PAIRWISE MODALITY EVIDENCE TRANSFER FOR GENERALIZED ZERO-SHOT EMOTION RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generalized zero-shot emotion recognition requires identifying both seen and unseen emotions using task-specific labeled examples only from seen categories. Multimodal observations provide complementary evidence for this task, but audio, text and visual predictors may favor different candidates. Moreover, a correction learned from errors on familiar categories may not remain useful when new categories enter the decision. We propose Pairwise Modality Evidence Transfer (PMET), a shared nonlinear correction of comparative multimodal evidence, to learn a reusable correction from these comparative preferences. We utilize a nonlinear antisymmetric scorer to combine modality margins, the baseline margin and description similarity, generated from semantic recognizer; uniform aggregation converts its pair corrections into residual class scores. To expose the learner to unfamiliar-category predictions during source training, we construct teachers that exclude selected source categories from fitting. Ground-truth query supervision jointly trains uniform and auxiliary relational aggregation, with only the uniform path retained at inference. The formulation changes candidate-level ordering while preserving the underlying semantic recognizer and its calibration policy. Experiments show that our proposed PMET achieves best performance on multiple emotion recognition datasets.

1 INTRODUCTION

Emotion recognition systems are usually trained to distinguish a fixed inventory of categories. In practice, however, the distinctions required by an application may extend beyond that inventory. And emotions that are relatively ambiguous and hard to collect, such as frustration and fear, brings additional difficulty into training models on wide range of emotions. Generalized zero-shot learning (GZSL) formalizes this setting: training examples come from seen categories, whereas evaluation includes both seen and unseen categories represented by semantic descriptions (Xian et al., 2017). The model must therefore transfer its interpretation of observed behavior to new semantic candidates while retaining its ability to recognize familiar emotions. Success requires both identifying evidence for an unseen label and comparing that label against the seen alternatives available at inference.

Multimodal observations offer a useful basis for this transfer because words, vocal expression and visible behavior provide complementary views of the same utterance. Their contribution is not uniform across candidate emotions. A modality that clearly separates sadness from neutrality may provide less useful evidence for distinguishing sadness from frustration, while another modality may support the opposite preference. A fixed weighting of modalities cannot respond to all such differences between candidate comparisons. Learning a separate correction for every named class is also insufficient in GZSL, because the output inventory used for training is incomplete at deployment. These observations motivate a decision rule that interprets modality evidence relative to the candidates under comparison and can reuse that interpretation when unfamiliar labels are introduced.

Existing research provides several foundations for such a rule. Semantic emotion models transfer knowledge through label descriptions and learned emotion representations (Dai et al., 2020; Qi et al., 2024; Peng et al., 2025; Jang et al., 2026). Audio–visual GZSL methods align modality representations with semantic spaces (Mercea et al., 2022; Kurzendörfer et al., 2024), while multimodal fusion and consistency methods model interaction or disagreement between observations (Rahman et al., 2020; Hazarika et al., 2020; Wang et al., 2026). Multi-modal open-vocabulary emotion recogni-

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.

tion solves the problem in another view by directly generating arbitrary emotion labels (Lian et al., 2025). However, it depends on large language models for reasoning and generating labels, which will struggle on compute-restricted deployment scenarios like edge computing.

Two main challenges reside in learning candidate-relative modality preference corrections with unseen label candidates. First, the correction must apply to a changing candidate vocabulary. A representation tied to named output slots would not specify how to handle a newly supplied emotion description. Second, the learning procedure must expose the correction to evidence generated for unfamiliar categories. Even a class-independent scorer could learn an unsuitable rule if it observed only the familiar-category errors of a single recognizer. These requirements concern different parts of the system: sharing the correction function addresses how it is applied to new candidates, whereas varying the categories available to its evidence generators addresses the predictions from which it learns.

We propose **Pairwise Modality Evidence Transfer (PMET)** to address these challenges. Here, a pair consists of two candidate emotions for the *same utterance*. Separate audio, text and visual probes compare the utterance with the descriptions of both candidates. The signed difference between their scores indicates which candidate each modality favors and by how much. A shared nonlinear function then combines these modality margins with the semantic recognizer's own margin and the similarity between candidate descriptions. Antisymmetrization ensures that reversing the order of a comparison reverses its correction. Uniform aggregation converts the resulting pair corrections into a residual score for each class. Because every pair uses the same scorer and there are no class-specific output weights, this computation can be applied to the supplied seen–unseen vocabulary.

To learn from predictions involving unfamiliar categories, we construct source teachers that each omit a subset of the seen categories during fitting. Labeled source queries then provide the correction learner with predictions involving both familiar and pseudo-unseen labels. Query ground truth supervises two aggregation paths over the same pair corrections. An auxiliary relational projection provides a second view during training, while direct supervision of uniform aggregation trains the path used at inference. This design permits a shared correction to be trained under source-category exclusions while keeping its deployment computation simple.

Our contributions are threefold:

- We formulate candidate-relative multimodal correction for generalized zero-shot emotion recognition. A shared nonlinear antisymmetric scorer combines signed modality preferences and semantic relations, then produces joint class residuals through uniform aggregation.
- We introduce a source-only correction-training procedure that combines category-excluded teachers with joint uniform/relational supervision. It exposes the learner to pseudo-unseen evidence while retaining a simple inference path over new candidate vocabularies.
- We evaluate our proposed algorithm on multiple multi-modal emotion recognition datasets. PMET achieves the best-performing method on GZSL multi-modal emotion recognition among evaluated methods and significant metric improvements among IEMOCAP, MELD and MER2024 datasets.

2 Related work

Semantic transfer and generalized zero-shot learning. Zero-shot recognition connects observations to a semantic representation of the label space, allowing classification to extend beyond a fixed supervised output layer. Semantic compatibility functions provide one approach to this connection (Akata et al., 2013), while generative and shared-latent methods synthesize or align representations for unseen categories (Xian et al., 2018; Schönfeld et al., 2019). When seen and unseen categories compete at inference, GZSL evaluation exposes the tendency to favor seen classes (Xian et al., 2017). Calibrated stacking, seen–unseen curves and adaptive confidence smoothing address this operating point (Chao et al., 2016; Atzmon & Chechik, 2019), and learning-to-adapt formulations provide another route to generalization (Zhang et al., 2022). PMET instead takes the semantic scores of a frozen recognizer as its starting point and learns a residual that changes candidate ordering.

Zero-shot emotion recognition. Emotion transfer has been studied through attributes, label descriptions, embedding relations and objectives designed for unseen emotions (Zhan et al., 2019; Xu et al., 2022; Banerjee et al., 2022; Yano et al., 2024). Modality-transferable emotion embeddings (Dai et al., 2020) and versatile multimodal learning (Qi et al., 2024) provide direct precedents for using multiple modalities in this setting. ProEmoTrans learns text-based emotion transfer and supplies the semantic recognition basis used in our system (Peng et al., 2025); SubT examines subspace tuning for transfer (Jang et al., 2026). Recent prototype and consistency approaches offer further mechanisms for relating observed evidence to emotion semantics (He et al., 2026; Yin et al., 2026). Since these studies differ in their corpora, label inventories and evaluation protocols, our adapted comparisons evaluate them under a common task rather than reproduce published leaderboard values. Within this shared setting, PMET focuses on learning a reusable correction of candidate-relative multimodal preferences from source-category exclusions.

Multimodal alignment, fusion and disagreement. Audio–visual GZSL methods align embeddings and transfer information through cross-modal semantic representations (Mazumder et al., 2021; Mercea et al., 2022; Kurzendörfer et al., 2024; Ma et al., 2026). More generally, low-rank fusion, cross-modal attention and modality-invariant/specific representations model complementary information across modalities (Liu et al., 2018; Rahman et al., 2020; Hazarika et al., 2020). Emotion-recognition surveys and benchmarks document the transfer challenges associated with these observations (Han et al., 2025; Marín et al., 2026), while conflict-aware cross-reconstruction explicitly addresses disagreement (Wang et al., 2026). PMET retains the underlying fusion model and uses modality evidence to make an additional decision-level comparison, focused on learning modality confliction among different emotion vocabularies.

