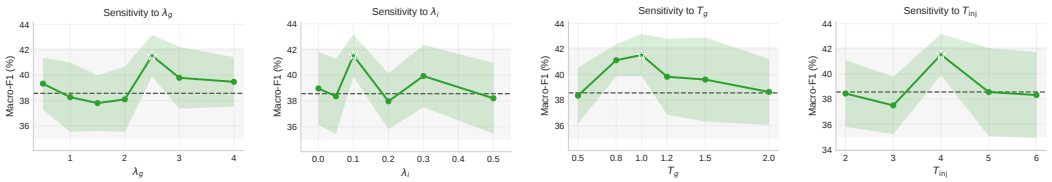
(a) Effect of TRD and MDA.

	EAV		MDMER	
Variant	Acc.	Macro-F1	Acc.	Macro-F1
Cls-only	40.81	38.57	53.04	49.21
+ MDA	41.31	39.98	53.82	52.30
+ TRD	41.64	41.04	54.30	52.93
Full model	**41.96**	**41.54**	**55.13**	**53.67**

(b) What to distill from privileged fusion.

	EAV		MDMER	
Variant	Acc.	Macro-F1	Acc.	Macro-F1
Cls-only	40.81	38.57	53.04	49.21
Feature distill	40.94	39.31	53.21	51.11
Graph distill (Ours)	**41.64**	**41.04**	**54.30**	**52.93**

Feature distill: $\tilde{\mathbf{R}}^t \rightarrow \mathbf{R}^{s\prime}$; Graph distill: $\mathbf{G}^t \rightarrow \mathbf{G}^s$



(a) λ_g (b) λ_i (c) T_g (d) T_{inj}

Figure 3: **Hyperparameter sensitivity on EAV.** Macro-F1 (%) under cross-subject EEG-only inference, averaged over three seeds (mean±std). We vary one hyperparameter at a time while keeping the others fixed to the default setting (star). The dashed line denotes the *Cls-only* baseline, and shaded regions indicate ±1 std.

4.4 Mechanism Analysis

We analyze why Topology-Aware Relational Distillation (TRD) improves cross-subject EEG-only emotion recognition. We focus on two mechanism questions aligned with our graph contribution: (i) whether transferable signal exists beyond a diagonal-dominance shortcut; and (ii) whether the gain reflects teacher-specific relational transfer rather than a generic regularization effect. Fig. 4 provides a qualitative view of the overall transferred off-diagonal dependency pattern, showing that TRD aligns the student with the teacher's edge preferences while suppressing student-only spurious edges to none under the displayed top-k on this subset (per-class maps are provided in Appendix C.2).

Table 4: Mechanism ablations on EAV.

Variant	Acc.	Macro-F1
Cls-only	40.81	38.57
+Diag-masked Graph	41.04	39.95
+Shuffled Graph	40.39	39.28
+Our Graph	**41.64**	**41.04**

4.4.1 Off-diagonal dependencies contribute transferable signal

Purpose (beyond diagonal shortcuts). Since our graph is induced by within-sample region similarity followed by row-wise normalization, diagonal dominance may create a shortcut. We test whether inter-region dependency mass still provides transferable supervision when diagonal terms are removed.

Setup. Inside the TRD loss, we mask the diagonal and renormalize each row:

$$\widetilde{\mathbf{G}} = \mathrm{RowNorm}(\mathbf{G} \odot (\mathbf{J} - \mathbf{I})), \qquad \mathrm{RowNorm}(\mathbf{A})_{ij} = \frac{A_{ij}}{\sum_{k=1}^{N} A_{ik} + \epsilon}, \tag{17}$$

where $\mathbf{I} \in \mathbb{R}^{N\times N}$ is the identity matrix and $\mathbf{J} \in \mathbb{R}^{N\times N}$ is the all-ones matrix. We then apply the same global edge-distribution KL objective (Eq. equation 12) to $\widetilde{\mathbf{G}}^t$ and $\widetilde{\mathbf{G}}^s$. All other training settings are unchanged.

