# TIE: TEXT-BOUNDED INCREMENTAL EVIDENCE MIGRATION FOR LLM-BASED MULTIMODAL AFFECTIVE COMPUTING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

LLM-based multimodal affective computing uses a large language model (LLM) to directly predict sentiment or emotion from multimodal inputs. Non-textual affective evidence is constructed, transformed into LLM-compatible representations, and propagated through the language-centered LLM. Along this path, a text-directed preference may be reinforced even when the constructed multimodal evidence supports the correct prediction, a phenomenon we term **text-prior reinforcement**. To address this problem, we formulate LLM-based multimodal affective prediction as **Text-Bounded Incremental Evidence Migration (TIE)**, treating the full path from evidence construction to its use in prediction as a unified process. Guided by this formulation, TIE uses textual semantics and text-bounded temporal routing to organize audio and visual evidence along text-referenced and text-residual pathways. To preserve the influence of this evidence beyond representation construction, TIE transforms the organized evidence into LLM-compatible representations and applies a migration-specific signal derived from the frozen final prediction block. For sentiment regression, an order-aware objective bridges label generation and ordered sentiment estimation. TIE outperforms same-backbone compact-prompt baselines on most of the five benchmarks. The combined training configuration also improves existing compact-prompt methods. Controlled ablations validate TIE's components, while migration analyses provide empirical support for text-prior reinforcement.

## 1 INTRODUCTION

Multimodal affective computing aims to infer affective states from heterogeneous signals. Multimodal sentiment analysis (MSA) estimates sentiment polarity and intensity, whereas emotion recognition in conversation (ERC) identifies categorical emotions in conversational contexts. Text usually provides explicit semantic content and contextual grounding, yet textual expressions can remain ambiguous, implicit, sarcastic, or context-dependent. Audio and visual modalities provide complementary affective cues, such as prosody, facial expression, hesitation, and emphasis, which can refine affective intensity, resolve ambiguity, or compensate for insufficient textual evidence. Effective multimodal affective computing therefore requires models to preserve modality-specific cues while interpreting them within a coherent semantic context.

Conventional multimodal affective computing generally follows a task-specific prediction paradigm. As illustrated in Figure 1(a), modality representations, cross-modal interactions, and a task-specific predictor are jointly optimized for affective prediction. This paradigm focuses on bridging modality heterogeneity and combining complementary, conflicting, or jointly informative signals.

Pretrained language models further strengthen textual modeling in multimodal affective computing. Compared with static word embeddings, models such as BERT and XLNet provide contextualized text representations. In earlier frameworks, they mainly provide textual representations, while downstream components perform multimodal modeling and affective prediction.

Recent LLM-based affective computing methods introduce a different paradigm. Rather than serving only as textual representation backbones, LLMs act as affective reasoners and prediction interfaces. Lightweight multimodal interfaces transform non-textual representations into compact LLM-



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.

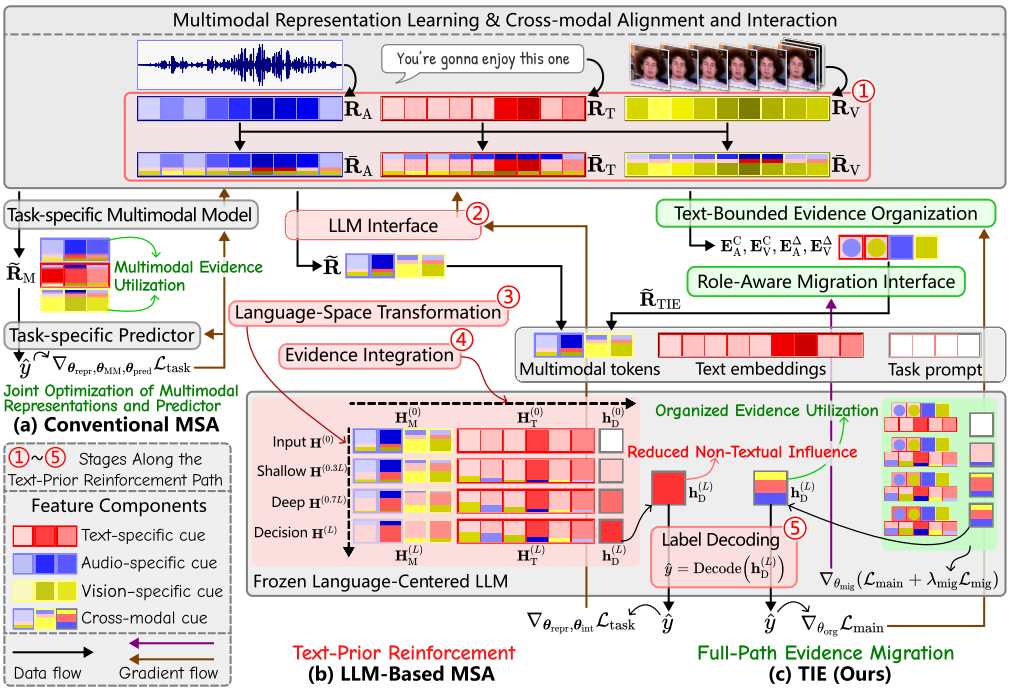



Figure 1: Motivation and overview of (a) conventional multimodal affective computing, (b) text-prior reinforcement along the evidence-migration path of LLM-based affective prediction, and (c) TIE's full-path evidence migration.

compatible tokens, retaining pretrained reasoning capabilities while reducing trainable parameters. As illustrated in Figure 1(b), these tokens are propagated with text through the frozen LLM before final label decoding.

However, constructing and injecting multimodal evidence does not ensure that the final prediction resolves conflicts with textual cues. Along the evidence-migration path, a text-directed preference may be reinforced even while the constructed evidence supports the correct multimodal label. We refer to this path-level phenomenon as **text-prior reinforcement**.

To address this problem, we formulate LLM-based multimodal affective prediction as **Text-Bounded Incremental Evidence Migration (TIE)**, treating the successive transformations from evidence construction to predictive utilization as a unified process. Under this formulation, text provides contextual boundaries for interpreting non-textual expressions, while audio and vision contribute evidence that may correspond to or extend beyond textual expression.

Guided by this formulation, TIE uses textual semantics and text-bounded temporal routing to encode audio and visual states alongside their deviations from modality-conditioned textual references, forming text-referenced and text-residual pathways (Figure 1(c)). To preserve the influence of this evidence beyond representation construction, TIE transforms the organized evidence into LLM-compatible representations and applies a migration-specific signal derived from the frozen final prediction block. For sentiment regression, an order-aware objective bridges label generation and ordered sentiment estimation.

We evaluate TIE on three MSA benchmarks (CMU-MOSI, CMU-MOSEI, and SIMS-V2) and two ERC benchmarks (MELD and CHERMA) with multiple LLM backbones. TIE outperforms same-backbone compact-prompt baselines on most benchmarks. The combined training configuration also improves existing compact-prompt methods while preserving their inference-time evidence sequences. Controlled ablations validate evidence organization, evidence transformation, migration supervision, and order-aware prediction, while migration analyses provide further empirical support for text-prior reinforcement.

Our main contributions are summarized as follows:

- We formulate **text-prior reinforcement** as a path-level problem in LLM-based affective prediction, trace text-directed preferences during propagation, and test source-position contributions to label prediction.
- We propose **Text-Bounded Incremental Evidence Migration (TIE)**, which uses textual semantics and text-bounded temporal routing to organize audio and visual evidence along text-referenced and text-residual pathways and promotes the continued influence of this evidence through the frozen LLM.
- We introduce migration supervision derived from the frozen final prediction block, together with an EMD-based order-aware objective for sentiment regression. Both training mechanisms can be applied to different compact-prompt methods without additional inference cost.
- Extensive experiments demonstrate the effectiveness and generality of TIE and provide empirical support for text-prior reinforcement.

## 2 Related Work

**Multimodal Representation and Interaction.** Task-specific multimodal affective models study representation factorization, cross-modal interaction, and adaptive fusion. Factorization separates shared and modality-specific information (Hazarika et al., 2020; Wang et al., 2025; Yang & Li, 2026; Jian et al., 2026; Liu & Wang, 2026; Meng et al., 2026). Cross-modal interaction builds on attention (Tsai et al., 2019), state-space models (Li et al., 2024; 2025b; Jiang et al., 2025b), and coarse-to-fine alignment (Zhao et al., 2026b). Further work addresses disagreement and reconstruction (Luo et al., 2025; Wang et al., 2026b), semantic distribution consistency (Wang et al., 2026a), and diffusion- or flow-based fusion (Lv et al., 2025; Mai & Han, 2026). Adaptive fusion uses uncertainty or confidence (Gao et al., 2024; Jiang et al., 2026), modality balancing and collaboration (Nguyen et al., 2024; He et al., 2026), or text-conditioned expert routing (Rao et al., 2026).

**Pretrained Language Models.** Pretrained language models provide contextualized text representations for multimodal modeling. MAG introduces non-textual information as gated displacements of contextual word representations, whereas CHFN models interactions between text and unaligned non-textual sequences (Rahman et al., 2020; Guo et al., 2022). UniMSE and UniSA further employ generative pretrained models to unify multiple sentiment-related tasks. Despite using smaller backbones than recent LLMs, task-specific or multi-task optimization can still incur substantial computational costs and specialize these models toward sentiment analysis (Hu et al., 2022; Li et al., 2023b). In these approaches, multimodal representations and prediction remain jointly optimized within task-specific architectures.

**LLM-Based Multimodal Affective Computing.** Multimodal LLMs typically connect non-textual encoders to language models through projection-based, query-based, or query-as-prompt interfaces (Eichenberg et al., 2022; Li et al., 2023a; Liang et al., 2024). Freezing the pretrained language model and optimizing only lightweight multimodal interfaces can reduce trainable parameters while preserving pretrained capabilities and mitigating catastrophic forgetting (Liu et al., 2022a; 2024). Recent studies extend LLMs to multimodal sentiment analysis and emotion understanding (Lian et al., 2025; Han et al., 2026; Zhang et al., 2026a). For compact-prompt adaptation, MSE-Adapter generates text-aligned pseudo tokens from acoustic and visual features, whereas PaP constructs text-conditioned sentiment prototypes for a frozen LLM (Yang et al., 2025a; Zhao et al., 2026a).

**Text Dominance and Modality Imbalance.** Prior work addresses unequal or divergent modality contributions through modality-specific supervision, conflict-aware interaction, and explicit modeling of text-corresponding and text-distant information (Liu et al., 2022b; Luo et al., 2025; Sun & Tian, 2025; Guo et al., 2024; Yang et al., 2025b). Related studies further identify textual dominance and the underutilization of non-textual information in general-purpose MLLMs (Wu et al., 2025; Zhao et al., 2025; Li et al., 2025c; 2026; Sun et al., 2026; Zhang et al., 2026b; Chen et al., 2026). Beyond modality imbalance during representation learning and interaction, we also study the continued influence of non-textual affective evidence within a pretrained language-centered model.