Candidate relations and pairwise ranking. Modeling dependencies between candidates and learning relations for recognition are established ideas (Gan et al., 2016; Sung et al., 2018), including candidate-aware mechanisms developed in other recognition settings (Bao et al., 2026). Pairwise comparison graphs can also be projected onto class potentials through least squares (Jiang et al., 2011). We use this standard construction to obtain coherent class-score residuals, without claiming the projection or its permutation properties as new theory. Our contribution lies in combining comparative modality evidence with category-excluded emotion transfer and a relational path used only during training.

3 METHODOLOGY

3.1 SEMANTIC BASIS AND COMPARATIVE EVIDENCE

Let $\mathcal{S}$ and $\mathcal{U}$ denote disjoint seen and unseen emotion sets. Labeled fitting examples belong only to $\mathcal{S}$, while deployment predicts over the joint candidate set $\mathcal{C} = \mathcal{S} \cup \mathcal{U}$. Each utterance $x = (x^a, x^t, x^v)$ comprises audio, transcript and video. Class descriptions provide label-side semantic information, distinct from the utterance transcript. PMET learns a shared correction with a semantic recognizer by comparing the evidence for pairs of candidate emotions.

Baseline model. Our algorithm is based on a visual–audio–text (VAT) adaptation of ProEmoTrans (Peng et al., 2025). It uses a shared BERT encoder for transcripts and class descriptions, retaining the source-label CRF training and emotion-prototype enrichment of the baseline. VAT incorporates audio and video through residual cross-attention, with text querying projected audio/video features. Comparing the resulting representation with each candidate description gives a semantic score v_c, so unseen scores arise from semantic compatibility rather than unfitted classifier slots. Three separately trained probes expose the evidence from individual modalities. Each probe maps its modality representation through a 64-dimensional GELU hidden layer into the 768-dimensional description space of its owning recognizer, followed by normalization. Class-balanced cross-entropy fits the probes using only their owning recognizer's fitted categories. The recognizer and probes remain frozen during correction learning, so query supervision updates the shared correction rather than the evidence generators. Appendix C describes their fitting.

A pair (c, k) compares two candidate emotions for the same utterance. Let $q_{m,c}$ denote the cosine similarity between modality probe $m \in \{a, t, v\}$ and candidate c. We express candidate preferences as signed margins, using temperature T, fixed source-derived scales σ_m and description cosine sim-

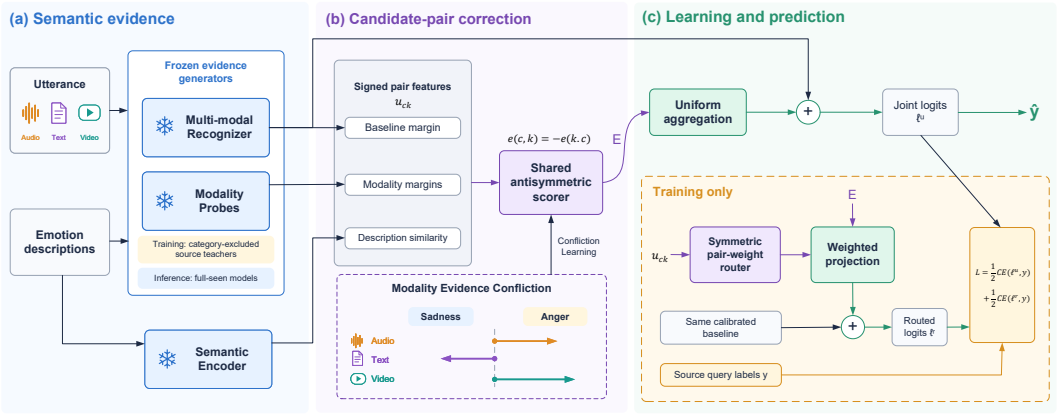


Figure 1: The architecture for our proposed method. PMET learns from category-excluded source teachers. The same pair corrections receive uniform and auxiliary relational supervision. Deployment retains the frozen full-seen evidence generators and uniform aggregation; the router is used only during training.

ilarity ρ_{ck}:

$$b_{ck} = \frac{v_c - v_k}{T}, \qquad d_{m,ck} = \frac{q_{m,c} - q_{m,k}}{\sigma_m}, \qquad u_{ck} = (b_{ck}, d_{a,ck}, d_{t,ck}, d_{v,ck}, \rho_{ck}). \tag{1}$$

Positive margins favor c, whereas negative margins favor k. Their signs distinguish agreement from competing modality preferences, and their magnitudes indicate preference strength. The shared nonlinear scorer uses this evidence and description similarity to condition each correction without a class-ID embedding. Each σ_m is the root-mean-square probe margin over query-training teacher presentations and source candidate pairs, floored at 10^{-4}. This label-free estimator excludes evaluation examples and measures margin scale, not confidence.

3.2 Antisymmetric Correction and Uniform Deployment

As correction is based on the modality tendency for candidate emotions, the correction should reverse direction when the candidate order is reversed. Reversing (c, k) negates its four margins while preserving ρ_{ck}. We enforce antisymmetry by applying a shared SiLU MLP network in both directions:

$$e_{ck} = \tfrac{1}{2}\big[f_\theta(u_{ck}) - f_\theta(u_{kc})\big], \qquad e_{cc} = 0. \tag{2}$$

To obtain class residuals, we aggregate each candidate's edges uniformly over all $K = |\mathcal{C}|$ candidates:

$$\delta_c^u = \frac{1}{K} \sum_{k \in \mathcal{C}} e_{ck}. \tag{3}$$

Antisymmetry makes the residual centered; this aggregation is the least-squares projection of the complete pair graph onto centered class potentials. We add the residual to the semantic score and retain the calibration penalty for fitted categories. For aggregation path p and the evidence generator's fitted category set $\mathcal{F}$, the calibrated logits are

$$\ell_c^p = \frac{v_c}{T} + \delta_c^p - \gamma \mathbf{1}[c \in \mathcal{F}]. \tag{4}$$

At inference, we use the uniform path $p = u$, set $\mathcal{F} = \mathcal{S}$, and predict the largest logit over the joint candidate set. Initializing the scorer's output to zero recovers the calibrated baseline, after which learning can change predictions for both seen and unseen emotions. Because the scorer is shared, permuting the candidates permutes the outputs consistently. The residuals nevertheless depend on the available comparisons, so changing the candidate roster can change them.

A globally affine f_θ reduces to ordinary linear score fusion: antisymmetrization removes its bias and symmetric similarity term, while uniform aggregation gives a linear combination of scores plus a candidate-common offset. Nonlinearity instead lets corrections depend jointly on the margins and semantic relationship. Appendix A derives the affine reduction.

3.3 CATEGORY-EXCLUDED TEACHING AND RELATIONAL SUPERVISION

We construct source teachers to expose correction learning to category exclusion. With all labeled source examples with $y \in \mathcal{S}$ which are disjoint from unseen evaluation set, we fit each teacher on $\mathcal{S}_j \subset \mathcal{S}$ and score queries against all descriptions in $\mathcal{S}$. A query's label is pseudo-unseen only relative to a teacher for which $y \notin \mathcal{S}_j$; the query remains a source-class example for the correction learner. Query ground truth supervises the correction directly, while teacher margins serve as inputs rather than distillation targets. Final-unseen examples and candidates in $\mathcal{U}$ are excluded from correction training. Each teacher and its probes exclude all correction-query utterances from fitting; excluded-category and query utterances are also removed from teacher context fragments. Scalar semantic margins provide a common interface across teachers without requiring aligned hidden spaces. During teaching, $K = |\mathcal{S}|$, and Eq. (4) uses $\mathcal{F} = \mathcal{S}_j$ to penalize only that teacher's familiar categories.