Findings. Diagonal-masked distillation still improves performance over *Cls-only* (Table 4), indicating transferable off-diagonal signal. Notably, keeping the diagonal further improves performance. This suggests that under row-softmax normalization, they encode each region's self-reliance and

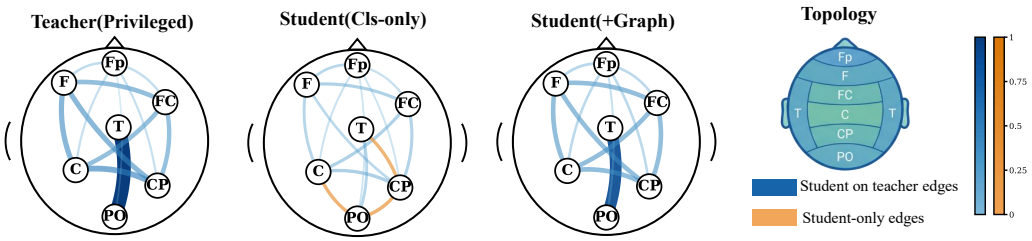


Figure 4: **Topology transfer map on EAV (off-diagonal only).** We visualize inter-region dependency on a fixed set of teacher top-k off-diagonal edges (k=12), averaged over the 50 highest-confidence correctly classified test samples. Blue edges depict the weights on the teacher edge set for each model (Teacher / Student). Orange edges indicate the student's strongest *spurious* connections, i.e., edges in the student top-k set but not in the teacher top-k set (top-3 shown). Width and color intensity indicate normalized strength under a shared mapping (colorbar: 0–1). The topology panel illustrates the predefined scalp-region partition used by both models; the temporal region **T** merges bilateral temporal electrodes and is placed centrally for readability (schematic layout). TRD reduces the per-sample edge-distribution KL to the teacher (0.48→0.21) and increases top-k overlap (63.8%→87.3%). On the averaged map, the +Graph top-k lies within the teacher edge set, so no student-only edges are shown.

the relative scale between self- and inter-region dependency mass, providing additional transferable supervision beyond off-diagonal preferences.

4.4.2 Teacher-specific relational transfer

Purpose (ruling out generic regularization). We test if TRD improves performance as a generic regularizer or relies on matching the correct teacher graph for each sample.

Setup. During training, we randomly permute teacher graphs within each mini-batch:

$$\mathbf{G}_b^t \leftarrow \mathbf{G}_{\pi(b)}^t, \qquad b = 1, \ldots, B, \tag{18}$$

which breaks sample-wise correspondence while keeping the batch-level teacher-graph distribution.

Findings. Shuffling teacher graphs fails to recover the gains of TRD(Table 4), indicating that TRD relies on teacher-specific, sample-wise relational supervision rather than generic regularization.

5 Conclusion

This work advances emotion recognition under the privacy constraint by leveraging identity-sensitive audio–visual signals as privileged training cues that are unavailable at deployment. We introduce a topology-aware privileged distillation framework that improves EEG-only cross-subject generalization via two mechanisms: (i) Topology-Aware Relational Distillation (TRD), which transfers multimodally induced inter-region dependency patterns by aligning teacher and student graphs, and (ii) Masked Decision-Space Alignment (MDA), which injects student EEG features into the frozen teacher decision pathway while masking non-EEG inputs, thereby enforcing semantic compatibility with the teacher's multimodal decision space. Empirically, our EEG-only student consistently outperforms the same backbone trained without distillation, improving accuracy by +1.15 pp and Macro-F1 by +2.97 pp on EAV, and by +2.09 pp and +4.46 pp on MDMER, respectively. Future work will study topology-preserving temporal extensions that retain our region-level spatial inductive bias while introducing lightweight time-window saliency to distill time-resolved inter-region interactions, better capturing short-term affective continuity.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation and to polish the writing. We did not use them to generate data, develop theory or proofs, propose hypotheses, design or implement the method, or interpret the experimental results. We have reviewed all AI-assisted text. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Experiments use the public EAV and MDMER datasets under their original release terms, and we rely on the consent and ethics approval of those source studies. We do not release new recordings or identity-linked audio–visual data.

REPRODUCIBILITY STATEMENT

Dataset definitions, the cross-subject splits, and the optimization setup are given in Section 4.1. Appendix A specifies the region-level interface, student hyperparameters, EEG-only baseline settings, the distillation target of each baseline, and the electrode-to-region maps. All main comparisons use the same three seeds.

REFERENCES

Yujian Cai, Xingguang Li, and Jinsong Li. Emotion recognition using different sensors, emotion models, methods and datasets: A comprehensive review. *Sensors*, 23(5):2455, 2023. doi: 10.3390/s23052455.

Yuqi Chen, Kan Ren, Kaitao Song, Yansen Wang, Yifan Wang, Dongsheng Li, and Lili Qiu. Eegformer: Towards transferable and explainable large-scale eeg foundation model. *ArXiv*, abs/2401.10278, 2024.

Yi Ding, Neethu Robinson, Chengxuan Tong, Qiuhao Zeng, and Cuntai Guan. Lggnet: Learning from local-global-graph representations for brain–computer interface. *IEEE Transactions on Neural Networks and Learning Systems*, 35(7):9773–9786, 2024. doi: 10.1109/TNNLS.2023.3236635.