# 3 METHOD

## 3.1 OVERVIEW

Given textual, acoustic, and visual sequences $X_T$, $X_A$, and $X_V$, we view multimodal affective prediction with a frozen LLM as an evidence-migration process:

$$(X_T, X_A, X_V) \xrightarrow{\mathcal{O}_{\theta_{\text{org}}}} \mathcal{E} \xrightarrow{\mathcal{M}_{\theta_{\text{mig}}}} \widetilde{\mathcal{E}} \xrightarrow{F_{\text{LLM}}} h_D \xrightarrow{\text{Decode}} \widehat{Y}. \tag{1}$$

Here, $\mathcal{O}_{\theta_{\text{org}}}$ organizes multimodal inputs into compact evidence $\mathcal{E}$, and the migration interface $\mathcal{M}_{\theta_{\text{mig}}}$ transforms it into LLM-compatible representations $\widetilde{\mathcal{E}}$. The frozen backbone $F_{\text{LLM}}$ produces the decision state $h_D$ for decoding $\widehat{Y}$. Existing compact-prompt methods train their interfaces using task-loss gradients backpropagated through the frozen LLM. However, this supervision leaves the intermediate evidence-migration process implicit. Effective evidence must remain compatible with the frozen model, propagate through it, and influence the decision state.

TIE addresses this problem through three coordinated components. **Text-Bounded Evidence Organization** uses textual semantics and soft temporal routing to organize temporally contextualized acoustic and visual representations into text-referenced and text-residual evidence. **Role-Aware Evidence Migration** transforms the organized evidence into compact LLM-compatible tokens through shared and role-conditioned migration directions. **Frozen-Model Migration Supervision** evaluates the propagated evidence through a compact compatibility computation derived from the frozen model's final prediction block and directs this feedback specifically to the migration interface.

## 3.2 TEXT-BOUNDED EVIDENCE ORGANIZATION

### 3.2.1 TEMPORAL MODALITY ENCODING AND TEXT-BOUNDED ROUTING

For acoustic and visual sequences $X_m \in \mathbb{R}^{L_m \times d_m}$, $m \in \{A, V\}$, where $L_m$ is the number of modality frames, independent bidirectional LSTMs produce temporally contextualized states $H_m \in \mathbb{R}^{L_m \times d_E}$. Here, $d_E$ is the evidence dimension. A shallow layer of the frozen LLM encodes the transcript into contextualized text representations, which a bottleneck adapter projects into the same evidence space as $T \in \mathbb{R}^{L_T \times d_E}$, where $L_T$ is the number of transcript-token positions.

We construct soft temporal routing weights $\Pi_{m,t,j} \in [0, 1]$ between modality frames and transcript positions, with $\sum_{j=1}^{L_T} \Pi_{m,t,j} = 1$. The tokenizer-independent lexical timeline provides a soft temporal reference, not exact frame-to-word alignment. Further construction details are provided in Appendix A.1.

### 3.2.2 TEXT-CONDITIONED EVIDENCE REFERENCE

We construct a local text-conditioned reference for each modality frame:

$$\widehat{H}_{m,t} = \sum_{j=1}^{L_T} \Pi_{m,t,j}\, g_m(T_j) \in \mathbb{R}^{d_E}, \tag{2}$$

where $g_m : \mathbb{R}^{d_E} \to \mathbb{R}^{d_E}$ is modality-specific. We train the reference with a masked Huber loss $\mathcal{L}_{\text{ref}}$ against the stop-gradient modality states. The full loss definition is given in Eq. (19).

### 3.2.3 TEXT-REFERENCED AND TEXT-RESIDUAL EVIDENCE RESAMPLING

Let $r \in \{C, \Delta\}$ index the text-referenced and text-residual roles. We use the complete modality representations and their deviations from the textual reference as two evidence sources:

$$X^C_{m,t} = H_{m,t} \in \mathbb{R}^{d_E}, \qquad X^{\Delta}_{m,t} = H_{m,t} - \text{sg}[\widehat{H}_{m,t}] \in \mathbb{R}^{d_E}, \tag{3}$$

Here, sg denotes stop-gradient. The text-referenced pathway retains the complete temporally contextualized modality representation, while the text-residual pathway encodes its deviation from the modality-conditioned textual reference. For each modality-role pair, a learned query resamples the full temporal sequence through masked multi-head attention. The role source supplies the values,

while the source, local reference, and their interaction form the keys. This produces four evidence representations $\mathcal{E} = [E_A^C, E_V^C, E_A^\Delta, E_V^\Delta] \in \mathbb{R}^{4\times d_E}$. Figure 5 illustrates the organization path; resampling details are given in Appendix A.2.

## 3.3 Role-Aware Evidence Migration

### 3.3.1 Role-Aware Migration Interface

Each evidence representation passes through a shared projector with a role-specific low-rank update. Let $Z_m^r = \mathrm{RMSNorm}(E_m^r)$:

$$\widetilde{E}_m^r = W_S Z_m^r + b_S + U_r V_r Z_m^r \in \mathbb{R}^{d_L}, \tag{4}$$

Here, $U_r V_r$ has rank at most $d_M$ and is shared by acoustic and visual evidence with role $r$. These role-specific residual updates instantiate the incremental migration over the shared mapping $W_S$.

The four migrated tokens precede the transcript in $\mathcal{I} = [\widetilde{E}_A^C; \widetilde{E}_V^C; \widetilde{E}_A^\Delta; \widetilde{E}_V^\Delta; X_T^{\mathrm{emb}}; \mathrm{Task}]$. During training, the target sequence $Y$ is appended for teacher forcing. Causal attention makes these evidence tokens available to all subsequent positions.

### 3.3.2 Frozen-Model Migration Supervision

The task loss updates evidence organization and the migration interface through the frozen LLM. We also compute a migration-specific signal using its final prediction block. Let $H_{\mathcal{E}}^{L-1}$ and $h_D^{L-1}$ denote the four evidence states and decision state at the input to the final block. We form

$$H_{\mathrm{mig}}^{L-1} = \left[H_{\mathcal{E}}^{L-1}; \mathrm{sg}[h_D^{L-1}]\right] \in \mathbb{R}^{5\times d_L}. \tag{5}$$

The detached decision state supplies predictive context; the auxiliary gradient passes through the evidence states to the migration interface. Applying the frozen final block $F_L$, normalization $N_L$, and native LM head gives

$$\mathcal{L}_{\mathrm{mig}} = -\log \mathrm{softmax}\Big(W_{\mathrm{LM}} N_L(F_L(H_{\mathrm{mig}}^{L-1})_D)\Big)_{y_1}. \tag{6}$$

Here, $y_1$ is the sentiment sign token for MSA or the class token for ERC, and the gradient is applied only to the migration interface. This objective requires an additional frozen final-block pass during training and adds no inference-time module; target, positional, and gradient details are provided in Appendix A.3.

## 3.4 Generative Sentiment and Emotion Prediction

TIE formulates MSA and ERC as label generation. Given multimodal input $\mathcal{I}$ and target sequence $Y = (y_1, \ldots, y_{L_Y})$, TIE is trained with the frozen LLM under teacher forcing:

$$\mathcal{L}_{\mathrm{tok}} = -\frac{1}{L_Y}\sum_{\ell=1}^{L_Y} \log p\left(y_\ell \mid \mathcal{I}, y_{<\ell}\right). \tag{7}$$

For ERC, $Y$ is a categorical emotion label; for MSA regression, it is a serialized scalar sentiment value.

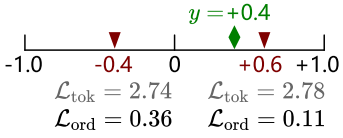



Figure 2: Similar token losses despite target distances of 0.8 and 0.2.

## 3.5 Order-Aware Sentiment Estimation

For MSA regression, token generation does not explicitly represent distances between sentiment values. Let $\mathcal{V}_Y = \{v_1 < \cdots < v_K\}$ be the dataset-specific uniformly spaced sentiment grid. Using label-position logits from the standard teacher-forcing forward, we define

$$s_k = \sum_{\ell=1}^{L_Y} \log p\left(\mathrm{tok}_\ell(v_k) \mid \mathcal{I}, y_{<\ell}\right), \qquad \widehat{p}_k = \frac{\exp(s_k)}{\sum_{j=1}^{K}\exp(s_j)}. \tag{8}$$

The normalization yields a teacher-forced ordinal surrogate over the sentiment values. For $k_y = \arg\min_k |v_k - y|$, we use the mean CDF discrepancy, equivalently a scaled one-dimensional Earth Mover's Distance:

$$\mathcal{L}_{\text{ord}} = \frac{1}{K}\text{EMD}(\widehat{p}, y) = \frac{1}{K}\sum_{k=1}^{K}|\text{CDF}_{\widehat{p}}(v_k) - \text{CDF}_y(v_k)| = \frac{1}{K}\sum_{j=1}^{K}\widehat{p}_j|j - k_y|. \tag{9}$$

The cumulative distributions and the equivalence are detailed in Appendix A.4. This objective complements exact-token supervision by assigning larger penalties to ordinal mass farther from the target. It reuses the standard forward logits and adds neither an additional LLM forward nor inference-time computation. We define $\mathcal{L}_{\text{task}} = \mathcal{L}_{\text{tok}} + \lambda_{\text{ord}}\mathcal{L}_{\text{ord}}$ for MSA regression and $\mathcal{L}_{\text{task}} = \mathcal{L}_{\text{tok}}$ for ERC.

### 3.6 Overall Objective and Gradient Routing

The training objectives are

$$\mathcal{L}_{\text{main}} = \mathcal{L}_{\text{task}} + \lambda_{\text{ref}}\mathcal{L}_{\text{ref}}, \qquad \mathcal{L}_{\text{total}} = \mathcal{L}_{\text{main}} + \lambda_{\text{mig}}\mathcal{L}_{\text{mig}}, \tag{10}$$

with distinct gradient scopes. For evidence-organization parameters $\theta_{\text{org}}$ and migration-interface parameters $\theta_{\text{mig}}$,

$$\begin{aligned} g_{\text{org}} &= \nabla_{\theta_{\text{org}}}\mathcal{L}_{\text{main}}, & \nabla_{\theta_{\text{org}}}\mathcal{L}_{\text{mig}} &= 0, \\ g_{\text{mig}} &= \nabla_{\theta_{\text{mig}}}\mathcal{L}_{\text{main}} + \lambda_{\text{mig}}\nabla_{\theta_{\text{mig}}}\mathcal{L}_{\text{mig}}, & \nabla_{\theta_{\text{LLM}}}\mathcal{L}_{\text{total}} &= 0. \end{aligned} \tag{11}$$

Thus, the main objective trains evidence organization and task prediction, while migration supervision updates only the interface through which organized evidence enters the frozen model. The reference and migration objectives are training-only and add no inference-time components.

## 4 Experiments

### 4.1 Experimental Setup

**Datasets and Metrics.** We evaluate MSA regression on the English CMU-MOSI and CMU-MOSEI benchmarks and the Chinese CH-SIMS v2.0 benchmark, and ERC on the English MELD and Chinese CHERMA benchmarks (Zadeh et al., 2016; Bagher Zadeh et al., 2018; Liu et al., 2022b; Poria et al., 2019; Sun et al., 2023). We follow the official train/validation/test splits. For MSA, we report mean absolute error (MAE), Pearson correlation (Corr), and the standard dataset-specific classification metrics derived from the continuous predictions. For ERC, we report accuracy and weighted F1. Dataset statistics, metric definitions, and feature preparation are in Appendices B.1, B.2, and C.1.

**Baselines.** We compare TIE with conventional task-specific methods, pretrained-language model methods, and compact-prompt LLM methods. The principal compact-prompt baselines are MSE-Adapter and PaP. Baseline descriptions and implementation details are in Appendices B.3 and C.3.

**Implementation Details.** We use frozen ChatGLM3-6B-Base (Team GLM et al., 2024) and Qwen3.5-2B/9B (Qwen Team, 2026) backbones, with pretrained weights obtained from ModelScope. Throughout this paper, "ChatGLM3-6B" denotes ChatGLM3-6B-Base. Only the evidence organization and migration modules are optimized. We monitor MAE for MSA or weighted F1 for ERC on the validation split to guide early stopping. We repeat training with five random seeds and report the mean across runs. Training details and hyperparameters are in Appendix C.4.