For example, a teacher fitted on neutral and sadness can score a happy query against all four source descriptions. Happy is pseudo-unseen to that teacher, although the query label is available for correction learning. A teacher fitted on happy can provide a familiar-category view of the same held-out query. These different exposure conditions prevent a source class from having a permanent "unseen" role in correction training. At inference, the same correction function receives evidence from a generator fitted on all seen emotions and evaluates comparisons involving final-unseen descriptions. The training procedure is intended to support this transfer across generators, although the pairwise parameterization alone does not guarantee it.

During training, an additional relational path weights comparisons differently while operating on the same predicted edges. Suppressing pair subscripts on the margins, the auxiliary router receives

$$r_{ck} = (|b|, |d_a|, |d_t|, |d_v|, |d_a - d_t|, |d_a - d_v|, |d_t - d_v|, \rho_{ck}). \tag{5}$$

All eight coordinates are unchanged when the pair is reversed. A small network followed by softplus and a positive floor produces symmetric weights w_{ck}, normalized to have mean one across pairs. The discrepancy terms let the router respond to differences between modality preferences without choosing an orientation for the edge. The weighted projection is

$$\delta^r = \arg\min_{\mathbf{1}^\top \delta = 0} \sum_{c<k} w_{ck} \left[(\delta_c - \delta_k) - e_{ck} \right]^2. \tag{6}$$

Let W have off-diagonal entries w_{ck} and a zero diagonal, define $L = \mathrm{diag}(W\mathbf{1}) - W$, and set $a_c = \sum_k W_{ck} e_{ck}$. The unique centered solution is obtained from

$$(L + \mathbf{1}\mathbf{1}^\top / K)\delta^r = a. \tag{7}$$

This differentiable solve passes query supervision to both the scorer and router. When pairwise preferences are inconsistent, the projection reconciles them into one set of class residuals rather than requiring every predicted edge to be exactly realizable.

Both paths receive ground-truth query supervision, with equal weighting across teachers and classes:

$$\mathcal{L} = \tfrac{1}{2} \left\langle \mathrm{CE}(\ell^r, y) + \mathrm{CE}(\ell^u, y) \right\rangle_{\text{teacher,class}}. \tag{8}$$

We average equally within teacher/class cells and then across cells to prevent larger classes from dominating. Repeated teacher views of one query are not independent observations. The uniform path is supervised directly because it is the path retained at deployment. The relational path provides an additional training signal for reconciling inconsistent pair evidence. After training, we discard the router and its weighted solve. Deployment therefore requires $O(K^2)$ pair scoring with a 315-parameter scorer and the three probes, which add 313,792 parameters; the 140-parameter router is used only during training. Appendix B gives the exact scale estimator, router architecture and minibatch weighting.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and category-transfer protocols. IEMOCAP (Busso et al., 2008) contains scripted and improvised conversations. In **IEMOCAP-7**, happiness, anger, neutral and sadness are seen categories; frustration, surprise and fear are unseen during task fitting. **IEMOCAP-4** comprises three

separately fitted tasks that hold out happiness, anger or sadness, while keeping neutral seen. All three use the same 1,096 endpoint utterances. We compute each metric within each task and then report its equal-weight mean. **MELD** (Poria et al., 2019) provides conversational emotion data from television dialogue. Our category split treats joy, anger, neutral and sadness as seen, and disgust, fear and surprise as unseen. **MER2024** (Lian et al., 2024) provides movie and television clips. We construct a category-transfer task from its labeled data, with neutral, anger, happiness and sadness seen, and worry and surprise unseen. All compared methods retain the same exclusions for missing English transcripts, including methods without transcript input.

Our protocols separate fitting and test utterance IDs and exclude final-unseen observations from every task-fitting stage. Each teacher excludes the query examples and its pseudo-unseen categories from fitting, whereas the deployment backbone uses all permitted source utterances, including the queries. Appendix D gives individual holdouts and external-dataset details.

Implementation details. PMET and its VAT foundation use BERT for transcript and label representations (Devlin et al., 2019), HuBERT-large features for audio (Hsu et al., 2021), and CLIP ViT-L/14 features for video (Radford et al., 2021). IEMOCAP and MELD use English HuBERT; MER2024 uses Chinese HuBERT and the release's English transcripts with English BERT. The compared methods retain their declared observation modalities and differ in encoder resources and parameter budgets. Class descriptions remain available as semantic side information even for methods without transcript input.

The text recognizer and its VAT adaptation are fitted for ten epochs with AdamW, learning rate 2×10^{-5}, 100 warmup updates and linear decay. Six teachers fitted on two source classes each generate scores over all four source candidates. We fit PMET for 100 epochs using AdamW, learning rate 10^{-3}, weight decay 10^{-4}, batch size 64 and gradient clipping at one. Temperature and the seen-class penalty are fixed across matched corrections after source-only selection. All experiments are conducted with Nvidia RTX 3090 GPUs. Appendix C details the recognizer and correction-fitting stages; Appendix D describes the external comparisons and VML adaptation.

Metrics and result provenance. Table 1 reports overall, seen-class and unseen-class weighted F1 (WF1; F, F_S and F_U, respectively), the seen–unseen recall harmonic mean (H), and recall-based AUSUC (Chao et al., 2016). We introduced WF1 for better presenting model recognition performance, as legacy recall-based metrics cannot cover biased predictions, for those models only excel on seen or unseen recognition, while predicting wrong on other categories, they can still pull up recall-based metrics with their high recall.

For a label group G, $\mathrm{WF1}_G = \sum_{c \in G} n_c F1_c / \sum_{c \in G} n_c$, with each class F1 computed from predictions over the full joint label space. Thus group scores retain false positives from the other group. While S_R and U_R denote mean-class recalls on seen and unseen categories, $H = 2S_R U_R/(S_R + U_R)$. AUSUC sweeps the seen-class penalty. WF1 reflects the observed class distribution, whereas H and AUSUC weight emotions equally within each group. Seen and unseen mean-class recalls and individual-holdout results are retained in the appendix.

4.2 Comparison with existing methods

We compare with a multiclass adaptation of MTEE (Dai et al., 2020), AVCA (Mercea et al., 2022), ClipClap (Kurzendörfer et al., 2024), VML (Qi et al., 2024), ProEmoTrans (Peng et al., 2025) and SubT (Jang et al., 2026). ProEmoTrans uses text, SubT uses audio, and AVCA, ClipClap and VML use audio and video. We additionally include their VAT adaptations; the MTEE adaptation and PMET also use all three modalities. All reported numbers are obtained for our task adaptations, rather than copied from the original papers.

On IEMOCAP-7, PMET obtains the highest reported overall WF1, H and AUSUC. Relative to ProEmoTrans-VAT, overall WF1 increases from 48.01 to 54.30, H from 34.53 to 41.60, and AUSUC from 0.2376 to 0.3123. The corresponding seen and unseen WF1 gains are 8.70 and 4.97 points. Across the three IEMOCAP-4 holdouts, PMET improves mean overall WF1 by 5.75 points over ProEmoTrans-VAT and leads all five aggregate measures in Table 1. Though AVCA achieved higher seen WF1 in this scenario, it failed to transfer the semantic information to unseen classes, resulting in much worse unseen and overall WF1 values.

Table 1: Complete-system comparisons on four category-transfer evaluations. F, F_S and F_U denote overall, seen and unseen WF1 (%); H is recall-based (%) and AUSUC is unscaled. IEMOCAP-4 averages the three affective holdouts. Bold denotes the highest reported value in each column.