Yi Ding, Chengxuan Tong, Shuailei Zhang, Muyun Jiang, Yong Li, Kevin Lim Jun Liang, and Cuntai Guan. Emt: A novel transformer for generalized cross-subject eeg emotion recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6):10381–10393, 2025. doi: 10.1109/TNNLS.2025.3552603.

Jianping Gou. Knowledge distillation: A survey. *International Journal of Computer Vision*, 2021.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *arXiv*, abs/1503.02531, 2015.

Fo Hu, Kailun He, Can Wang, Qinxu Zheng, Bin Zhou, Gang Li, and Yu Sun. Strflnet: Spatio-temporal representation fusion learning network for eeg-based emotion recognition. *IEEE Transactions on Affective Computing*, 2025. doi: 10.1109/TAFFC.2025.3611173.

Fushuo Huo, Wenchao Xu, Jingcai Guo, Haozhao Wang, and Song Guo. C^2KD: Bridging the modality gap for cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16006–16015, 2024. doi: 10.1109/CVPR52733.2024.01515.

Rauf Izmailov, Blerta Lindqvist, and Peter Lin. Feature selection in learning using privileged information. In *Proceedings of the IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 957–963, 2017. doi: 10.1109/ICDMW.2017.131.

Hyo-Jeong Jang, Hye-Bin Shin, and Seong-Whan Lee. Uncertainty-aware cross-modal knowledge distillation with prototype learning for multimodal brain-computer interfaces. arXiv:2507.13092, 2025.

Herbert H. Jasper. The ten twenty electrode system of the international federation. *Electroencephalography and Clinical Neurophysiology*, 10(2):371–375, 1958.

Mingi Ji, Byeongho Heo, and Sungrae Park. Show, attend and distill: Knowledge distillation via attention-based feature matching. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pp. 7945–7952, 2021. doi: 10.1609/aaai.v35i9.16969.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data (labram). In *The Twelfth International Conference on Learning Representations*, 2024.

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. Eegnet: A compact convolutional network for eeg-based brain-computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018. doi: 10.1088/1741-2552/aace8c.

Min-Ho Lee, Adai Shomanov, Balgyn Begim, Zhuldyz Kabidenova, Aruna Nyssanbay, Adnan Yazici, and Seong-Whan Lee. Eav: Eeg-audio-video dataset for emotion recognition in conversational contexts. *Scientific Data*, 11(1):1026, 2024. doi: 10.1038/s41597-024-03838-4.

Deng Li, Bohao Xing, Xin Liu, Baiqiang Xia, Bihan Wen, and Heikki Kälviäinen. Deemo: Deidentity multimodal emotion recognition and reasoning. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pp. 5707–5716, 2025a. doi: 10.1145/3746027.3755411.

Peiyang Li, Huan Liu, Yajing Si, Cunbo Li, Fali Li, Xuyang Zhu, Xiaoye Huang, Ying Zeng, Dezhong Yao, Yangsong Zhang, and Peng Xu. Eeg-based emotion recognition by combining functional connectivity network and local activations. *IEEE Transactions on Biomedical Engineering*, 66(10):2869–2881, 2019. doi: 10.1109/TBME.2019.2897651.

Zixing Li, Chao Yan, Zhen Lan, Xiaojia Xiang, Han Zhou, Jun Lai, and Dengqing Tang. Adaptive modality balanced online knowledge distillation for brain–eye–computer-based dim object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2025b.

Huan Liu, Tianyu Lou, Yuzhe Zhang, Yixiao Wu, Yang Xiao, Christian S. Jensen, and Dalin Zhang. Eeg-based multimodal emotion recognition: A machine learning perspective. *IEEE Transactions on Instrumentation and Measurement*, 73:1–29, 2024a. doi: 10.1109/TIM.2024.3369130.

Qile Liu, Weishan Ye, Lingli Zhang, and Zhen Liang. Eeg-scmm: Soft contrastive masked modeling for cross-corpus eeg-based emotion recognition. arXiv:2408.09186, 2025.

Xucheng Liu, Ting Li, Cong Tang, Tao Xu, Peng Chen, Anastasios Bezerianos, and Hongtao Wang. Emotion recognition and dynamic functional connectivity analysis based on eeg. *IEEE Access*, 7: 143293–143302, 2019. doi: 10.1109/ACCESS.2019.2945059.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. arXiv:2310.06625, 2024b.

Yucheng Liu, Ziyu Jia, and Haichao Wang. Emotionkd: A cross-modal knowledge distillation framework for emotion recognition based on physiological signals. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pp. 6122–6131, 2023. doi: 10.1145/3581783.3612277.