### 4.2 Main Results

Baseline results are from the cited sources unless marked as our reproductions.

With ChatGLM3-6B, TIE reduces MAE relative to PaP by $2.5\%$ on CMU-MOSI, $1.2\%$ on CMU-MOSEI, and $9.5\%$ on CH-SIMS v2.0 (Tables 1–2). It also improves the reported classification

Table 1: Performance comparison on CMU-MOSI and CMU-MOSEI.

| | CMU-MOSI | | | | | CMU-MOSEI | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | MAE↓ | F1↑ | Acc-2↑ | Acc-7↑ | Corr↑ | MAE↓ | F1↑ | Acc-2↑ | Acc-7↑ | Corr↑ |
| **Task-Specific Models** | | | | | | | | | | |
| $\text{MAG}_{\text{BERT}}$ [ACL20] | 0.712 | 84.10 | 84.20 | – | 0.796 | 0.583 | 82.77 | 82.51 | 50.41 | 0.741 |
| MISA [ACMMM20] | 0.783 | 81.70 | 81.80 | 42.30 | 0.761 | 0.555 | 83.80 | 83.60 | 52.20 | 0.756 |
| $\text{CHFN}_{\text{BERT}}$ [ACMMM22] | 0.689 | 84.20 | 84.30 | 48.60 | 0.809 | 0.525 | 83.90 | 83.70 | 54.30 | 0.778 |
| UniMSE [EMNLP22] | 0.691 | 85.83 | 85.85 | 48.68 | 0.809 | 0.523 | 85.79 | 85.86 | 54.39 | 0.773 |
| $\text{UniSA}_{\text{BART}}$ [ACMMM23] | 0.742 | – | 84.11 | 48.54 | – | 0.587 | – | 84.93 | 50.03 | – |
| ITHP [ICLR24] | 0.643 | 88.60 | 88.70 | – | 0.852 | 0.564 | 87.40 | 87.30 | – | 0.813 |
| EUAR [ACMMM24] | – | 86.30 | 86.30 | 46.10 | – | – | 86.40 | 86.60 | 54.90 | – |
| CMamba [NIPS24] | – | – | – | – | – | 0.547 | 85.60 | 85.70 | – | 0.758 |
| DLF [AAAI25] | 0.731 | 85.04 | 85.06 | 47.08 | 0.781 | 0.536 | 85.27 | 85.42 | 53.90 | 0.764 |
| EMOE [CVPR25] | 0.697 | 85.30 | 85.40 | 47.80 | – | 0.530 | 85.50 | 85.50 | 53.90 | – |
| AlignMamba [CVPR25] | – | 86.90 | 86.90 | – | – | – | 86.50 | 86.60 | – | – |
| DDSE [ACMMM25] | 0.699 | 86.58 | 86.74 | 47.52 | 0.794 | 0.537 | 85.38 | 85.47 | 53.81 | 0.767 |
| DiffuFuse [ACMMM25] | 0.672 | 85.90 | 86.10 | 50.70 | 0.791 | 0.521 | 86.40 | 86.70 | 55.40 | 0.776 |
| MDF [AAAI26] | 0.692 | 85.03 | 84.99 | 48.25 | 0.810 | 0.525 | 83.92 | 85.01 | 54.80 | 0.788 |
| FINE [AAAI26] | – | 86.94 | 86.95 | 48.50 | – | – | 87.68 | 87.70 | 55.59 | – |
| CCAF [WWW26] | 0.743 | 85.75 | 85.82 | – | 0.790 | 0.592 | 85.91 | 85.90 | – | 0.794 |
| TCMax [ICLR26] | 0.778 | 84.13 | 84.27 | – | 0.775 | 0.589 | 85.52 | 85.61 | – | 0.789 |
| FUSE-Net [CVPR26] | 0.688 | 83.75 | 83.53 | 49.27 | 0.798 | 0.527 | 84.50 | 84.52 | 54.32 | 0.772 |
| CICA [CVPR26] | 0.630 | 88.14 | 88.19 | 49.56 | 0.855 | 0.489 | 84.72 | 85.15 | 55.29 | 0.856 |
| CACR [CVPR26] | 0.710 | 86.90 | 87.04 | 48.69 | – | 0.532 | 85.36 | 85.37 | 54.41 | – |
| TSD [CVPR26] | 0.691 | 86.60 | 86.50 | 49.00 | – | 0.525 | 86.20 | 86.20 | 54.60 | – |
| EBMC [CVPR26] | 0.694 | 86.20 | 86.26 | 50.34 | 0.833 | 0.505 | 86.23 | 86.04 | 57.32 | 0.824 |
| CaReFlow [CVPR26] | 0.616 | 89.70 | 89.80 | 50.60 | 0.858 | 0.504 | 88.00 | 87.90 | 55.70 | 0.799 |
| ConsMSA [ICML26] | 0.698 | 86.06 | 86.06 | 47.96 | – | 0.525 | 86.65 | 86.67 | 54.92 | – |
| **ChatGLM3-6B** | | | | | | | | | | |
| MAG[1] [ACL20] | – | – | – | – | – | 0.588 | 84.73 | 85.10 | 50.52 | 0.725 |
| CHFN[1] [ACMMM22] | – | – | – | – | – | 0.563 | 85.26 | 85.58 | 51.50 | 0.745 |
| MSE[2] [AAAI25] | 0.607 | 88.13 | 89.04 | 46.79 | 0.860 | 0.515 | 86.77 | 86.91 | 54.56 | 0.783 |
| PaP[2] [CVPR26] | 0.563 | 88.46 | 89.50 | 52.24 | 0.878 | 0.495 | 86.87 | 87.16 | 54.82 | 0.802 |
| TIE [Ours] | 0.549 | 90.34 | 90.38 | 53.21 | 0.883 | 0.489 | 88.04 | 88.21 | 55.19 | 0.807 |
| **Qwen3.5-2B** | | | | | | | | | | |
| TIE [Ours] | 0.631 | 86.23 | 86.27 | 49.01 | 0.844 | 0.550 | 85.17 | 85.48 | 52.03 | 0.744 |
| **Qwen3.5-9B** | | | | | | | | | | |
| TIE [Ours] | 0.526 | 90.45 | 90.47 | 53.64 | 0.895 | 0.488 | 88.89 | 88.87 | 55.35 | 0.826 |

metrics and correlation over PaP on these datasets, including weak-intensity accuracy on CH-SIMS v2.0.

For emotion recognition, TIE achieves $68.35\%$ accuracy and $67.08\%$ weighted F1 on MELD with ChatGLM3-6B (Table 3). The results indicate benefits across sentiment and emotion prediction, without uniformly leading every benchmark or metric.

### 4.3 Component Analysis

**Evidence Organization.** On CH-SIMS v2.0, resampling complete modality representations and their reference residuals reduces MAE by $11.0\%$ relative to two complete-representation branches of equal capacity (Table 17). Enabling the reference loss further reduces MAE by $9.6\%$ relative to its removal (Table 18). These comparisons support the use of a supervised local reference and distinct resampling sources.

**Coordinated Migration Training.** We compare each constructor's native configuration (Org) with Full, which adds the applicable migration interface, frozen-model migration supervision, and order-aware objective while preserving its evidence constructor and inference-time prompt sequence.

[1]ChatGLM3-6B results for MAG and CHFN are reported by MSE-Adapter (Yang et al., 2025a); their original publications did not use this backbone.

[2]The CMU-MOSI results for MSE and PaP are obtained from our reproductions.

[3]AV-MC and LFMIM use modality-specific annotations during training; TIE and the other baselines use only multimodal target annotations.

Table 2: Performance comparison on CH-SIMS v2.0.

| Method | MAE↓ | F1↑ | Acc-2↑ | Acc-2.w↑ | Corr↑ |
|---|---|---|---|---|---|
| **Task-Specific Models** | | | | | |
| MulT [ACL19] | 0.317 | 79.59 | 79.50 | 69.61 | 0.703 |
| $\text{MAG}_{\text{BERT}}$ [ACL20] | 0.334 | 79.78 | 79.79 | 71.87 | 0.691 |
| MISA [ACMMM20] | 0.314 | 80.63 | 80.53 | 70.50 | 0.725 |
| AV-$\text{MC}^3$ [ICMI22] | 0.297 | 82.55 | 82.50 | 74.54 | 0.732 |
| CMamba [NIPS24] | 0.287 | 83.50 | 83.40 | – | 0.758 |
| FUSE-Net [CVPR26] | 0.297 | 79.53 | 79.59 | – | 0.712 |
| CaReFlow [CVPR26] | 0.277 | 82.90 | 82.90 | – | 0.745 |
| CICA [CVPR26] | 0.245 | 85.89 | 85.98 | – | 0.842 |
| **ChatGLM3-6B** | | | | | |
| MAG [ACL20] | 0.341 | 78.34 | 78.34 | 69.77 | 0.649 |
| CHFN [ACMMM22] | 0.303 | 81.00 | 81.02 | 72.57 | 0.700 |
| MSE [AAAI25] | 0.296 | 83.76 | 83.77 | 75.24 | 0.720 |
| PaP [CVPR26] | 0.264 | 84.80 | 84.75 | 76.18 | 0.769 |
| TIE [Ours] | 0.239 | 84.88 | 84.82 | 77.23 | 0.802 |
| **Qwen3.5-2B** | | | | | |
| TIE [Ours] | 0.274 | 81.44 | 81.68 | 74.08 | 0.725 |
| **Qwen3.5-9B** | | | | | |
| TIE [Ours] | 0.231 | 85.88 | 85.90 | 78.26 | 0.803 |

Table 3: Performance comparison on MELD and CHERMA.

| | MELD | | CHERMA | |
|---|---|---|---|---|
| Method | Acc↑ | WF1↑ | Acc↑ | WF1↑ |
| **Task-Specific Models** | | | | |
| MulT [ACL19] | – | – | – | 69.24 |
| UniMSE [EMNLP22] | 65.09 | 65.51 | – | – |
| $\text{LFMIM}^3$ [ACL23] | – | – | – | 70.54 |
| $\text{UniSA}_{\text{BART}}$ [ACMMM23] | 62.34 | 62.22 | – | – |
| MAGTKD [IJCAI25] | 66.36 | 65.32 | – | – |
| CMAD [ICCV25] | – | – | 72.40 | 72.20 |
| DRKF [ACMMM25] | 66.70 | 65.40 | – | – |
| **ChatGLM3-6B** | | | | |
| MAG [ACL20] | 60.38 | 59.81 | 71.20 | 71.02 |
| CHFN [ACMMM22] | 63.03 | 62.08 | 71.41 | 71.37 |
| MSE [AAAI25] | 66.23 | 65.13 | 72.90 | 72.73 |
| PaP [CVPR26] | 67.47 | 65.97 | 75.33 | 75.28 |
| TIE [Ours] | 68.35 | 67.08 | 74.41 | 74.34 |
| **Qwen3.5-2B** | | | | |
| TIE [Ours] | 66.90 | 64.84 | 72.27 | 72.29 |
| **Qwen3.5-9B** | | | | |
| TIE [Ours] | 69.07 | 67.71 | 75.44 | 75.42 |

Table 4: Org/Full comparison across compact-prompt constructors on CH-SIMS v2.0 with ChatGLM3-6B. Parameter counts include all trainable modules. TIE training uses cached contextualized text representations; its inference timing includes recomputing these representations and lexical routing online. All timings exclude shared offline acoustic/visual feature extraction; see Appendix C.5.

| Constructor | Variant | MAE↓ | Acc-2↑ | Acc-2.w↑ | Corr↑ | Trainable Params↓ | Train (ms/sample)↓ | Infer (ms/sample)↓ |
|---|---|---|---|---|---|---|---|---|
| MSE-Adapter† | Org | 0.2930 | 82.94 | 75.28 | 0.7316 | 2.681M | 26.70 | 59.60 |
| MSE-Adapter† | Full | 0.2723 | 84.56 | 77.14 | 0.7664 | 2.681M | 39.47 | 59.56 |
| PaP‡ | Org | 0.2610 | 83.37 | 73.58 | 0.7808 | 5.673M | 28.13 | 65.31 |
| PaP‡ | Full | 0.2524 | 84.18 | 75.03 | 0.7949 | 5.820M | 42.52 | 65.37 |
| TIE | Org | 0.2472 | 84.14 | 76.81 | 0.7922 | 2.540M | 31.13 | 65.36 |
| TIE | Full | 0.2393 | 84.82 | 77.23 | 0.8024 | 2.679M | 44.91 | 65.19 |

†Reproduced from the released implementation. ‡Reproduced from the paper because a complete implementation is unavailable. See Appendix C.3.