	IEMOCAP-7					IEMOCAP-4 average				
Method	F	F_S	F_U	H	AUSUC	F	F_S	F_U	H	AUSUC
MTEE	38.72	27.10	45.11	22.89	0.1070	38.87	44.24	19.61	28.10	0.2878
AVCA	28.73	36.11	24.66	12.14	0.1435	37.58	40.26	21.27	14.74	0.3092
AVCA-VAT	39.55	**41.02**	38.74	22.46	0.1915	48.04	55.03	21.98	30.17	0.3514
ClipClap	27.63	20.99	31.29	22.13	0.1068	41.36	48.75	14.06	19.17	0.2878
ClipClap-VAT	34.60	39.71	31.80	27.96	0.1518	53.97	64.25	17.29	22.52	0.4011
VML	26.80	15.35	33.10	19.19	0.0945	39.90	44.47	20.95	27.80	0.2942
VML-VAT	38.62	28.01	44.46	25.64	0.1421	40.40	43.67	25.34	31.91	0.3467
ProEmoTrans	48.11	34.67	55.50	36.96	0.2504	54.70	59.07	39.55	52.17	0.5156
ProEmoTrans-VAT	48.01	30.47	57.65	34.53	0.2376	54.91	59.36	39.27	51.12	0.5164
SubT	10.60	14.54	8.43	15.22	0.1169	48.54	53.55	34.58	35.12	0.5301
SubT-VAT	32.03	20.12	38.57	21.85	0.1615	53.85	60.06	33.16	39.74	0.5403
PMET (ours)	**54.30**	39.17	**62.62**	**41.60**	**0.3123**	**60.66**	**65.67**	**43.05**	**58.89**	**0.5776**
	MELD					MER2024				
Method	F	F_S	F_U	H	AUSUC	F	F_S	F_U	H	AUSUC
MTEE	36.92	41.02	14.23	22.25	0.1038	67.42	75.84	23.55	36.65	0.2638
AVCA	36.42	41.04	10.86	15.37	0.0645	65.75	75.31	15.89	23.71	0.2555
AVCA-VAT	43.54	49.54	10.30	21.97	0.0982	66.57	76.06	17.08	21.58	0.2671
ClipClap	21.91	23.50	13.08	15.27	0.0644	34.73	37.89	18.29	30.67	0.2226
ClipClap-VAT	38.43	43.86	8.39	22.40	0.0943	56.75	65.87	9.18	30.27	0.2577
VML	32.69	36.16	13.44	19.08	0.0907	10.66	9.59	16.24	18.33	0.2271
VML-VAT	48.36	54.77	12.82	23.27	0.1176	18.68	19.34	15.24	19.63	0.1801
ProEmoTrans	35.05	36.83	25.16	22.70	0.1558	35.05	37.17	24.00	**38.88**	0.2959
ProEmoTrans-VAT	41.78	44.70	25.54	27.93	0.1543	60.75	67.53	25.40	37.43	0.2954
SubT	22.59	24.35	12.85	12.88	0.0566	39.63	46.32	4.76	26.44	0.1725
SubT-VAT	30.28	33.13	14.49	20.32	0.1052	42.61	50.12	3.48	27.27	0.1984
PMET (ours)	**52.00**	**56.77**	**25.56**	**29.98**	**0.1560**	**71.33**	**79.21**	**30.26**	31.96	**0.3440**

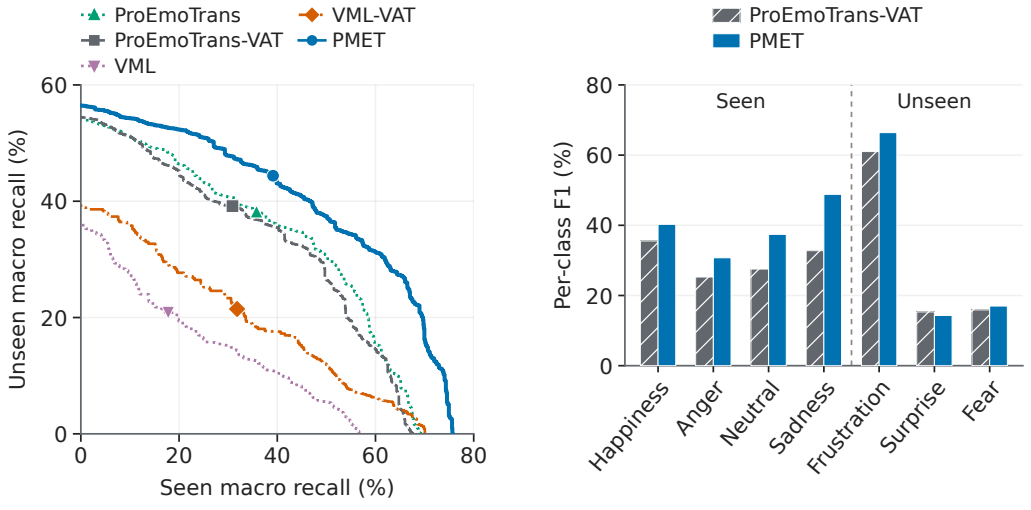
MER2024 provides a comparison on a verified common evaluation population, with some of the models failing to transfer emotional evidences to unseen categories in a more strict environment. PMET reaches 71.33 WF1 and 0.3440 AUSUC, the highest values among the listed systems. Against ProEmoTrans-VAT, the gains are 10.58 WF1 points and 0.0486 AUSUC; seen and unseen WF1 improve by 11.68 and 4.86 points. ProEmoTrans achieved highest H with high unseen recall due to recognizing most samples as unseen categories, this results in low WF1 scores with a high H. MELD summaries show 52.00 WF1 for PMET, compared with 41.78 for ProEmoTrans-VAT and 48.36 for VML-VAT. PMET achieved much better seen recognition performance while holding similar semantic transfer and unseen recognition performance.

4.3 Ablation study

Table 2 compares evidence, correction and teacher-design variants. *VAT baseline* denotes ProEmoTrans-VAT, and *Direct MLP fusion* directly combines normalized VAT and probe scores with a shared candidate-wise network. *Single all-seen teacher* replaces category-excluded teaching with one teacher fitted on all source categories. *Matched all-seen bank* uses six independently fitted all-seen teachers, matching PMET's teacher count on each dataset in this table. Both teacher controls retain query exclusions, the correction architecture, total query presentations, head-update budget and deployment calibration.

Table 2: Ablations of correction evidence, architecture and teacher design. F denotes overall weighted F1 (%). Higher is better, and bold marks the best value in each column.

	IEMOCAP-7		MELD		MER2024	
Variant	F	AUSUC	F	AUSUC	F	AUSUC
VAT baseline	48.01	0.2376	41.78	0.1543	60.75	0.2954
w/o probe evidence	48.63	0.2632	46.44	0.1454	60.74	0.2873
Unsigned probe margins	49.07	0.2586	37.17	0.1200	58.56	0.2828
Direct MLP fusion	50.22	0.2867	49.02	0.1529	69.26	0.3377
Single all-seen teacher	19.88	0.2039	48.65	0.1498	66.43	0.3219
Matched all-seen bank	25.50	0.2367	47.92	0.1528	67.14	0.3144
w/o relational loss	53.47	0.2950	50.49	0.1549	70.56	0.3432
w/o uniform loss	53.09	0.3045	49.34	0.1526	70.79	0.3416
PMET (ours)	**54.30**	**0.3123**	**52.00**	**0.1560**	**71.33**	**0.3440**



(a) Seen–unseen recall tradeoff.

(b) Per-emotion F1.