Ghulam Muhammad, Sumayah Almuntasheri, Fadia Alenezi, Nwraan Alhadi, and Victor C. M. Leung. Eeg-based multimodal emotion recognition: Recent progress, challenges, and future directions. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 22 (2):1–28, 2026. doi: 10.1145/3774428.

Jianyuan Ni, Raunak Sarbajna, Yang Liu, Anne H. H. Ngu, and Yan Yan. Cross-modal knowledge distillation for vision-to-sensor action recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4448–4452, 2022.

Robert Oostenveld and Peter Praamstra. The five percent electrode system for high-resolution eeg and erp measurements. *Clinical Neurophysiology*, 112(4):713–719, 2001. doi: 10.1016/S1388-2457(00)00527-7.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3962–3971, 2019. doi: 10.1109/CVPR.2019.00409.

Nikolaos Passalis, Maria Tzelepi, and Anastasios Tefas. Probabilistic knowledge transfer for lightweight deep representation learning. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5):2030–2039, 2021. doi: 10.1109/TNNLS.2020.2995884.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. arXiv:1412.6550, 2015.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. arXiv:1910.10699, 2019.

Fred Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1365–1374, 2019. doi: 10.1109/ICCV.2019.00145.

Jiabao Wang, Yuming Chen, Zhaohui Zheng, Xiang Li, Ming-Ming Cheng, and Qibin Hou. Crosskd: Cross-head knowledge distillation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Qian Wang, Mou Wang, Yan Yang, and Xiaolei Zhang. Multi-modal emotion recognition using EEG and speech signals. *Computers in Biology and Medicine*, 149:105907, 2022. doi: 10.1016/j.compbiomed.2022.105907.

Shuang Wu, Heng Liang, Yong Zhang, Yanlin Chen, and Ziyu Jia. A cross-modal densely guided knowledge distillation based on modality rebalancing strategy for enhanced unimodal emotion recognition, 2025.

Xun Wu, Wei-Long Zheng, and Bao-Liang Lu. Investigating eeg-based functional connectivity patterns for multimodal emotion recognition. arXiv:2004.01973, 2020.

Zihui Xue, Zhengqi Gao, Sucheng Ren, and Hang Zhao. The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation. In *International Conference on Learning Representations (ICLR)*, 2023.

Pei Yang, Niqi Liu, Xinge Liu, Yezhi Shu, Wenqi Ji, Ziqi Ren, Jenny Sheng, Minjing Yu, Ran Yi, Dan Zhang, and Yong-Jin Liu. A multimodal dataset for mixed emotion recognition. *Scientific Data*, 11(1):847, 2024. doi: 10.1038/s41597-024-03676-4.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv:1612.03928, 2017.

A Model and Implementation Details

This appendix documents the architectural details and default hyperparameters used in our experiments to facilitate reproducibility and fair comparison.

A.1 Supplementary Architecture Notes

Backbone-agnostic region-level EEG interface. What is essential is that the EEG encoder is adapted to expose a unified *region-level* interface shared by teacher and student:

$$(\mathbf{R}, \mathbf{g}) = f_{\text{eeg}}(\mathbf{X}), \qquad \mathbf{R} \in \mathbb{R}^{N \times D},\ \mathbf{g} \in \mathbb{R}^{d_g}, \tag{19}$$

where $\mathbf{R} = [\mathbf{r}_1; \ldots; \mathbf{r}_N]$ follows a fixed sensor-space region partition and region order shared by both models (Appendix A.5), and $\mathbf{g}$ is a global EEG feature.

Common EEG-only inference pathway. All student variants perform EEG-only inference through the global feature:

$$\hat{\mathbf{y}}^s = h^s_{\text{cls}}(\mathbf{g}^s), \tag{20}$$

and deployment uses only this EEG-only student branch (no privileged modalities).

A.2 Default Hyperparameters

Table A.1 summarizes the default hyperparameters used for training the EEG-only student in the main experiments. Unless otherwise stated, we use AdamW with cosine learning-rate decay and lightweight EEG augmentations.

Table A.1: Default hyperparameter settings (student). Values correspond to the configuration used for the models reported in the main results.

Hyperparameter	**EAV**	**MDMER**
EEG input (channels × time)	30×500	18×300
Regions	7	5
Classes	5	2
Batch size	32	64
Training epochs (max)	150	120
Early-stop patience	8	8
Optimizer	AdamW	AdamW
Learning rate	5×10^{-4}	5×10^{-4}
Weight decay	1×10^{-2}	1×10^{-2}
Gradient clipping (max-norm)	1.0	1.0
LR schedule / warm-up	cosine / 5 epochs	cosine / 5 epochs
Min LR ($\eta_{\min}$)	1×10^{-5}	1×10^{-5}
Dropout	0.35	0.35
Classification loss	focal + label smoothing	focal + label smoothing
Focal γ	2.0	2.0
Label smoothing	0.10	0.15
EEG augmentation: Gaussian noise std	0.02	0.02
EEG augmentation: time mask prob	0.10	0.10
EEG augmentation: channel dropout prob	0.05	0.05

A.3 EEG-only Baseline Hyperparameters

For every baseline, architecture/loss HPs follow the official implementation / original paper; we then lightly adapt a few optimization HPs (learning rate, batch size, dropout, early-stopping) to each dataset, with model selection always performed on that dataset's own validation split. Method-specific key settings are summarized in Table A.2.