Full reduces MAE relative to Org by $7.1\%$, $3.3\%$, and $3.2\%$ for MSE-Adapter, PaP, and TIE with ChatGLM3-6B (Table 4). With Qwen3.5-9B, the reductions are $5.9\%$, $5.5\%$, and $9.8\%$ (Table 15), supporting the combined configuration across constructors and backbones.

**Migration Supervision and Ordered Prediction.** Frozen-terminal supervision yields $2.9\%$ lower mean MAE than ordinary-forward first-token cross-entropy with the same target, weight, and interface-only gradient scope (Table 20). Restricting migration gradients to the interface reduces MAE by $9.8\%$ relative to also updating evidence organization (Table 19).

On CMU-MOSI with Qwen3.5-9B, restoring the role-aware interface, order-aware objective, or migration supervision to its respective ablated configuration reduces mean MAE by $1.9\%$, $1.5\%$, or $0.8\%$; Full has the lowest mean MAE (Table 21).

**Computational Cost.** Full increases training time per sample by $44.3\%$–$51.2\%$ over Org, while measured inference latency changes by less than $0.3\%$ (Table 4).

### 4.4 Evidence-Migration Analysis

We examine text-directed predictions, evidence propagation, and predictive influence. On CH-SIMS v2.0, let $d_i = |y_{T,i} - y_i|$ and $B_{T,i} = (\widehat{y}_i - y_i)/(y_{T,i} - y_i)$, where $y_i$ and $y_{T,i}$ are the multimodal and text annotations. Here, $B_T = 0$ points to the multimodal target and $B_T = 1$ to the text target.

Figure 3 shows text-directed shifts and growing errors with modality discrepancy. TIE is below MSE-Adapter and MISA in most bins, but PaP performs better in the highest-discrepancy bin. Predictions near neutral can also yield positive $B_T$, so this pattern alone does not locate its source.

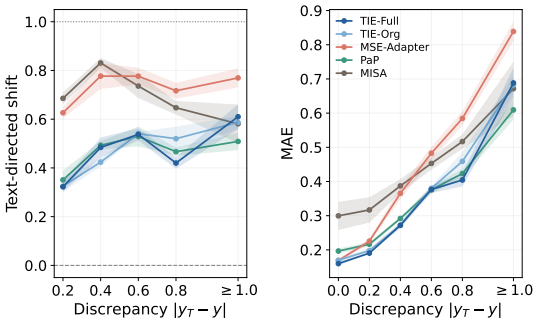



Figure 3: Text-directed behavior on CH-SIMS v2.0.

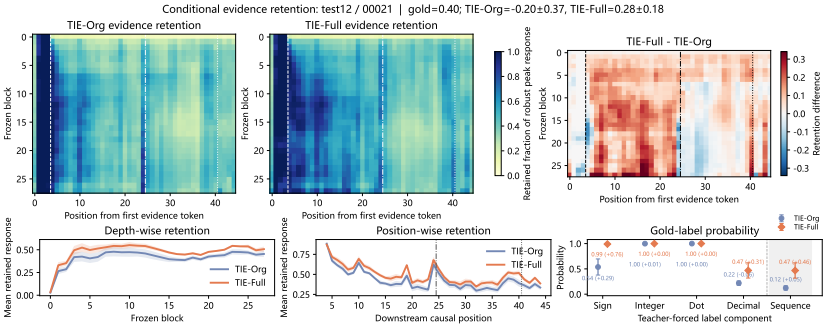



Figure 4: Conditional evidence propagation for a CH-SIMS v2.0 discrepancy case. Parenthesized gold-label probabilities show the change relative to evidence replacement.

In the illustrative case (Figure 4), Full shows more persistent normalized responses to evidence replacement and higher clean gold-label probability than Org. Parentheses give the clean-minus-replacement change (Appendix E).

On 76 polarity-conflict samples, the diagnostic terminal-head readout shows similar initial text-directed margins for Org and Full, followed by positive middle-layer margins. Their final margins are $0.158$ and $-0.250$, respectively (Figure 7).

On conflict samples, attenuating Full's middle-layer transcript contributions increases the correct-direction sign margin by $0.2009$, whereas attenuating evidence contributions decreases it by $0.1573$; matched whole-output and random controls do not reproduce this pattern (Table 22). Evidence replacement also lowers gold-sign probability by $18.11$ percentage points for Org and $23.70$ for Full (Appendix E.4).

Text-directed contributions thus coexist with evidence supporting the multimodal target. These are source-position effects, and the margin changes rarely flip the predicted sign.

## 5 Conclusion

We presented TIE, which connects text-bounded evidence organization, role-aware migration, and frozen-model migration supervision for multimodal affective prediction. An order-aware objective further connects label generation with sentiment estimation. Experiments show gains across multiple benchmarks and compact-prompt constructors. Propagation and intervention analyses indicate that text-directed contributions can coexist with evidence supporting multimodal targets, motivating the coordinated treatment of evidence construction, migration, and predictive influence.

## REFERENCES

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1208. URL https://aclanthology.org/P18-1208/.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 59–66, 2018. doi: 10.1109/FG.2018.00019. URL https://doi.org/10.1109/FG.2018.00019.

Linqing Chen, Hanmeng Zhong, Wentao Wu, and Peng Zhou. The dominance of text space: Unveiling the asymmetric nature of cross-modal alignment in large language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 36666–36678, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1699. URL https://aclanthology.org/2026.acl-long.1699/.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarep — a collaborative voice analysis repository for speech technologies. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 960–964, 2014. doi: 10.1109/ICASSP.2014.6853739. URL https://doi.org/10.1109/ICASSP.2014.6853739.

Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE MultiMedia*, 19(3):34–41, July 2012. ISSN 1070-986X. doi: 10.1109/MMUL.2012.26. URL https://doi.org/10.1109/MMUL.2012.26.

Constantin Eichenberg, Sidney Black, Samuel Weinbach, Letitia Parcalabescu, and Anette Frank. MAGMA – multimodal augmentation of generative models through adapter-based finetuning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2416–2428, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.179. URL https://aclanthology.org/2022.findings-emnlp.179/.

Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pp. 1459–1462, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605589336. doi: 10.1145/1873951.1874246. URL https://doi.org/10.1145/1873951.1874246.

Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, and Mang Ye. Emoe: Modality-specific enhanced dynamic emotion experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fang_EMOE_Modality-Specific_Enhanced_Dynamic_Emotion_Experts_CVPR_2025_paper.html.

Zixian Gao, Disen Hu, Xun Jiang, Huimin Lu, Heng Tao Shen, and Xing Xu. Enhanced experts with uncertainty-aware routing for multimodal sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 9650–9659, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3680949. URL https://doi.org/10.1145/3664647.3680949.

Jiwei Guo, Jiajia Tang, Weichen Dai, Yu Ding, and Wanzeng Kong. Dynamically adjust word representations using unaligned multimodal information. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, pp. 3394–3402, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392037. doi: 10.1145/3503161.3548137. URL https://doi.org/10.1145/3503161.3548137.

Zirun Guo, Tao Jin, Jingyuan Chen, and Zhou Zhao. Classifier-guided gradient modulation for enhanced multimodal learning. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385. URL https://openreview.net/forum?id=oe5ZEqTOaz.

Zhiyuan Han, Beier Zhu, Wenwen Tong, Pengyang Shao, Peipei Song, Xinyi Wang, Jiangnan Chen, Lewei Lu, and Xun Yang. Omni-perception policy optimization for multimodal emotion reasoning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=4dwe2wlxx5.

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*, MM '20, pp. 1122–1131, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379885. doi: 10.1145/3394171.3413678. URL https://doi.org/10.1145/3394171.3413678.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Kang He, Yuzhe Ding, Xinrong Wang, Fei Li, Chong Teng, and Donghong Ji. Enhance-then-balance modality collaboration for robust multimodal sentiment analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30183–30193, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/He_Enhance-then-Balance_Modality_Collaboration_for_Robust_Multimodal_Sentiment_Analysis_CVPR_2026_paper.html.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 7837–7851, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.534. URL https://aclanthology.org/2022.emnlp-main.534/.

Zhongquan Jian, Wenhan Lv, Yanhao Chen, Guanran Luo, Wentao Qiu, Shaopan Wang, Bingbing Hu, and Qingqiang Wu. MDF: A modality-aware disentanglement and fusion framework for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31292–31300, 2026. doi: 10.1609/aaai.v40i37.40392. URL https://doi.org/10.1609/aaai.v40i37.40392.

Haoyu Jiang, Xiaoliang Chen, Duoqian Miao, Xiaolin Qin, Xianyong Li, and Yajun Du. CICA: Coupling confidence-aware pretraining with confidence-informed attention for robust multimodal sentiment analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 37842–37851, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Jiang_CICA_Coupling_Confidence-Aware_Pretraining_with_Confidence-Informed_Attention_for_Robust_Multimodal_CVPR_2026_paper.html.

Peiyuan Jiang, Yao Liu, Qiao Liu, Zongshun Zhang, Jiaye Yang, Lu Liu, and Daibing Yao. Drkf: Decoupled representations with knowledge fusion for multimodal emotion recognition. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 955–964, New York, NY, USA, 2025a. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3754758. URL https://doi.org/10.1145/3746027.3754758.

Shenjie Jiang, Zhuoyu Wang, Xuecheng Wu, Hongru Ji, Mingxin Li, Xianghua Li, and Chao Gao. Ddse: A decoupled dual-stream enhanced framework for multimodal sentiment analysis with text-centric ssm. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 5893–5902, New York, NY, USA, 2025b. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3754817. URL https://doi.org/10.1145/3746027.3754817.

Bozhou Li, Xinda Xue, Sihan Yang, Yang Shi, Xinlong Chen, Yushuo Guan, Yuanxing Zhang, and Wentao Zhang. The unseen bias: How norm discrepancy in pre-norm MLLMs leads to visual information loss. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=GVVNG2EMQv.

Jie Li, Shifei Ding, Lili Guo, and Xuan Li. Multi-modal anchor gated transformer with knowledge distillation for emotion recognition in conversation. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 8141–8149. International Joint Conferences on Artificial Intelligence Organization, 8 2025a. doi: 10.24963/ijcai.2025/905. URL https://doi.org/10.24963/ijcai.2025/905. Main Track.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 23–29 Jul 2023a. URL https://proceedings.mlr.press/v202/li23q.html.

Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2852–2861, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Li_Reliable_Crowdsourcing_and_CVPR_2017_paper.html.

Wenbing Li, Hang Zhou, Junqing Yu, Zikai Song, and Wei Yang. Coupled mamba: Enhanced multimodal fusion with coupled state space model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=UXEo3uNNIX.

Yan Li, Yifei Xing, Xiangyuan Lan, Xin Li, Haifeng Chen, and Dongmei Jiang. AlignMamba: Enhancing Multimodal Mamba with Local and Global Cross-modal Alignment. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24774–24784, Los Alamitos, CA, USA, June 2025b. IEEE Computer Society. doi: 10.1109/CVPR52734.2025.02307. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52734.2025.02307.

Zaijing Li, Ting-En Lin, Yuchuan Wu, Meng Liu, Fengxiao Tang, Ming Zhao, and Yongbin Li. Unisa: Unified generative framework for sentiment analysis. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, pp. 6132–6142, New York, NY, USA, 2023b. Association for Computing Machinery. ISBN 9798400701085. doi: 10.1145/3581783.3612336. URL https://doi.org/10.1145/3581783.3612336.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35799–35819. PMLR, 13–19 Jul 2025c. URL https://proceedings.mlr.press/v267/li25ca.html.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, Jiangyan Yi, and Jianhua Tao. AffectGPT: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=xmbdACI0xu.

Tian Liang, Jing Huang, Ming Kong, Luyuan Chen, and Qiang Zhu. Querying as Prompt: Parameter-Efficient Learning for Multimodal Language Model . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26845–26855, Los Alamitos, CA, USA, June 2024. IEEE Computer Society. doi: 10.1109/CVPR52733.2024.02536. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52733.2024.02536.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 61–68, Dublin, Ireland, May 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.8. URL https://aclanthology.org/2022.acl-short.8/.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *AI Open*, 5:208–215, 2024. ISSN 2666-6510. doi: https://doi.org/10.1016/j.aiopen.2023.08.012. URL https://www.sciencedirect.com/science/article/pii/S2666651023000141.

Yadong Liu and Shangfei Wang. FINE: Factorized multimodal sentiment analysis via mutual information estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 1955–1963, 2026. doi: 10.1609/aaai.v40i3.37176. URL https://doi.org/10.1609/aaai.v40i3.37176.

Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. Make acoustic and visual cues matter: Ch-sims v2.0 dataset and av-mixup consistent module. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, ICMI '22, pp. 247–258, New York, NY, USA, 2022b. Association for Computing Machinery. ISBN 9781450393904. doi: 10.1145/3536221.3556630. URL https://doi.org/10.1145/3536221.3556630.

Yuanyi Luo, Wei Liu, Qiang Sun, Sirui Li, Jichunyang Li, Rui Wu, and Xianglong Tang. TriagedMSA: Triaging Sentimental Disagreement in Multimodal Sentiment Analysis . *IEEE Transactions on Affective Computing*, 16(03):1557–1569, July 2025. ISSN 1949-3045. doi: 10.1109/TAFFC.2024.3524789. URL https://doi.ieeecomputersociety.org/10.1109/TAFFC.2024.3524789.

Xiongjian Lv, Yimin Wen, and Hang Yu. Diffufuse: Diffusion-driven dual-stream fusion framework for multimodal sentiment analysis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 8458–8467, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3755508. URL https://doi.org/10.1145/3746027.3755508.

Sijie Mai and Shiqin Han. CaReFlow: Cyclic adaptive rectified flow for multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 37799–37809, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Mai_CaReFlow_Cyclic_Adaptive_Rectified_Flow_for_Multimodal_Fusion_CVPR_2026_paper.html.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. *SciPy 2015*, 2015. doi: 10.25080/Majora-7b98e3ed-003. URL https://doi.org/10.25080/Majora-7b98e3ed-003.

Chunlei Meng, Jiabin Luo, Zhenglin Yan, Zhenyu Yu, Rong Fu, Zhongxue Gan, and Chun Ouyang. Tri-subspaces disentanglement for multimodal sentiment analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8791–8800, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Meng_Tri-Subspaces_Disentanglement_for_Multimodal_Sentiment_Analysis_CVPR_2026_paper.html.

Cam-Van Thi Nguyen, The-Son Le, Anh-Tuan Mai, and Duc-Trong Le. Ada2i: Enhancing modality balance for multimodal conversational emotion recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 9330–9339, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3681648. URL https://doi.org/10.1145/3664647.3681648.

Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In Xianghua Xie, Mark W. Jones, and Gary K. L. Tam (eds.), *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 41.1–41.12. BMVA Press, September 2015. ISBN 1-901725-53-7. doi: 10.5244/C.29.41. URL https://dx.doi.org/10.5244/C.29.41.

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. URL https://aclanthology.org/P19-1050/.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2359–2369, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.214. URL https://aclanthology.org/2020.acl-main.214/.

Dongning Rao, Yunbiao Zeng, Zhihua Jiang, and Jujian Lv. A text-routed sparse mixture-of-experts model with explanation and temporal alignment for multi-modal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 32799–32807, 2026. doi: 10.1609/aaai.v40i39.40559. URL https://doi.org/10.1609/aaai.v40i39.40559.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised Pre-Training for Speech Recognition. In *Interspeech 2019*, pp. 3465–3469, 2019. doi: 10.21437/Interspeech.2019-1873. URL https://www.isca-archive.org/interspeech_2019/schneider19_interspeech.html.

Jun Sun, Shoukang Han, Yu-Ping Ruan, Xiaoning Zhang, Shu-Kai Zheng, Yulong Liu, Yuxin Huang, and Taihao Li. Layer-wise fusion with modality independence modeling for multi-modal emotion recognition. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 658–670, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.39. URL https://aclanthology.org/2023.acl-long.39/.

Kaiwei Sun and Mi Tian. Sequential fusion of text-close and text-far representations for multimodal sentiment analysis. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 40–49, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.4/.

Yueru Sun, Yimeng Zhang, Haoyu Gu, Nuo Chen, Dong She, Xianrong Yao, Yang Gao, and Zhanpeng Jin. EmoMM: Benchmarking and steering MLLM for multimodal emotion recognition under conflict and missingness. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 20351–20371, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1018. URL https://aclanthology.org/2026.findings-acl.1018/.

Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/tan19a.html.

Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, pp. 3927–3935, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450386517. doi: 10.1145/3474085.3475587. URL https://doi.org/10.1145/3474085.3475587.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Jingyu Sun, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024. URL https://arxiv.org/abs/2406.12793.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1656. URL https://aclanthology.org/P19-1656/.

Pan Wang, Qiang Zhou, Yawen Wu, Tianlong Chen, and Jingtong Hu. Dlf: disentangled-language-focused multimodal sentiment analysis. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i20.35416. URL https://doi.org/10.1609/aaai.v39i20.35416.

Pan Wang, Lipeng Ke, Huajun Ying, Pritish Mohapatra, Rohan Sarkar, Suresh Lakhani, Sankar Venkataraman, and Jingtong Hu. ConsMSA: Semantic distribution consistency learning for multimodal sentiment analysis. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=rWTqAgyZcd.

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 (CVPR)*, pp. 15880–15889, June 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wang_Conflict-Aware_Adaptive_Cross-Reconstruction_for_Multimodal_Sentiment_Analysis_CVPR_2026_paper.html.

Huyu Wu, Meng Tang, Xinhan Zheng, and Haiyun Jiang. When language overrules: Revealing text dominance in multimodal large language models, 2025. URL https://arxiv.org/abs/2508.10552.

Xiongye Xiao, Gengshuo Liu, Gaurav Gupta, Defu Cao, Shixuan Li, Yaxing Li, Tianqing Fang, Mingxi Cheng, and Paul Bogdan. Neuro-inspired information-theoretic hierarchical perception for multimodal learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Z9AZsU1Tju.

Yang Yang, Xunde Dong, and Yupeng Qiang. Mse-adapter: A lightweight plugin endowing llms with the capability to perform multimodal sentiment analysis and emotion recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25642–25650, Apr. 2025a. doi: 10.1609/aaai.v39i24.34755. URL https://ojs.aaai.org/index.php/AAAI/article/view/34755.

Yang Yang, Hongpeng Pan, Qing-Yuan Jiang, Yi Xu, and Jinhui Tang. Learning to rebalance multi-modal optimization by adaptively masking subnetworks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(6):4553–4566, June 2025b. ISSN 0162-8828. doi: 10.1109/TPAMI.2025.3547417. URL https://doi.org/10.1109/TPAMI.2025.3547417.

Zhilu Yang and Mingcheng Li. Factorize, reconstruct, enhance: A unified framework for multimodal sentiment analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15860–15869, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/

Yang_Factorize_Reconstruct_Enhance_A_Unified_Framework_for_Multimodal_Sentiment_Analysis_CVPR_2026_paper.html.

Feng Yu, Xiangyu Wu, Yang Yang, and Jianfeng Lu. Multimodal classification via total correlation maximization. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=MbQhdzAhSl.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10790–10797, 2021. URL https://doi.org/10.1609/aaai.v35i12.17289.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31 (6):82–88, November 2016. ISSN 1541-1672. doi: 10.1109/MIS.2016.94. URL https://doi.org/10.1109/MIS.2016.94.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, Di Wu, and Zhendong Peng. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6182–6186, 2022. doi: 10.1109/ICASSP43922.2022.9746682. URL https://doi.org/10.1109/ICASSP43922.2022.9746682.

Chengsheng Zhang, Chenghao Sun, Zhining Xie, and Xinmei Tian. Interpreting and enhancing emotional circuits in large vision-language models via cross-modal information flow. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=wnKgWnx8mH.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. doi: 10.1109/LSP.2016.2603342. URL https://doi.org/10.1109/LSP.2016.2603342.

Zhuoran Zhang, Tengyue Wang, Xilin Gong, Yang Shi, Haotian Wang, Di Wang, and Lijie Hu. Beyond fixed biases: Decoding the role of reasoning uncertainty in MLLM modality conflicts. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=CQtHecicN0.

Haozhe Zhao, Shuzheng Si, Liang Chen, Yichi Zhang, Maosong Sun, Baobao Chang, and Minjia Zhang. Looking beyond text: Reducing language bias in large vision-language models via multimodal dual-attention and soft-image guidance. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19666–19690, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.995. URL https://aclanthology.org/2025.emnlp-main.995/.

Xianbing Zhao, Lan Luo, Hengyang Lu, and Buzhou Tang. Prototype-as-prompt: Multimodal sentiment prototypes endowing large language models the capability to perform multimodal sentiment analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23010–23020, June 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhao_Prototype-as-Prompt_Multimodal_Sentiment_Prototypes_Endowing_Large_Language_Models_the_Capability_CVPR_2026_paper.html.

Xianbing Zhao, Shengzun Yang, and Buzhou Tang. Ccaf: Coarse-to-fine cross-modal alignment and fusion for multimodal sentiment analysis. In *Proceedings of the ACM Web Conference 2026*, WWW '26, pp. 7421–7430, New York, NY, USA, 2026b. Association for Computing Machinery. ISBN 9798400723070. doi: 10.1145/3774904.3792569. URL https://doi.org/10.1145/3774904.3792569.

Yan Zhuang, Minhao Liu, Wei Bai, Yanru Zhang, Xiaoyue Zhang, Jiawen Deng, and Fuji Ren. CMAD: Correlation-aware and modalities-aware distillation for multimodal sentiment analysis with missing modalities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4626–4636, October 2025. doi: 10.1109/ICCV51701.2025.00440. URL https://doi.org/10.1109/ICCV51701.2025.00440.