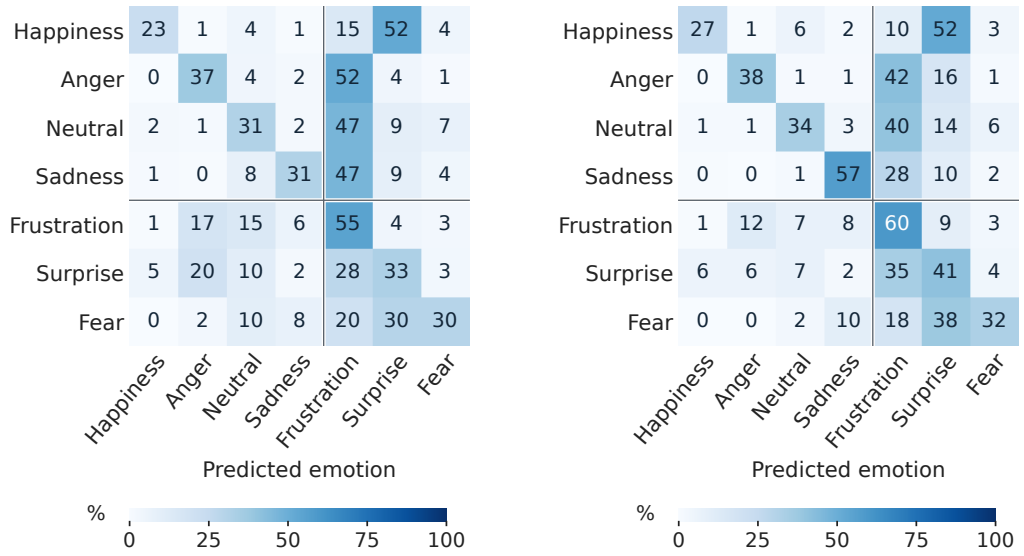
Figure 2: IEMOCAP-7 comparison. (a) Five-method recall curves; markers show recorded calibrated operating points. (b) Class-wise F1 for ProEmoTrans-VAT and PMET at those points. The first four emotions are seen; frustration, surprise and fear are unseen.

PMET achieves the highest reported F and AUSUC in Table 2. It exceeds direct MLP fusion in F by 4.08, 2.98 and 2.07 points on IEMOCAP-7, MELD and MER2024, respectively. Removing probe evidence, margin direction or either training loss reduces both metrics relative to PMET.

On IEMOCAP-7, correction learning with teachers that lack pseudo-unseen categories fails to achieve effective unseen transfer, even when the teacher count is matched. The correction learned redirects unseen examples toward seen classes despite improving source-validation loss, and finally yields unseen WF1 of 7.28% for the single teacher and 15.16% for the matched bank, accompanied by overall WF1 of 19.88% and 25.50%, respectively. The matched-bank result shows that the transfer deficit cannot be explained by teacher count alone, supporting the importance of pseudo-unseen exposure during correction learning.

4.4 Class coverage and error patterns

Figures 2 and 3 use the IEMOCAP-7 result in Table 1, with their recorded calibration for point predictions.



(a) ProEmoTrans-VAT. (b) PMET.

Figure 3: Confusion matrices for ProEmoTrans-VAT and PMET. Rows are true emotions; columns are joint-label predictions at each model's recorded calibration. Cells show integer-rounded row percentages, with recall on the diagonal.

Figure 2a extends the comparison beyond a single operating point. Among these five systems, PMET has the largest AUSUC (0.3123), compared with 0.2504 for text-only ProEmoTrans, 0.2376 for ProEmoTrans-VAT, 0.1421 for VML-VAT and 0.0945 for VML. The grouped bars in Figure 2b show how aggregate performance is distributed across emotions. PMET has higher F1 than ProEmoTrans-VAT on six of seven classes, with increases from 32.77 to 48.82 for sadness and from 27.53 to 37.42 for neutral. Frustration F1 increases from 61.01 to 66.40, whereas surprise F1 decreases a little from 15.32 to 14.31; fear changes from 15.89 to 16.99. The class-wise comparison therefore exposes differences hidden by weighted averages. In particular, frustration accounts for most of unseen utterances, so its performance strongly influences unseen WF1 while surprise and fear remain difficult.

Figure 3 compares the class confusions of the two reported systems. Neutral utterances predicted as frustration account for 46.7% under ProEmoTrans-VAT and 39.9% under PMET; for sadness, the corresponding proportions are 46.8% and 28.2%. Sadness recall is 31.5% for ProEmoTrans-VAT and 57.4% for PMET. These patterns are consistent with PMET's stronger seen-class performance in Table 1. Surprise illustrates why the confusion diagonal and class F1 must be interpreted together: its recall rises from 32.7% to 41.1%, but its F1 falls because the model predicts surprise more often for other emotions.

5 Conclusion

We presented PMET, a method for learning transferable corrections from comparative audio, text and visual evidence in generalized zero-shot emotion recognition. Category-excluded source teachers provide predictions for both familiar and pseudo-unseen emotions, allowing a shared nonlinear scorer to learn from these different exposure conditions. Antisymmetric pair corrections are trained through uniform and auxiliary relational aggregation, while deployment uses only the uniform path on top of a frozen semantic recognizer.

Experiments across multiple emotion datasets show improved aggregate recognition under the joint seen–unseen setting, while class-level analysis reveals heterogeneous gains across individual emotions, with matched analyses highlighting the importance of separate directional evidence and nonlinear processing.

AI USE STATEMENT

Generative AI tools were used for literature-search assistance and language editing. Human authors independently verified the retrieved information against the original sources, methodological decisions, implementations, experimental results, and claims.

REPRODUCIBILITY STATEMENT

We specified algorithm design, training algorithms, used datasets, evaluation protocols and implementation settings within main text, and provides derivations, detailed implementation, as well as detailed results in the appendix. We attached code for PMET in supplementary material.

REFERENCES

Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-Embedding for Attribute-Based Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

Yuval Atzmon and Gal Chechik. Adaptive Confidence Smoothing for Generalized Zero-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Abhishek Banerjee, Uttaran Bhattacharya, and Aniket Bera. Learning Unseen Emotions from Gestures via Semantically-Conditioned Zero-Shot Perception with Adversarial Autoencoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 3–10, 2022. doi: 10.1609/aaai.v36i1.19873.

Zehao Bao, Shujun Guo, and Bruce X. B. Yu. Visual Anchoring in Diffusion: Multimodal Zero-Shot Skeleton Action Recognition, 2026.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335, November 2008. ISSN 1574-0218. doi: 10.1007/s10579-008-9076-6.

Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An Empirical Study and Analysis of Generalized Zero-Shot Learning for Object Recognition in the Wild. In *Computer Vision – ECCV 2016*, volume 9906 of *Lecture Notes in Computer Science*, pp. 52–68. Springer, 2016. doi: 10.1007/978-3-319-46475-6_4.

Wenliang Dai, Zihan Liu, Tiezheng Yu, and Pascale Fung. Modality-Transferable Emotion Embeddings for Low-Resource Multimodal Emotion Recognition. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pp. 269–280. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.aacl-main.30.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423.

Chuang Gan, Ming Lin, Yi Yang, Gerard de Melo, and Alexander G. Hauptmann. Concepts Not Alone: Exploring Pairwise Relationships for Zero-Shot Video Activity Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 3487–3493, 2016.

Zhiyuan Han, Beier Zhu, Yanlong Xu, Peipei Song, and Xun Yang. Benchmarking and Bridging Emotion Conflicts for Multimodal Emotion Reasoning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 5528–5537, Dublin, Ireland, October 2025. Association for Computing Machinery. ISBN 979-8-4007-2035-2. doi: 10.1145/3746027.3754856.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: Modality-Invariant and - Specific Representations for Multimodal Sentiment Analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1122–1131. Association for Computing Machinery, 2020. doi: 10.1145/3394171.3413678.

Kang He, Boyu Chen, Yuzhe Ding, Fei Li, Chong Teng, and Donghong Ji. PaSE: Prototype-aligned Calibration and Shapley-based Equilibrium for Multimodal Sentiment Analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 30960–30968, 2026. doi: 10.1609/aaai.v40i37.40355.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. doi: 10.1109/TASLP.2021.3122291.

Jaehyuk Jang, Kangwook Ko, Wonjun Lee, and Changick Kim. SubT: Subspace Tuning for Few-shot Generalization of Audio-Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2026*. Association for Computational Linguistics, 2026.

Xiaoye Jiang, Lek-Heng Lim, Yuan Yao, and Yinyu Ye. Statistical ranking and combinatorial Hodge theory. *Mathematical Programming*, 127(1):203–244, 2011. doi: 10.1007/s10107-010-0419-x.