Table A.2: EEG-only baseline hyperparameters. Official recipes are the starting point; remaining entries are validation-selected.

Method	EAV	MDMER
EEGNet	F_1=8, D=2, dropout 0.5 wd 10^{-2}, bs 32, ≤150 ep, patience 20	F_1=16, SAME-pad kernel 64 lr 5×10^{-4}, bs 64, ≤120 ep patience 8, focal + LS 0.15
EEGFormer	patch 25, dim 128, depth 4 heads 8, dropout 0.3	dim 64, heads 4, depth 2+2 patch 10; otherwise as EEGNet
EmT	graph + transformer topology prior, dropout 0.35	18-ch / 300 Hz, 6-region topology binary head; otherwise as EAV
iTransformer	d_{model}=128, heads 4, layers 2 d_{ff}=256, dropout 0.1 patch 100/50, bs 64, 50 ep	same as EAV
LaBraM	finetune labram-base lr 10^{-4}, 100 ep, warmup 10, wd 0.05	finetune labram-base lr 5×10^{-4}, 50 ep, warmup 5 wd 0.05, patience 15, patch 200
STRFLNet	embed 64, ST-ODE 4, dropout 0.3 wd 10^{-3}, LS 0.1, + PLI	embed 32, ST-ODE 2, dropout 0.5 lr 5×10^{-4}, 100 ep, LS 0.10, + PLI
EEG-SCMM	5-band DE; pretrain 100 ep (lr 5×10^{-4}) + finetune 50 ep (lr 10^{-4}), bs 128, mask 0.4	DE, 18-ch; pretrain 100 ep + finetune 80 ep, patience 15

Table A.3: What each KD baseline aligns.

Method	Student signal	Teacher signal	Objective
Soft-label	logits	fusion logits	temp. KL
Global-feat	global feature	EAV: fusion input MDMER: EEG global feature	cosine (+proj. if needed)
RKD	global feature	EAV: fusion input MDMER: EEG global feature	distance + angle
SPKD	global feature	same as RKD	batch similarity MSE
CRD	global feature	fusion	in-batch contrastive
CDGKD	projected features + logits	refined region features + fusion logits	MMD + DKD
AMBOKD	logits	fusion / audio / vision logits	offline KL + adaptive OKD
Ours	co-response graph + inject path	teacher graph / fusion logits	TRD + MDA

A.4 Conventional KD Baselines

All methods in Table 2 share the same the identical teacher–student backbones, data splits, optimization protocol, and EEG-only inference; the only controlled factor is the distillation objective. Privileged modalities are used only inside the frozen teacher to produce supervision (fusion logits or teacher features); the student always consumes EEG alone. We adapt AMBOKD to this fair setting by removing student-side A/V heads and transferring only from teacher modality logits. Table A.3 lists the student signal, teacher signal, and objective aligned by each method.

Table A.4: Electrode-to-region mappings used in our experiments. Region orders are fixed and shared by the teacher and student.

Region	Electrodes
EAV (30 channels)	
Fp	Fp1, Fp2
F	F7, F3, Fz, F4, F8
FC	FC5, FC1, FC2, FC6
T	T7, T8
C	C3, Cz, C4
CP	CP5, CP1, CP2, CP6
PO	P7, P3, Pz, P4, P8, PO9, O1, Oz, O2, PO10
MDMER (18 channels)	
Fp	Fp1, Fp2
F	F3, Fz, F4, F7, F8
C	C3, Cz, C4
T	T3, T4, T5, T6
PO	P3, P4, O1, O2

Table B.1: **Privileged teacher fusion vs. EEG-only student (test set).** Each cell reports **Acc / Macro-F1** (%).

Dataset	Student (Cls-only, EEG-only)	Teacher (Fusion, privileged)
EAV	40.81 / 38.57	51.25 / 50.31
MDMER	53.04 / 49.21	55.72 / 54.58

A.5 Sensor-space Region Topology (Electrode-to-Region Mapping)

Scope. We define a coarse *sensor-space* topographic prior by grouping electrodes into a small set of predefined *scalp topographic regions* based on electrode labels in the extended 10–20/10–10 system (Jasper, 1958; Oostenveld & Praamstra, 2001). This "region" definition refers to electrode groups in sensor space rather than anatomical brain parcels. The teacher and student share the *same* electrode-to-region mapping and a fixed region order, given in Table A.4.