# APPENDIX CONTENTS

**A Method Details and Derivations** **19**

A.1 Text-Bounded Routing and Reference Construction . . . . . 19

A.2 Role-Conditioned Evidence Resampling . . . . . 20

A.3 Migration Interface and Supervision Details . . . . . 21

A.4 Order-Aware Objective Details . . . . . 21

**B Datasets, Evaluation Metrics, and Baselines** **23**

B.1 Datasets . . . . . 23

B.2 Evaluation Metrics . . . . . 23

B.3 Baselines . . . . . 24

**C Reproducibility Details** **26**

C.1 Feature Extraction . . . . . 26

C.2 Prompts and Label Generation . . . . . 27

C.3 Baseline Implementations . . . . . 27

C.4 Implementation and Hyperparameters . . . . . 28

C.5 Efficiency Measurement Protocol . . . . . 29

**D Additional Experimental Results** **30**

D.1 Temporal Routing and Migration Context . . . . . 30

D.2 Controlled Comparisons with Qwen3.5 Backbones . . . . . 30

D.3 Additional Component Ablations . . . . . 31

**E Evidence-Migration Diagnostics** **32**

E.1 Text-Directed Behavior . . . . . 33

E.2 Layerwise Text-Directed Readout . . . . . 33

E.3 Conditional Evidence Propagation and Gold-Label Support . . . . . 34

E.4 Source-Path Intervention and Predictive Influence . . . . . 36

**F AI Use Statement** **37**

# A METHOD DETAILS AND DERIVATIONS

## A.1 TEXT-BOUNDED ROUTING AND REFERENCE CONSTRUCTION

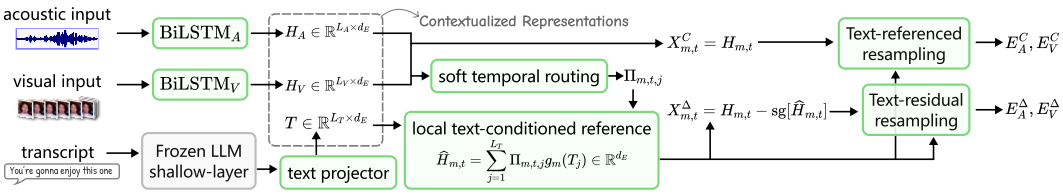



Figure 5: Text-bounded evidence organization. Soft routing supplies token weights used with projected text features to construct a local reference for each acoustic or visual frame. Complete modality representations and their reference residuals are resampled into four evidence vectors.

For acoustic and visual input sequences $X_m \in \mathbb{R}^{L_m \times d_m}$, $m \in \{A, V\}$, where $L_m$ is the number of modality frames, independent bidirectional LSTMs produce temporally contextualized states in the $d_E$-dimensional evidence space:

$$H_m = \mathrm{LN}\left(W_m^E \,\mathrm{BiLSTM}_m(X_m) + b_m^E\right) \in \mathbb{R}^{L_m \times d_E}. \tag{12}$$

Separately, let $Z \in \mathbb{R}^{L_T \times d_L}$ denote the cached contextualized text representations produced by a shallow layer of the frozen LLM from the transcript. The trainable text projector is

$$T = W_{\mathrm{up}}\,\mathrm{GELU}(W_{\mathrm{down}}\,\mathrm{LN}(Z) + b_{\mathrm{down}}) + b_{\mathrm{up}} \in \mathbb{R}^{L_T \times d_E}, \tag{13}$$

where $W_{\mathrm{down}}$ maps to a bottleneck of dimension $d_R$, and $W_{\mathrm{up}}$ maps to the $d_E$ evidence space. The frozen text representations $Z$ receive no gradient.

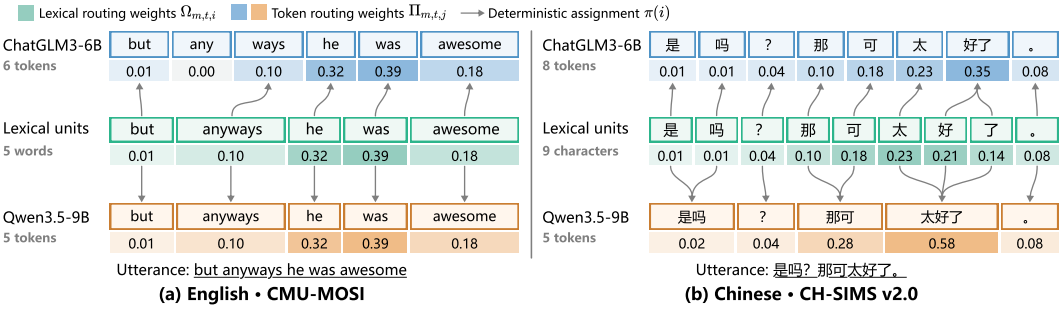



Figure 6: Illustration of lexical-to-token routing for English and Chinese utterances. Each lexical unit is assigned to its last overlapping LLM token; token weights aggregate the assigned lexical routing weights.

Subword tokenization does not necessarily follow lexical boundaries. Figure 6(a) illustrates English lexical units, defined as words (including internal apostrophes) or individual punctuation marks; Figure 6(b) shows the corresponding Chinese units, defined as non-whitespace characters. Character offsets deterministically assign each unit $i$ to its last overlapping LLM-token position $\pi(i)$; the assignment is not learned. For modality frame $t$, its normalized temporal center is

$$\tau_{m,t} = \frac{t - \frac{1}{2}}{L_m} \in (0, 1), \qquad t = 1, \ldots, L_m. \tag{14}$$

A lightweight geometry function conditioned on the modality state predicts a center displacement and routing width:

$$[\delta_{m,t}, \eta_{m,t}] = W_m^G H_{m,t} + b_m^G \in \mathbb{R}^2, \qquad \mu_{m,t} = \sigma(\mathrm{logit}(\tau_{m,t}) + \delta_{m,t}) \in (0, 1), \tag{15}$$

$$w_{m,t} = \frac{\exp(\mathrm{clip}(\eta_{m,t}, -6, 6))}{L_X} \in \left[\frac{e^{-6}}{L_X}, \frac{e^{6}}{L_X}\right], \tag{16}$$

where $L_X$ is the number of valid lexical units. For lexical center $\xi_i = (i - \frac{1}{2})/L_X$, $i = 1, \ldots, L_X$, the routing field is

$$\Omega_{m,t,i} = \mathrm{softmax}_i \left[ -\frac{1}{2} \left( \frac{\xi_i - \mu_{m,t}}{w_{m,t}} \right)^2 \right] \in (0,1), \qquad \sum_{i=1}^{L_X} \Omega_{m,t,i} = 1. \tag{17}$$

Lexical units assigned to the same LLM token are aggregated and renormalized:

$$\Pi_{m,t,j} = \frac{\sum_{i:\pi(i)=j} \Omega_{m,t,i}}{\sum_{j'=1}^{L_T} \sum_{i:\pi(i)=j'} \Omega_{m,t,i}} \in [0,1], \qquad \sum_{j=1}^{L_T} \Pi_{m,t,j} = 1. \tag{18}$$

The normalized coordinates initialize a temporal reference, while learned offsets and widths provide modality-conditioned soft routing rather than timestamp-level alignment or a guaranteed monotonic mapping. Routing mass remains conserved for each valid frame.

These routing weights form a local text-conditioned reference for each modality frame:

$$\widehat{H}_{m,t} = \sum_{j=1}^{L_T} \Pi_{m,t,j} g_m(T_j) \in \mathbb{R}^{d_E},$$

where $g_m : \mathbb{R}^{d_E} \to \mathbb{R}^{d_E}$ is modality-specific.

The local text-conditioned reference is optimized with the masked Huber loss

$$\mathcal{L}_{\mathrm{ref}} = \frac{1}{2} \sum_{m \in \{A,V\}} \Big\langle \mathrm{Huber}\Big( \widehat{H}_{m,t}, \mathrm{sg}[H_{m,t}] \Big) \Big\rangle_{M_m}, \tag{19}$$

where $\langle \cdot \rangle_{M_m}$ denotes a masked mean over valid frames of modality $m$, and sg denotes stop-gradient. The target is detached, while routing and modality encoding are optimized through $\widehat{H}_{m,t}$.

### A.2 Role-Conditioned Evidence Resampling

Recall that $X^C_{m,t} = H_{m,t} \in \mathbb{R}^{d_E}$ and $X^\Delta_{m,t} = H_{m,t} - \mathrm{sg}[\widehat{H}_{m,t}] \in \mathbb{R}^{d_E}$. For role $r \in \{C, \Delta\}$, we first construct a relation representation

$$G^r_{m,t} = \mathrm{GELU}\Big( W_X X^r_{m,t} + W_B \widehat{H}_{m,t} + W_I (X^r_{m,t} \odot \widehat{H}_{m,t}) + b_G \Big) \in \mathbb{R}^{d_E}. \tag{20}$$

The source $X^r_{m,t}$ supplies the attention values, while $G^r_{m,t}$ conditions key construction on the source, its local reference, and their element-wise relation. A learned query $q^r_m$ resamples each modality-role sequence. For attention head $h$,

$$Q^h_{m,r} = W^h_Q q^r_m, \quad K^h_{m,r,t} = W^h_K G^r_{m,t}, \quad V^h_{m,r,t} = W^h_V X^r_{m,t}, \tag{21}$$

and the valid-frame mask $M_{m,t}$ is applied before the temporal softmax:

$$\alpha^h_{m,r,t} = \mathrm{softmax}_{t:M_{m,t}=1} \left( \frac{Q^h_{m,r} (K^h_{m,r,t})^\top}{\sqrt{d_h}} \right), \qquad E^r_m = W_O \left[ \Big\Vert_{h=1}^{H_q} \sum_t \alpha^h_{m,r,t} V^h_{m,r,t} \right]. \tag{22}$$

We use $H_q = 4$ heads with $d_h = d_E / H_q$. The four modality-role outputs form

$$\mathcal{E} = [E^C_A, E^C_V, E^\Delta_A, E^\Delta_V] \in \mathbb{R}^{4 \times d_E}. \tag{23}$$

**Interpretation of evidence roles.** The two sources are overlapping views rather than a complementary decomposition. The residual may contain complementary affect, temporal mismatch, or noise; the relevance of each to prediction depends on the task. The roles are defined by their relation and value sources rather than by a fixed orthogonality or separation constraint. The text-referenced and text-residual design defines two conditioned resampling pathways rather than an additive or orthogonal decomposition of $H_m$. In particular, $X^C_m = H_m$ preserves the complete modality representation, which is related to the local text-conditioned reference when forming $G^C_m$ and when

resampled by the text-referenced-role query. The source $X_m^{\Delta} = H_m - \mathrm{sg}[\widehat{H}_m]$ represents the deviation from that modality-conditioned textual reference. The role names describe these distinct organization operations; they do not assert that the first source contains only text-supported information, that the residual is necessarily useful, or that the two resulting tokens are statistically independent.

SFTTR (Sun & Tian, 2025) combines contrastively learned text-close/text-far projections with sequential cross-modal fusion and prediction losses. TIE instead resamples complete modality states and residuals relative to modality-conditioned textual references, and trains the migration interface with supervision from the frozen LLM's terminal prediction block.

### A.3 Migration Interface and Supervision Details

For the migration interface, $W_S \in \mathbb{R}^{d_L \times d_E}$ is shared by all evidence roles, while $V_r \in \mathbb{R}^{d_M \times d_E}$ and $U_r \in \mathbb{R}^{d_L \times d_M}$ define the role-specific low-rank residual. Acoustic and visual evidence share the same residual transformation for a given role. Each $U_r$ is initialized to zero, so the interface begins as a shared projector and learns role-specific increments during training.