David Kurzendörfer, Otniel-Bogdan Mercea, A. Sophia Koepke, and Zeynep Akata. Audio-Visual Generalized Zero-Shot Learning using Pre-Trained Large Multi-Modal Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2627–2638, 2024.

Zheng Lian, Haiyang Sun, Licai Sun, Zhuofan Wen, Siyuan Zhang, Shun Chen, Hao Gu, Jinming Zhao, Ziyang Ma, Xie Chen, Jiangyan Yi, Rui Liu, Kele Xu, Bin Liu, Erik Cambria, Guoying Zhao, Björn W. Schuller, and Jianhua Tao. MER 2024: Semi-supervised learning, noise robustness, and open-vocabulary multimodal emotion recognition. In *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*, MRAC '24, pp. 41–48. Association for Computing Machinery, 2024. doi: 10.1145/3689092.3689959.

Zheng Lian, Haiyang Sun, Licai Sun, Haoyu Chen, Lan Chen, Hao Gu, Zhuofan Wen, Shun Chen, Zhang Siyuan, Hailiang Yao, Bin Liu, Rui Liu, Shan Liang, Ya Li, Jiangyan Yi, and Jianhua Tao. OV-MER: Towards Open-Vocabulary Multimodal Emotion Recognition. In *Forty-Second International Conference on Machine Learning*, June 2025.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2247–2256. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1209.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Siteng Ma, Xiaoyu Niu, Haocheng Tang, Zhe Yang, Jisheng Chu, and Wenrui Li. Fusion-regularized alignment modality-adaptive audio-visual network for audio-visual zero-shot learning. *Neurocomputing*, 685:133693, 2026.

Oriol Marín, Roger Marí, Gloria Haro, and Rafael Redondo. Multimodal Emotion Recognition in Conversations via Class-Wise Adaptive Modality Fusion and Affective Geometry. In *11th Workshop and Competition on Affective Behavior Analysis In-the-Wild (ABAW), ECCV 2026*, 2026.

Pratik Mazumder, Pravendra Singh, Kranti Kumar Parida, and Vinay P. Namboodiri. AVGZSLNet: Audio-Visual Generalized Zero-Shot Learning by Reconstructing Label Features From Multi-Modal Embeddings. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3090–3099, 2021.

Otniel-Bogdan Mercea, Lukas Riesch, A. Sophia Koepke, and Zeynep Akata. Audio-Visual Generalised Zero-Shot Learning With Cross-Modal Attention and Language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10553–10563, 2022.

Kun Peng, Cong Cao, Hao Peng, Guanlin Wu, Zhifeng Hao, Lei Jiang, Yanbing Liu, and Philip S. Yu. Emotion Transfer with Enhanced Prototype for Unseen Emotion Recognition in Conversation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 596–608. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.31.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1050.

Fan Qi, Huaiwen Zhang, Xiaoshan Yang, and Changsheng Xu. A Versatile Multimodal Learning Framework for Zero-Shot Emotion Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7):5728–5741, 2024. doi: 10.1109/TCSVT.2024.3362270.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Wasifur Rahman, Md. Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating Multimodal Information in Large Pretrained Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2359–2369. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.214.

Edgar Schönfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized Zero- and Few-Shot Learning via Aligned Variational Autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8247–8255, 2019.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. Learning to Compare: Relation Network for Few-Shot Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1199–1208, 2018.

Yan Wang, Fuyuan Cao, and Xingwang Zhao. Conflict-Aware Adaptive Cross-Reconstruction for Multimodal Sentiment Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15880–15889, 2026.

Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-Shot Learning—The Good, the Bad and the Ugly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4582–4591, 2017.

Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature Generating Networks for Zero-Shot Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5542–5551, 2018.

Xinzhou Xu, Jun Deng, Nicholas Cummins, Zixing Zhang, Li Zhao, and Björn W. Schuller. Exploring Zero-Shot Emotion Recognition in Speech Using Semantic-Embedding Prototypes. *IEEE Transactions on Multimedia*, 24:2752–2765, 2022. doi: 10.1109/TMM.2021.3087098.

Taro Yano, Kunihiro Takeoka, and Masafumi Oyamada. Relevance, Diversity, and Exclusivity: Designing Keyword-augmentation Strategy for Zero-shot Classifiers. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pp. 106–119. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.starsem-1.9.

Wen Yin, Siyu Zhan, Cencen Liu, Xin Hu, Guiduo Duan, Xiurui Xie, Yuan-Fang Li, and Tao He. TiCAL: Typicality-Based Consistency-Aware Learning for Multimodal Emotion Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 17948–17956, 2026. doi: 10.1609/aaai.v40i21.38854.

Chi Zhan, Dongyu She, Sicheng Zhao, Ming-Ming Cheng, and Jufeng Yang. Zero-Shot Emotion Recognition via Affective Structural Embedding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1151–1160, 2019.

Yiwen Zhang, Caixia Yuan, Xiaojie Wang, Ziwei Bai, and Yongbin Liu. Learn to Adapt for Generalized Zero-Shot Text Classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 517–527. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.39.

A AGGREGATION PROPERTIES

We first derive uniform aggregation from the projection objective. The complete candidate graph with unit off-diagonal weights has Laplacian $KI - \mathbf{1}\mathbf{1}^\top$. Because the centered class potentials sum to zero, differentiating the least-squares objective in Eq. (6) gives $K\delta_c = \sum_k e_{ck}$, which yields Eq. (3). Antisymmetry ensures that the resulting correction is centered:

$$\sum_c \delta_c^u = \frac{1}{K} \sum_{c,k} e_{ck} = 0. \tag{A1}$$

The construction also preserves the correspondence between candidates and their scores when candidate order changes. Under any consistent permutation of scores and descriptions, the shared scorer evaluates the same corresponding pair features. The edge matrix is therefore permuted in both dimensions, and uniform aggregation applies the same permutation to the correction vector. Candidate ordering consequently leaves the associated scores and prediction unchanged, apart from implementation-dependent ties. This property and centering hold even when the learned edges are not transitively consistent: the least-squares projection reconciles their inconsistent component without a separately trained cycle constraint.

Linear pair scoring reduces to ordinary score fusion. Write $z_c = v_c/T$ and $r_{m,c} = q_{m,c}/\sigma_m$. For a globally affine pair scorer, antisymmetrization removes its bias and the term linear in symmetric description similarity. Defining $g_c = a_b z_c + \sum_m a_m r_{m,c}$ gives

$$e_{ck} = g_c - g_k. \tag{A2}$$

Uniform aggregation then yields

$$\delta_c^u = g_c - \frac{1}{K} \sum_k g_k. \tag{A3}$$

Fixed calibration parameters can have a changed effective operating point. Suppose $e_{ck} = \alpha(z_c - z_k)$ with one global scalar α. The final scores, up to a candidate-common term, become $(1+\alpha)z_c - \gamma \mathbf{1}[c \in \mathcal{F}]$. When $1 + \alpha > 0$, the prediction is equivalent to using the original semantic scores with

$$\gamma_{\text{eff}} = \frac{\gamma}{1+\alpha}. \tag{A4}$$

In this example, the correction changes the effective seen penalty without changing any within-group ordering. Sweeping gamma over its complete range recovers the original tradeoff curve under a positive global rescaling, subject to tie and numerical conventions. A sample-dependent rescaling need not preserve this curve. Keeping the numeric calibration parameter fixed therefore does not by itself guarantee an unchanged operating point for seen–unseen discrimination.

Projection and edge interpretation. If every edge has the form $e_{ck} = r_c - r_k$, any positive edge weighting in Eq. (6) recovers the centered vector $r - \bar{r}\mathbf{1}$. Weighting matters when the edge field contains inconsistent comparisons, because it determines how the projection reconciles them. Moreover, equal forward outputs at a consistent configuration do not generally imply equal derivatives for an unconstrained scorer: the projection operators can respond differently to perturbations outside the consistent subspace.