Relation to prior work. Our region-wise grouping follows the spirit of local–global EEG graph modeling (e.g., LGGNet Ding et al. (2024)), with dataset-specific adaptations to the available montages.

B Supplementary Experimental Results

This section provides supplementary results referenced by the main text, including teacher diagnostics and robustness checks.

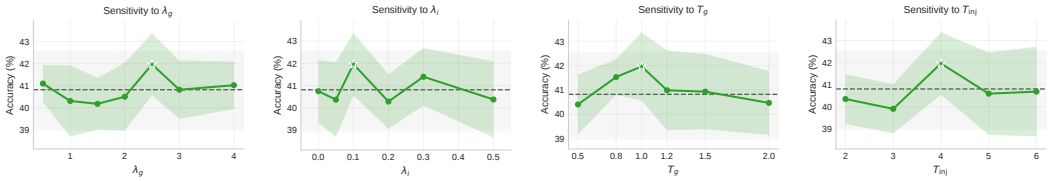
B.1 Teacher Diagnostics for Privileged Supervision

Privileged teacher as an upper bound. Table B.1 compares the privileged teacher fusion branch with the EEG-only *Cls-only* student under the same cross-subject protocol. The fusion teacher consistently provides a stronger supervisory signal, motivating privileged distillation.

Teacher graph-verification branch (label-supervised diagnostic). To avoid a trivial or unstable dependency structure during teacher training, we optimize an auxiliary graph-verification classifier on a graph-guided representation with the ground-truth labels. In addition, the teacher includes an *EEG-only* classification head attached to the EEG encoder, trained with label supervision to keep EEG representations discriminative and EEG-grounded. Table B.2 reports branch-level test performance as a diagnostic that the teacher's graph-guided pathway is label-informed. On EAV, the graph-guided branch is notably stronger than the teacher EEG-only branch, suggesting that privileged refinement can induce a discriminative relational pathway. On MDMER, the graph-guided

Table B.2: **Teacher branch diagnostics (test set).** Each cell reports **Acc / Macro-F1** (%).

Dataset	Teacher EEG-only	Teacher Graph-guided	Teacher Fusion
EAV	32.33 / 32.13	41.11 / 40.38	51.25 / 50.31
MDMER	52.35 / 45.97	52.12 / 43.17	55.72 / 54.58



(a) λ_g (b) λ_i (c) T_g (d) T_{inj}

Figure B.1: **Hyperparameter sensitivity on EAV (accuracy counterparts).** Accuracy (%) under cross-subject EEG-only inference, averaged over three seeds (mean±std). We vary one hyperparameter at a time while keeping the others fixed to the default setting (star). The dashed line denotes the *Cls-only* baseline, and shaded regions indicate ±1 std.

branch is weaker but remains non-trivial; this is consistent with MDMER using a single privileged modality (vision only) and a binary cross-subject setting. Importantly, the student improvements are driven by distilling the teacher-induced relations (TRD) and decision-space alignment (MDA), rather than requiring the teacher graph-verification classifier itself to be the best-performing branch (Sec. 4.3).

B.2 Accuracy Counterparts for Hyperparameter Sensitivity

Fig. B.1 reports accuracy for the same full-model sweeps as Fig. 3. The reported default is the highest mean on each slice. For λ_g, smaller weights fall toward Cls-only, and weights above 2.5 move back toward it, so topology transfer uses a large graph weight without being increased without bound. For λ_i, accuracy is non-monotonic and highest at λ_i=0.1, above turning injection off; away from this value, the gain is weaker and less consistent. T_g stays above Cls-only from 0.8 to 1.5. T_{inj} is searched on a coarse grid and is highest at the reported default, so neither temperature is finely tuned.

C Supplementary Analyses

We provide additional analyses to interpret the behavior of MDA and TRD.

C.1 In-distribution Masked Interface for MDA

Purpose. MDA queries a frozen fusion classifier with privileged modalities replaced by zeros (Eq. equation 14). If this missing-modality pattern is never seen during teacher training, the zero-padded query can induce a substantial distribution shift inside the frozen head, limiting the effectiveness of semantic alignment. We therefore train the teacher with modality dropout so that the masked input already lies in the support of f^t_{cls}.