The compact terminal computation preserves the absolute positions occupied by the four evidence states and the decision state in the original sequence. For rotary-position backbones, their original position indices are used to construct the corresponding rotary embeddings; the five-state computation does not renumber the decision state as the fifth sequence position. A causal mask retains the evidence-to-decision ordering, and no key–value cache from the full forward is reused. The same positional and causal semantics are applied to the Qwen3.5 and ChatGLM3 implementations.

Although $h_D^{L-1}$ carries the joint predictive context accumulated before the terminal block, it is detached in $H_{\mathrm{mig}}^{L-1}$. Consequently, a prediction made solely through the decision state's residual path yields no migration-loss gradient to $\theta_{\mathrm{mig}}$. The interface receives an auxiliary gradient only through terminal-block interactions from the evidence states to the decision output:

$$\nabla_{\theta_{\mathrm{mig}}} \mathcal{L}_{\mathrm{mig}} = \frac{\partial \mathcal{L}_{\mathrm{mig}}}{\partial H_{\mathcal{E}}^{L-1}} \frac{\partial H_{\mathcal{E}}^{L-1}}{\partial \theta_{\mathrm{mig}}}, \qquad \frac{\partial h_D^{L-1}}{\partial \theta_{\mathrm{mig}}} = 0 \quad \text{on the auxiliary path.} \tag{24}$$

The decision state therefore supplies sample-conditioned context without becoming an independently trainable shortcut. This does not by itself quantify the causal contribution of each evidence token; the intervention analyses separately measure their conditional propagation and decoding response.

The supervised decision represented by $y_1$ differs by task. In MSA, it is the leading sign token and constrains polarity for nonzero targets rather than the full sentiment magnitude; zero is serialized as $+0.0$ for a unique fixed-length representation. Magnitude and complete-sequence supervision are supplied by the order-aware and token objectives on the normal prediction path. In ERC, the emotion category is represented by a single token, so the auxiliary target covers the complete categorical decision. This auxiliary computation requires an additional frozen final-block pass during training and adds no inference-time module.

### A.4 Order-Aware Objective Details

Let $\mathcal{V}_Y = \{v_1 < \cdots < v_K\}$ be the uniformly spaced sentiment grid introduced in the main text, and let $\mathrm{tok}(v_k) = (\mathrm{tok}_1(v_k), \ldots, \mathrm{tok}_{L_Y}(v_k))$ denote the serialized token sequence of $v_k$. Under each evaluated tokenizer, all valid MSA values share the same four-token effective serialization. They can therefore be evaluated at the same label positions without length-dependent padding or length correction.

The standard teacher-forcing forward conditions position $\ell$ on the gold prefix $y_{<\ell}$. At these positions, we collect the log-probabilities assigned to the tokens of each sentiment value and define

$$s_k = \sum_{\ell=1}^{L_Y} \log p\left(\mathrm{tok}_\ell(v_k) \mid \mathcal{I}, y_{<\ell}\right), \qquad \widehat{p}_k = \frac{\exp(s_k)}{\sum_{j=1}^{K} \exp(s_j)} \in (0,1), \qquad \sum_{k=1}^{K} \widehat{p}_k = 1. \tag{25}$$

Thus, $s = (s_1, \ldots, s_K) \in \mathbb{R}^K$ is a collection of teacher-forced value scores obtained from the normal prediction forward, and $\widehat{p} = (\widehat{p}_1, \ldots, \widehat{p}_K)$ is their normalized ordinal surrogate. It is not

the autoregressive predictive distribution over complete sentiment strings: the scores reuse logits conditioned on the same gold prefix at each label position rather than separate rollouts conditioned on $\mathrm{tok}_{<\ell}(v_k)$.

For a scalar annotation $y$, let $k_y = \arg\min_k |v_k - y|$ identify its nearest valid sentiment value. The target is a point mass at $k_y$. Its cumulative distribution and that of $\widehat{p}$ are

$$\mathrm{CDF}_{\widehat{p}}(v_k) = \sum_{j=1}^{k} \widehat{p}_j, \qquad \mathrm{CDF}_y(v_k) = \mathbf{1}[k \geq k_y]. \tag{26}$$

Here, $\mathrm{CDF}_y$ is the cumulative distribution of the point mass induced by $y$ at index $k_y$. Using $\mathrm{EMD}(\widehat{p}, y)$ as shorthand for the one-dimensional Earth Mover's Distance between $\widehat{p}$ and this point target, the mean CDF discrepancy, equivalently the EMD scaled by $1/K$, is

$$\mathcal{L}_{\mathrm{ord}} = \frac{1}{K}\mathrm{EMD}(\widehat{p}, y) = \frac{1}{K}\sum_{k=1}^{K} |\mathrm{CDF}_{\widehat{p}}(v_k) - \mathrm{CDF}_y(v_k)| = \frac{1}{K}\sum_{j=1}^{K} \widehat{p}_j |j - k_y|. \tag{27}$$

The final equality can be seen by splitting the cumulative-distance sum around $k_y$:

$$\sum_{k=1}^{K} |\mathrm{CDF}_{\widehat{p}}(v_k) - \mathrm{CDF}_y(v_k)| = \sum_{k<k_y} \mathrm{CDF}_{\widehat{p}}(v_k) + \sum_{k\geq k_y} (1 - \mathrm{CDF}_{\widehat{p}}(v_k)) = \sum_{j=1}^{K} \widehat{p}_j |j - k_y|. \tag{28}$$

Equivalently, transporting probability mass from index $j$ to the point target at $k_y$ incurs the one-dimensional distance $|j - k_y|$. Because the sentiment grid is uniformly spaced, index distance is proportional to numerical sentiment distance and therefore preserves its ordering up to a dataset-specific constant. The factor $1/K$ averages the discrepancy over the $K$ cumulative positions; it scales the EMD but does not force its maximum to equal one.

The objective introduces only score aggregation, normalization, and cumulative-distribution operations. Its gradient passes through the frozen LLM computation to the trainable evidence-organization and migration modules, while the LLM parameters remain fixed. It requires no additional LLM forward and adds no inference-time computation. ERC does not use this objective because its emotion classes have no predefined ordinal relation.

**Constrained autoregressive comparison.** The default objective above uses teacher-forced value scores. Our strict variant uses each value's own prefix and normalizes next-token probabilities over its valid continuations. For a prefix $b$, let $\mathcal{A}(b)$ contain the next tokens that can extend $b$ to a valid sentiment value. We define

$$p_{\mathrm{valid}}(a \mid \mathcal{I}, b) = \frac{p(a \mid \mathcal{I}, b)}{\sum_{u \in \mathcal{A}(b)} p(u \mid \mathcal{I}, b)}, \qquad a \in \mathcal{A}(b). \tag{29}$$

The resulting value score is

$$s_k^{\mathrm{strict}} = \sum_{\ell=1}^{L_Y} \log p_{\mathrm{valid}}\left(\mathrm{tok}_\ell(v_k) \mid \mathcal{I}, \mathrm{tok}_{<\ell}(v_k)\right). \tag{30}$$

We implement this variant with a shared-prefix tree that shares branching-prefix scores across values. A prefix with a single valid continuation contributes zero log-probability under $p_{\mathrm{valid}}$ and needs no separate scoring forward; its token remains in the full prefix supplied to later branching evaluations. This score is the log-probability of a constrained autoregressive path, not the original full-vocabulary sequence log-likelihood. The comparison therefore changes both prefix conditioning and local normalization. Its branching states require additional frozen-LLM computation that grows with the valid label set, whereas the default surrogate reuses one standard forward. Table 5 evaluates this trade-off on CMU-MOSEI with ChatGLM3-6B.

Including final test evaluation, the corresponding times are $4.242$ and $26.634$ hours per seed. Strict scoring improves the MAE and Corr while increasing end-to-end training and evaluation time to approximately $6.3\times$ that of the teacher-forced setting. We therefore retain teacher-forced scoring as the default efficiency–performance trade-off. Neither objective changes inference-time decoding or cost.

Table 5: Teacher-forced and constrained autoregressive order-aware scoring on CMU-MOSEI with ChatGLM3-6B. Training times exclude final test evaluation. Parenthesized percentages are relative to Teacher-forced; negative values indicate degradation.

| Scoring | MAE↓ | Acc-2↑ | Corr↑ | Training time↓ | |
|---|---|---|---|---|---|
| | | | | Per seed (h) | Per epoch (min) |
| Teacher-forced | $0.4894_{\pm.0076}$ | $88.21_{\pm0.30}$ | $0.8071_{\pm.0035}$ | $3.946_{\pm0.401}$ | $14.09_{\pm0.03}$ |
| Strict (constrained) | $0.4799_{(+1.9\%)\pm.0054}$ | $88.12_{(-0.1\%)\pm0.51}$ | $0.8185_{(+1.4\%)\pm.0049}$ | $26.341_{(-567.5\%)\pm4.103}$ | $102.65_{(-628.5\%)\pm0.20}$ |

# B Datasets, Evaluation Metrics, and Baselines

## B.1 Datasets

We evaluate TIE on three MSA benchmarks and two ERC benchmarks spanning English and Chinese. We follow the standard training, validation, and test splits used in prior work, as summarized in Table 6.

**CMU-MOSI** contains 2,199 short opinion segments drawn from 93 English video blogs on YouTube featuring 89 distinct speakers. The videos cover diverse topics, including movies, books, and products. Each segment was rated for sentiment intensity by five annotators, and the ratings were averaged to obtain its final annotation (Zadeh et al., 2016).

**CMU-MOSEI** is a large-scale collection of English monologue videos from YouTube, covering more than 1,000 online speakers and 250 different topics over approximately 66 hours of recordings. Each segment was rated by three annotators for sentiment and six Ekman emotions; our experiments use only the sentiment annotations (Bagher Zadeh et al., 2018).

**CH-SIMS v2.0** contains 4,403 labeled Chinese video clips collected from 11 diverse scenarios, with an emphasis on expressive acoustic and visual behaviors. Each clip provides a multimodal sentiment annotation together with textual, acoustic, and visual annotations obtained under modality-isolated conditions. TIE uses only the multimodal annotations: modality-specific annotations are never provided to the model during training or inference and are used solely for post-hoc analysis (Liu et al., 2022b).

**MELD** extends EmotionLines into a multimodal setting and contains 13,708 utterances from 1,433 multi-party dialogues in the television series *Friends*. Each utterance was re-annotated by three annotators using its corresponding video clip and assigned by majority vote to one of seven emotions: anger, disgust, fear, joy, neutral, sadness, or surprise. MELD also provides three-class sentiment annotations, which are not used in our experiments (Poria et al., 2019).

**CHERMA** contains 28,717 Chinese utterances collected from 148 television dramas, seven variety shows, and two movies. The utterances are categorized as happiness, sadness, fear, anger, surprise, disgust, or neutrality. Each sample receives textual, acoustic, visual, and multimodal annotations independently. As with CH-SIMS v2.0, TIE uses only the multimodal annotations: modality-specific annotations are never provided during training or inference (Sun et al., 2023).