Uniform deployment discards every antisymmetric edge component whose row sums are zero, so distinct internal edge fields can produce identical class residuals. We therefore interpret predictions primarily through the effective class-margin change $\delta_c - \delta_k$. Individual edges and router weights provide additional internal diagnostics, but do not uniquely determine the deployed correction.

B CORRECTION IMPLEMENTATION AND DATA AUDIT

B.1 SCALES, ROUTER AND WEIGHTED PROJECTION

To make margins from different probes comparable, we use a fixed scale for each modality. For modality m, let Ω collect query-training teacher presentations and unordered source pairs. We compute the label-free RMS scale as

$$\sigma_m = \max\left(10^{-4}, \sqrt{|\Omega|^{-1} \sum_{(i,c,k)\in\Omega} (q_{m,c}^{(i)} - q_{m,k}^{(i)})^2}\right). \tag{B1}$$

This scale measures the magnitude of source margins; it is neither a centered standard deviation nor an estimate of confidence. Its computation excludes all evaluation examples. The router uses the eight reversal-invariant coordinates in Eq. (5). An $8 \to 14 \to 1$ SiLU network, with hidden bias and bias-free output, gives

$$\widetilde{w}_{ck} = 10^{-3} + \mathrm{softplus}(h_\phi(r_{ck})), \qquad w_{ck} = \widetilde{w}_{ck} / \operatorname{mean}_{a<b} \widetilde{w}_{ab}. \tag{B2}$$

The resulting weights are positive and symmetric, with mean one. We initialize the final network weights to zero, so the scorer initially applies no correction and the router assigns uniform weights. The differentiable centered solve in Eq. (7) allows the routed training loss to update both the scorer and the router. Uniform deployment uses the closed-form aggregation and bypasses the solve. To prevent larger teacher/class cells from dominating training, we average equally over teachers and source classes. For J teachers, K_s source classes and n_{jy} queries in teacher/class cell (j, y), this average is

$$\langle g \rangle_{\text{teacher,class}} = \frac{1}{JK_s} \sum_{j,y} \frac{1}{n_{jy}} \sum_{i\in Q_{jy}} g_i. \tag{B3}$$

We implement this average with minibatch weights $N/(JK_s n_{jy})$, where N is the number of pooled presentations. These weights balance the loss but do not make repeated teacher views of the same utterance independent observations. The seen penalty is defined by each teacher's fitted categories during teaching and by the complete seen set at deployment. Inference requires neither ground-truth labels nor a constraint that preserves seen predictions.

B.2 PROBES AND CONTROLS

The probes receive 1,024-dimensional English HuBERT-large-ll60k audio features, 768-dimensional CLIP ViT-L/14 video features and the owning model's BERT CLS text representation. Each probe consists of Linear(input,64), GELU and Linear(64,768), followed by normalization. We construct the direction-removal control by replacing signed probe margins with their magnitudes in both training paths. Under pair reversal, only the baseline margin then changes sign, while unsigned probe evidence and description similarity remain unchanged. This control retains baseline direction and removes the direction of both agreement and opposition in the probe evidence. The no-probe control instead masks modality-specific information in both paths. We also evaluate a magnitude-only router, but treat it as an alternative formulation because matching its parameter count requires changing the hidden width in addition to its inputs.

C PROEMOTRANS-VAT BASELINE

C.1 REPRESENTATIONS, FUSION AND SEMANTIC INFERENCE

We initialize the reference model from `bert-base-uncased` and use its 768-dimensional final CLS representation for both transcripts and emotion descriptions. Inputs are padded or truncated to 128 tokens. To control memory use, we encode utterances in chunks of at most 16, while each optimizer batch contains four conversation fragments. Description vectors remain connected to the shared encoder and receive gradients during native fitting.

For the seven-class setting, we use the authors' descriptions, each containing a definition and two examples. Anger, happiness, neutral, sadness and frustration use the IEMOCAP descriptions, while

surprise and fear use the authors' MELD descriptions. All descriptions fit within 128 tokens. The surprise and fear descriptions are supplied class-side resources and are not fitted to evaluation utterances. At inference, the owning model encodes the unseen descriptions alongside those of the seen categories.

VAT independently projects the L2-normalized audio and video features to 768 dimensions, allowing the text representation to attend to both modalities. With text vector x_i^t and projected tokens a'_i, v'_i, the fusion uses eight attention heads and zero attention dropout:

$$o_i = \mathrm{MHA}(\mathrm{LN}(x_i^t), [a'_i, v'_i], [a'_i, v'_i]), \tag{C1}$$

$$z_i = x_i^t + P_o(o_i). \tag{C2}$$

We initialize P_o to zero, so fusion initially preserves the text representation. The inherited Gaussian context step then uses distance zero on the diagonal and $|i-j|+1$ off-diagonal, together with a unit-width Gaussian density G_{ij} and valid-position masking:

$$A_{ij} = \mathrm{softmax}_j(G_{ij}\cos(z_i, z_j)/0.02)\,, \qquad h_i = \sum_j A_{ij} z_j. \tag{C3}$$

For fitted categories $\mathcal{F}$, the inherited CRF receives probability-valued emissions instead of raw cosine logits:

$$E_{is} = \mathrm{softmax}_{s\in\mathcal{F}}(\cos(h_i, t_s)/0.02), \tag{C4}$$

$$\mathcal{L}_{\mathrm{base}} = -B^{-1}\sum_{b=1}^{B} \log P_{\mathrm{CRF}}(y^{(b)} \mid E^{(b)}). \tag{C5}$$

During native training, the likelihood updates both BERT and the source CRF. VAT training subsequently uses cached BERT features and descriptions, freezes the CRF parameters and updates the fusion module through the same likelihood.

Attention Viterbi Decoding (AVD) uses source transition structure M to enrich representations:

$$\alpha_i = \mathrm{softmax}(\mathrm{AVD}(E, M)_i), \tag{C6}$$

$$\widetilde{h}_i = \tfrac{1}{2}h_i + \tfrac{1}{2}\sum_{s\in\mathcal{F}} \alpha_{is} t_s, \tag{C7}$$

$$v_{ic} = \cos(\widetilde{h}_i, t_c), \quad c \in \mathcal{C}. \tag{C8}$$

C.2 Optimization and selection

Both baseline stages use legacy Transformers AdamW (Loshchilov & Hutter, 2019) with betas $(0.9, 0.999)$, epsilon 10^{-8} and bias correction. Bias and LayerNorm weights are excluded from weight decay. Training uses float32, fragment shuffling without replacement and BERT gradient checkpointing. When selection scores tie, we retain the earliest checkpoint.

Final probes exclude queries from their own updates, although their BERT owner is fitted on the permitted 4,091-example pool that includes queries. Teacher probes and their owning models both exclude queries. Probe cross-entropy uses the calibrated temperature of the owning model. The RMS normalization subsequently applied in Eq. (B1) is a separate transformation that scales pairwise margins.

D Detailed comparisons and reproduction protocols

This appendix retains the full metric breakdown underlying Table 1, including individual IEMOCAP-4 holdouts.

D.1 Protocol and reproduction details

MER2024 protocol and representations. We construct a six-class category-transfer task from the gold-labelled MER2024 clips, with neutral, anger, happiness and sadness as seen categories

and worry and surprise as unseen categories. For PMET and its VAT foundation, the release's English transcripts are used with the English BERT text encoder, while frozen Chinese HuBERT-large replaces the English acoustic encoder; video representations use frozen CLIP ViT-L/14. We retain 4,813 clips with available English transcripts and apply a class-stratified clip split that keeps exact-file duplicates together.

VML adaptations. We implement the affective graph, private/shared modality decomposition, two-stage co-attention and grouped semantic decoder of VML as a paper-based adaptation to our category-transfer protocols. The audio–visual model uses frozen HuBERT and CLIP features at eight ordered temporal positions, 300-dimensional GloVe class embeddings and external lexical valence, arousal and dominance priors. Audio uses the English HuBERT model for IEMOCAP and MELD and the Chinese model for MER2024. The VAT extension adds frozen contextual MPNet transcript embeddings pooled into up to eight ordered lexical bins and extends the private/shared branches and fusion stage to three modalities. Both VML variants use 150-dimensional hidden representations, five attention heads and a fixed group capacity of two. We train with AdamW for 100 epochs, learning rate 3×10^{-4}, weight decay 10^{-4} and batch size 64.