Setup. During multimodal teacher training, privileged audio and visual embeddings are replaced by zeros with probability 0.5, while the fusion classifier remains under label supervision. MDA then feeds the student EEG representation into the frozen classifier under the same zero-masked audio–visual input (Eq. equation 14). To test whether this dropout is necessary, we retrain the EAV teacher without it (same architecture and protocol otherwise), freeze the teacher, and train a student with MDA only, using the same cross-subject split and three seeds.

Findings. Removing modality dropout degrades the MDA-only student (Table C.1) to about the *Cls-only* level, indicating that semantic alignment holds only when the zero-masked query remains in-distribution for the frozen fusion classifier.

Table C.1: **Modality-dropout ablation for MDA** (EAV, inject-only student, 3 seeds). The default row matches + *MDA* in Table 3a.

Teacher training	Acc. (%)	Macro-F1 (%)
w/ modality dropout (p=0.5; default)	41.31 ± 1.16	39.98 ± 2.29
w/o modality dropout	40.49 ± 1.39	38.27 ± 3.23
Δ (w/ − w/o)	**+0.82**	**+1.71**

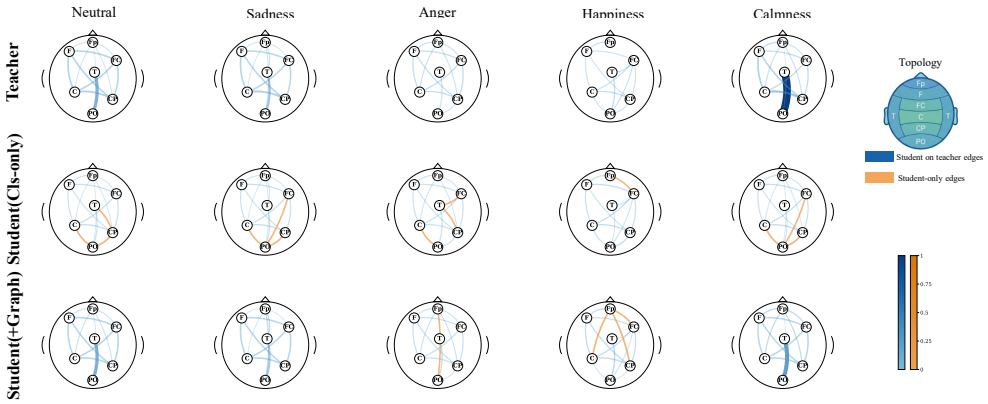


Figure C.1: **Per-class topology transfer maps on EAV (off-diagonal only).** Columns correspond to the 5 emotion classes; rows correspond to Teacher(Privileged), Student(*Cls-only*), and Student(+*Graph*). For each class, the teacher reference set is the class-specific top-k undirected off-diagonal edges (k=12) computed from the mean teacher graph over n=50 teacher-correct high-confidence test samples. Blue edges show each model's weights restricted to the teacher edge set; orange edges show the student's strongest student-only edges outside the teacher set (top-3 maximum).

C.2 Class-Conditional Topology Transfer Maps

Protocol. For each class c, we select n=50 test samples that are (i) correctly classified by the teacher and (ii) have the highest teacher confidence within that class. We compute the mean teacher graph $\bar{\mathbf{G}}^t$, the mean graph-student graph $\bar{\mathbf{G}}^s$, and the mean *Cls-only* baseline graph $\bar{\mathbf{G}}^{\text{cls}}$ (all graphs are induced by Eq. equation 5). For top-k visualization/overlap, we rank symmetrized off-diagonal edges using $\bar{W}_{ij} = \frac{1}{2}(\bar{G}_{ij} + \bar{G}_{ji})$ (diagonal removed). We define the class-specific teacher reference set as the top-k edges from $\bar{\mathbf{W}}^t$ with k=12. In each column of Fig. C.1, blue edges show a model's weights restricted to this teacher reference set, and orange edges show the student's strongest *student-only* edges (up to top-3) among the student's own top-k edges that are not in the teacher set. This subset is used only for diagnostic visualization and does not affect training or reported test metrics.

Quantitative summary. Table C.2 reports (i) top-k undirected edge overlap with the class-specific teacher reference set, (ii) the mean weight of the displayed student-only edges (lower is better), and (iii) the global edge-distribution KL used by TRD (Eq. equation 12) computed on the mean directed graphs (and its diagonal-masked version). Overall, +*Graph* reduces the divergence to the teacher (mean KL 0.159→0.038; mean off-diagonal KL 0.185→0.039) and suppresses student-only edges (mean 0.145→0.058), consistent with the overall mechanism analysis. One class (Happiness) shows limited improvement under this subset; notably, *Cls-only* already achieves near-zero divergence to the teacher for this class (KL ≈ 0), so the remaining discrepancy is dominated by small ranking perturbations around the top-k cutoff.