D.2 IEMOCAP-7

	Weighted F1 (%)			Recall-based (%)			
Method	All	Seen	Unseen	S_R	U_R	H_R	AUSUC
MTEE	38.72	27.10	45.11	26.42	20.19	22.89	0.1070
AVCA	28.73	36.11	24.66	58.87	6.77	12.14	0.1435
AVCA-VAT	39.55	**41.02**	38.74	57.21	13.97	22.46	0.1915
ClipClap	27.63	20.99	31.29	19.11	26.27	22.13	0.1068
ClipClap-VAT	34.60	39.71	31.80	**44.36**	20.41	27.96	0.1518
VML	26.80	15.35	33.10	17.79	20.84	19.19	0.0945
VML-VAT	38.62	28.01	44.46	31.82	21.47	25.64	0.1421
ProEmoTrans	48.11	34.67	55.50	35.78	38.22	36.96	0.2504
ProEmoTrans-VAT	48.01	30.47	57.65	30.88	39.16	34.53	0.2376
SubT	10.60	14.54	8.43	10.31	29.08	15.22	0.1169
SubT-VAT	32.03	20.12	38.57	16.04	34.25	21.85	0.1615
PMET	**54.30**	39.17	**62.62**	39.14	**44.39**	**41.60**	**0.3123**

D.3 IEMOCAP-4-HAPPINESS

	Weighted F1 (%)			Recall-based (%)			
Method	All	Seen	Unseen	S_R	U_R	H_R	AUSUC
MTEE	35.44	40.05	24.42	42.56	27.97	33.76	0.3233
AVCA	19.28	9.37	43.02	8.22	**85.86**	15.00	0.3786
AVCA-VAT	45.20	49.29	35.42	49.57	42.11	45.54	0.4594
ClipClap	44.22	48.94	32.95	47.63	42.00	44.64	0.4254
ClipClap-VAT	56.42	65.25	35.30	**68.23**	30.13	41.80	0.5148
VML	34.92	34.11	36.86	33.95	56.04	42.29	0.3821
VML-VAT	36.31	32.98	44.26	29.99	73.99	42.68	0.4178
ProEmoTrans	55.63	60.16	44.80	57.57	47.37	51.97	0.5165
ProEmoTrans-VAT	56.14	60.34	46.09	57.81	49.23	53.17	0.5234
SubT	53.25	64.80	25.63	75.60	15.79	26.12	0.5360
SubT-VAT	56.86	**65.03**	37.33	71.32	30.03	42.26	0.5427
PMET	**59.03**	63.83	**47.53**	57.00	60.99	**58.93**	**0.5764**

D.4 IEMOCAP-4-ANGER

Method	Weighted F1 (%)			Recall-based (%)			AUSUC
	All	Seen	Unseen	S_R	U_R	H_R	
MTEE	37.96	41.01	25.72	37.18	36.53	36.85	0.2954
AVCA	37.10	43.34	12.12	46.19	11.26	18.11	0.2579
AVCA-VAT	39.84	45.29	18.01	43.83	21.92	29.22	0.3153
ClipClap	37.95	45.13	9.24	47.20	7.46	12.88	0.2272
ClipClap-VAT	51.42	62.66	6.38	67.18	6.39	11.67	0.3542
VML	35.10	39.12	18.98	33.94	29.68	31.67	0.2786
VML-VAT	43.38	48.29	23.72	43.68	33.79	38.10	0.3597
ProEmoTrans	52.33	57.26	32.58	57.28	39.27	46.59	0.4775
ProEmoTrans-VAT	54.11	59.51	32.50	**59.10**	35.62	44.45	0.4815
SubT	35.89	34.33	42.14	28.84	86.30	43.23	0.4956
SubT-VAT	47.20	48.45	**42.20**	45.55	**70.78**	55.43	**0.5336**
PMET	**58.11**	**63.23**	37.60	58.38	54.34	**56.28**	0.5254

D.5 IEMOCAP-4-SADNESS

Method	Weighted F1 (%)			Recall-based (%)			AUSUC
	All	Seen	Unseen	S_R	U_R	H_R	
MTEE	43.20	51.67	8.70	52.77	7.87	13.70	0.2448
AVCA	56.36	68.07	8.67	73.13	6.02	11.12	0.2910
AVCA-VAT	59.08	70.52	12.50	74.69	8.80	15.74	0.2794
ClipClap	41.90	52.19	0.00	59.19	0.00	0.00	0.2107
ClipClap-VAT	54.08	64.85	10.18	67.32	7.87	14.09	0.3342
VML	49.69	60.16	7.01	65.14	5.09	9.45	0.2220
VML-VAT	41.53	49.74	8.05	49.91	8.80	14.96	0.2625
ProEmoTrans	56.13	59.78	41.26	55.48	**60.65**	57.95	0.5528
ProEmoTrans-VAT	54.48	58.22	39.22	55.93	55.56	55.74	0.5442
SubT	56.48	61.52	35.96	67.63	24.54	36.01	0.5586
SubT-VAT	57.50	66.71	19.95	**72.28**	12.65	21.53	0.5446
PMET	**64.83**	**69.94**	**44.01**	65.53	57.87	**61.46**	**0.6310**

D.6 MELD

Method	Weighted F1 (%)			Recall-based (%)			AUSUC
	All	Seen	Unseen	S_R	U_R	H_R	
MTEE	36.92	41.02	14.23	19.00	26.83	22.25	0.1038
AVCA	36.42	41.04	10.86	24.99	11.10	15.37	0.0645
AVCA-VAT	43.54	49.54	10.30	32.92	16.49	21.97	0.0982
ClipClap	21.91	23.50	13.08	11.00	24.97	15.27	0.0644
ClipClap-VAT	38.43	43.86	8.39	25.26	20.12	22.40	0.0943
VML	32.69	36.16	13.44	16.05	23.54	19.08	0.0907
VML-VAT	48.36	54.77	12.82	35.20	17.39	23.27	0.1176
ProEmoTrans	35.05	36.83	25.16	16.02	**38.89**	22.70	0.1558
ProEmoTrans-VAT	41.78	44.70	25.54	22.90	35.79	27.93	0.1543
SubT	22.59	24.35	12.85	8.81	23.96	12.88	0.0566
SubT-VAT	30.28	33.13	14.49	16.49	26.46	20.32	0.1052
PMET	**52.00**	**56.77**	**25.56**	**37.34**	25.04	**29.98**	**0.1560**

D.7 MER2024

Method	Weighted F1 (%)			Recall-based (%)			AUSUC
	All	Seen	Unseen	S_R	U_R	H_R	
MTEE	67.42	75.84	23.55	69.56	24.88	36.65	0.2638
AVCA	65.75	75.31	15.89	75.85	14.05	23.71	0.2555
AVCA-VAT	66.57	76.06	17.08	74.39	12.62	21.58	0.2671
ClipClap	34.73	37.89	18.29	25.56	38.33	30.67	0.2226
ClipClap-VAT	56.75	65.87	9.18	61.65	20.06	30.27	0.2577
VML	10.66	9.59	16.24	11.76	41.49	18.33	0.2271
VML-VAT	18.68	19.34	15.24	14.77	29.23	19.63	0.1801
ProEmoTrans	35.05	37.17	24.00	29.38	**57.44**	**38.88**	0.2959
ProEmoTrans-VAT	60.75	67.53	25.40	62.12	26.79	37.43	0.2954
SubT	39.63	46.32	4.76	39.71	19.82	26.44	0.1725
SubT-VAT	42.61	50.12	3.48	45.86	19.40	27.27	0.1984
PMET	**71.33**	**79.21**	**30.26**	**80.42**	19.94	31.96	**0.3440**