Table C.2: **Per-class topology transfer metrics on EAV** (same subset and k=12 as Fig. C.1). **Overlap**: undirected off-diagonal top-k overlap between the student's top-k edges and the class-specific teacher reference set (computed on $\bar{\mathbf{W}}$). **Spurious**: mean weight of the displayed student-only edges (up to top-3; fewer if less exist outside the teacher set). **KL**: global edge-distribution KL (Eq. equation 12) computed on mean *directed* graphs $\bar{\mathbf{G}}$. $\mathbf{KL_{off}}$: the same KL after diagonal masking and row renormalization (Eq. equation 17) *for analysis only*.

Class	Overlap↑ (Cls)	Overlap↑ (+Graph)	Spurious↓ (Cls)	Spurious↓ (+Graph)	KL↓ (Cls)	KL↓ (+Graph)	KL_{off}↓ (Cls)	KL_{off}↓ (+Graph)
Neutral	0.500	**1.000**	0.151	**0.000**	0.243	**0.036**	0.252	**0.026**
Sadness	0.583	**1.000**	0.151	**0.000**	0.148	**0.024**	0.151	**0.017**
Anger	0.667	**0.917**	0.144	**0.136**	0.022	**0.011**	0.019	**0.009**
Happiness	0.917	0.750	0.128	0.154	0.009	0.029	0.009	0.030
Calmness	0.667	**1.000**	0.152	**0.000**	0.372	**0.090**	0.495	**0.114**
Mean	0.667	**0.933**	0.145	**0.058**	0.159	**0.038**	0.185	**0.039**

C.3 Class-wise Performance Balance of the Student

We report Macro-F1 together with class-wise recall for the EEG-only student trained with *Cls-only* supervision and with our full method, averaged over the same three seeds as the main results, to characterize class balance and localize the source of improvements. For EAV, the test set is class-balanced, so the mean recall equals overall accuracy; for MDMER (binary, class-imbalanced), we additionally summarize the mean recall (balanced accuracy) for completeness.

EAV (5-class). Table C.3 shows that *Cls-only* under-predicts Sadness (C1) and Calmness (C4), while over-predicting Neutral (C0). Our method lifts the two weakest classes (C1: $22.04 \rightarrow 36.06$; C4: $18.98 \rightarrow 24.21$) and reduces Neutral dominance ($61.02 \rightarrow 49.03$), yielding a more even recall profile (std: $17.13 \rightarrow 11.11$; min: $18.98 \rightarrow 24.21$). This redistribution explains why Macro-F1 improves more than overall accuracy.

Table C.3: **EAV student class-wise recall (test set)** (%; mean over 3 seeds). C0–C4 correspond to {Neutral, Sadness, Anger, Happiness, Calmness}. Std/Min/Range are computed over the five class recalls.

Method	C0	C1	C2	C3	C4	Std↓	Min↑	Range↓
Student (Cls-only)	61.02	22.04	54.26	47.78	18.98	17.13	18.98	42.04
Student (Ours)	49.03	36.06	56.57	43.94	24.21	11.11	24.21	32.36

MDMER (binary). As shown in Table C.4, *Cls-only* is strongly biased toward the positive class. Our method raises negative-class recall ($29.99 \rightarrow 39.61$) at a moderate cost in positive-class recall ($74.45 \rightarrow 69.54$), improving mean recall (balanced accuracy) and Macro-F1.

Table C.4: **MDMER student class-wise recall (test set)** (%; mean over 3 seeds). Mean recall equals balanced accuracy for binary classification.

Method	Positive (C0)	Negative (C1)	Mean Recall↑	Macro-F1↑
Student (Cls-only)	74.45	29.99	52.22	49.21
Student (Ours)	69.54	39.61	54.57	53.67

D Limitations

Montage dependence. The region interface assumes a labeled 10–20/10–10 montage, so that electrodes can be assigned to topographic regions. On unlabeled, non-standard, or extremely sparse consumer/BCI caps, position-based assignment may fail or yield overly coarse regions, and deployment is therefore not plug-and-play across devices. Future work will evaluate sparse but labeled montages and recover or learn spatial assignment when standard labels are unavailable, while retaining the topographic prior.

Static, segment-level graphs. TRD constructs one dependency graph per window and may miss short-timescale affective dynamics. Future work will study topology-preserving temporal extensions with lightweight window saliency, as noted in the conclusion.

Distilled graph as a transfer target. The dependency-mass graph is induced from learned region representations and is used as a structured distillation target, rather than a direct estimate of neurophysiological connectivity. Accordingly, our contributions concern privileged transfer and cross-subject EEG-only generalization, not mechanistic brain interpretation. Future work will incorporate stronger neurophysiological priors without sacrificing these goals.