Table 6: Statistics of the standard dataset splits.

| Dataset | Task | Language | Train | Validation | Test | Label space |
|---|---|---|---|---|---|---|
| CMU-MOSI | MSA | English | 1,284 | 229 | 686 | $[-3, 3]$ |
| CMU-MOSEI | MSA | English | 16,326 | 1,871 | 4,659 | $[-3, 3]$ |
| CH-SIMS v2.0 | MSA | Chinese | 2,722 | 647 | 1,034 | $[-1, 1]$ |
| MELD | ERC | English | 9,989 | 1,109 | 2,610 | 7 classes |
| CHERMA | ERC | Chinese | 17,230 | 5,743 | 5,744 | 7 classes |

## B.2 Evaluation Metrics

Let $y_i$ and $\hat{y}_i$ denote the gold and predicted sentiment values for sample $i$, respectively, and let $N$ be the number of evaluated samples. For all regression datasets, we report mean absolute error (MAE)

and Pearson correlation (Corr):

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|, \tag{31}$$

$$\mathrm{Corr} = \frac{\sum_i(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_i(\hat{y}_i - \bar{\hat{y}})^2}\sqrt{\sum_i(y_i - \bar{y})^2}}. \tag{32}$$

Lower MAE and higher Corr indicate better regression performance.

**CMU-MOSI and CMU-MOSEI.** F1 is support-weighted under the non-negative/negative partition, with class intervals $[0, 3]$ and $[-3, 0)$; Acc-2 uses the same partition. For Acc-7, we clip predictions and annotations to $[-3, 3]$ and round them to the nearest integer sentiment level.

**CH-SIMS v2.0.** Predictions and annotations are first clipped to $[-1, 1]$. F1 is support-weighted under the non-positive/positive partition, with class intervals $[-1, 0]$ and $(0, 1]$; Acc-2 uses the same partition. Acc-2.w applies it only to samples whose gold intensity lies in $[-0.4, 0.4]$.

For any discretization function $g$, its $K$-class accuracy is

$$\text{Acc-}K = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}[g(\hat{y}_i) = g(y_i)]. \tag{33}$$

**MELD and CHERMA.** For the seven-class ERC datasets, we report classification accuracy and weighted F1. Given class support $n_c$ and the per-class precision $P_c$ and recall $R_c$, these are

$$\mathrm{Acc} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}[\hat{y}_i = y_i], \tag{34}$$

$$\mathrm{WF1} = \sum_{c=1}^{7}\frac{n_c}{N}\frac{2P_cR_c}{P_c + R_c}, \tag{35}$$

where a zero denominator contributes zero. Weighted F1 accounts for the substantial class imbalance in both ERC benchmarks.

### B.3 Baselines

**MulT.** MulT models unaligned multimodal sequences through directional cross-modal attention (Tsai et al., 2019).

**MAG-BERT.** MAG-BERT introduces a Multimodal Adaptation Gate to incorporate word-aligned acoustic and visual information into contextual BERT representations (Rahman et al., 2020).

**MISA.** MISA factorizes each modality into modality-invariant and modality-specific representations, which are subsequently integrated by a self-attention fusion module for sentiment prediction (Hazarika et al., 2020).

**CHFN.** CHFN is a Transformer-based fusion model that processes unaligned multimodal sequences and integrates non-verbal context into word representations (Guo et al., 2022).

**AV-MC.** AV-MC augments acoustic and visual signals and regularizes predictions across the resulting multimodal views (Liu et al., 2022b).

**UniMSE.** UniMSE formulates MSA and ERC as a unified text-generation task and incorporates encoded acoustic and visual features into multiple T5 layers through modality-fusion modules. We report results obtained under its original cross-task training protocol (Hu et al., 2022).

**LFMIM.** LFMIM maintains modality-specific encoders and fuses their representations through a multimodal Transformer (Sun et al., 2023).

**UniSA-BART.** UniSA jointly models multiple sentiment-analysis tasks with a multimodal generative Transformer. We report its BART-based variant under the original multi-task protocol (Li et al., 2023b).

**ITHP.** ITHP organizes multimodal information through an information-bottleneck-based hierarchical perception model with a designated prime modality and sequential detector modalities (Xiao et al., 2024).

**EUAR.** EUAR employs a mixture-of-experts architecture with uncertainty-aware routing for multimodal sentiment prediction (Gao et al., 2024).

**Coupled Mamba (CMamba).** CMamba couples modality-specific state-space chains through cross-modal hidden-state transitions while retaining independent intra-modal state evolution. The results reported in our tables use its aligned data setting (Li et al., 2024).

**DLF.** DLF combines disentangled multimodal representations, cross-modal feature refinement, and hierarchical prediction for multimodal sentiment analysis (Wang et al., 2025).

**EMOE.** EMOE uses modality-specific dynamic experts and unimodal distillation for sample-adaptive multimodal emotion prediction. We report its results under the unaligned data setting (Fang et al., 2025).

**AlignMamba.** AlignMamba combines local and global cross-modal alignment with Mamba-based multimodal fusion (Li et al., 2025b).

**MAGTKD.** MAGTKD combines prompt-enhanced textual representations, cross-modal knowledge distillation, and anchor-gated multimodal fusion for conversational emotion recognition (Li et al., 2025a).

**CMAD.** CMAD transfers correlation structure from a complete-modality teacher to students operating under different modality-availability conditions. We report its complete audio–visual–text result on CHERMA (Zhuang et al., 2025).

**DDSE.** DDSE employs a decoupled dual-stream state-space framework for multimodal representation learning and sentiment prediction (Jiang et al., 2025b).

**DiffuFuse.** DiffuFuse employs a diffusion-based dual-stream architecture for multimodal feature fusion (Lv et al., 2025).

**DRKF.** DRKF decouples shared and modality-specific representations before knowledge-based multimodal fusion (Jiang et al., 2025a).

**MDF.** MDF uses text as an anchor to disentangle acoustic representations and adaptively weights the resulting modality features for sentiment prediction (Jian et al., 2026).

**FINE.** FINE factorizes each modality into shared and unique components and uses mutual-information objectives to retain task-relevant representations (Liu & Wang, 2026).

**CCAF.** CCAF performs coarse-to-fine cross-modal alignment and fusion over global and local multimodal representations (Zhao et al., 2026b).

**TCMax.** TCMax augments multimodal learning with a total-correlation maximization objective. We report its MAG-BERT-based regression results on CMU-MOSI and CMU-MOSEI (Yu et al., 2026).

**FUSE-Net.** FUSE-Net factorizes multimodal representations and combines reconstruction and modality-difference enhancement for sentiment prediction (Yang & Li, 2026).

**CICA.** CICA pretrains modality encoders to estimate representation reliability and uses the resulting confidence signals to modulate multimodal attention (Jiang et al., 2026).

**CACR.** CACR uses conflict-aware cross-reconstruction to learn shared and modality-specific representations under cross-modal disagreement (Wang et al., 2026b).

**TSD.** TSD decomposes multimodal representations into modality-specific, pairwise shared, and globally shared subspaces before cross-modal fusion. We report its results under the unaligned data setting (Meng et al., 2026).

**EBMC.** EBMC combines semantic disentanglement, complementary enhancement, modality coordination, and reliability-aware distillation for multimodal prediction (He et al., 2026).

**CaReFlow.** CaReFlow uses cyclic adaptive rectified flows to map modality distributions before multimodal fusion (Mai & Han, 2026).

**ConsMSA.** ConsMSA models intra-modal and inter-modal semantic distribution consistency to retain sentiment-relevant multimodal representations (Wang et al., 2026a).

# C Reproducibility Details

## C.1 Feature Extraction

Except for CMU-MOSI, we use the unaligned acoustic and visual feature sequences distributed through the official MSE-Adapter data release (Yang et al., 2025a), rather than re-extracting features from the raw videos. CMU-MOSI uses the standard MMSA-Standard representation described below. TIE and the reproduced compact-prompt baselines use the same acoustic and visual inputs. Table 7 summarizes the dimensions of these representations.

**CMU-MOSI.** We use the standard unaligned representation adopted by MMSA-Standard, with COVAREP acoustic descriptors and FACET visual descriptors (Degottex et al., 2014; Zadeh et al., 2016).

**CMU-MOSEI.** Following the standard unaligned setting (Yu et al., 2021), COVAREP extracts the acoustic sequence, while OpenFace locates faces and extracts the corresponding frame-level visual descriptors (Degottex et al., 2014; Baltrusaitis et al., 2018).

**CH-SIMS v2.0.** We follow the feature pipeline released by the dataset publisher. TalkNet first localizes the active speaker, after which OpenFace extracts facial descriptors from the localized video sequence (Tao et al., 2021; Baltrusaitis et al., 2018). OpenSMILE processes audio sampled at 16 kHz to produce the acoustic sequence (Eyben et al., 2010).

**MELD.** Following UniMSE and UniSA (Hu et al., 2022; Li et al., 2023b), an EfficientNet model with supervised pretraining on VGGFace and AFEW extracts the visual sequence (Tan & Le, 2019; Parkhi et al., 2015; Dhall et al., 2012). The raw waveform is processed with librosa to obtain a sequence of Mel-spectrogram features (McFee et al., 2015).

**CHERMA.** We follow the feature pipeline released by the dataset publisher (Sun et al., 2023). MTCNN first detects and aligns faces, and a ResNet pretrained for facial-expression recognition on RAF-DB extracts the frame-level visual sequence (Zhang et al., 2016; He et al., 2016; Li et al., 2017). A pretrained wav2vec model extracts the frame-level acoustic representation (Schneider et al., 2019; Zhang et al., 2022).

Table 7: Acoustic and visual input features. Dimensions refer to one temporal feature vector before the modality encoders.

| Dataset | Acoustic feature | $d_A$ | Visual feature | $d_V$ |
|---|---|---|---|---|
| CMU-MOSI | COVAREP | 5 | FACET | 20 |
| CMU-MOSEI | COVAREP | 74 | OpenFace | 35 |
| CH-SIMS v2.0 | OpenSMILE | 25 | TalkNet + OpenFace | 177 |
| MELD | Mel spectrogram | 64 | EfficientNet | 64 |
| CHERMA | wav2vec | 1,024 | ResNet | 2,048 |

For the textual modality, each transcript is tokenized by the corresponding frozen LLM. Its selected shallow layer produces contextualized text representations, while the original token embeddings remain the textual input to the language model. During training, these representations are cached. For an unseen sample, TIE first runs the transcript through the selected shallow layer to construct the evidence tokens and then performs the full multimodal prediction forward. Accordingly, the reported TIE efficiency includes this constructor-side frozen computation; the order-aware and migration objectives themselves remain training-only.

### C.2 Prompts and Label Generation

The task prompt follows the transcript and specifies only the prediction task, valid label space, and response prefix. We use no in-context demonstrations or chain-of-thought instructions. Table 8 lists the prompt templates. For Chinese datasets, it gives the prompt used in the experiments followed by an English translation.

For MELD, the LLM input retains the contextual text supplied by the released data, while evidence organization uses the target utterance; speaker identifiers are not provided to the model. CHERMA is evaluated at the utterance level without dialogue history or speaker identifiers.

Table 8: Task prompts and label-generation formats.

| Dataset | Task prompt | Output |
|---|---|---|
| CMU-MOSI / CMU-MOSEI | Please predict the sentiment intensity of the above multimodal content in the range $[-3.0, +3.0]$. Assistant: The sentiment is | Numeric |
| CH-SIMS v2.0 | 请对上述多模态内容的情感强度进行预测，范围在$[-1.0, 1.0]$之间。响应：情感为<br>*Please predict the sentiment intensity of the above multimodal content in the range* $[-1.0, 1.0]$. *Response: The sentiment is* | Numeric |
| MELD | Please recognize the emotion of the above multimodal content from the target set: neutral:0, surprise:1, fear:2, sadness:3, joy:4, disgust:5, anger:6. Assistant: The emotion is | Class ID |
| CHERMA | 请选择适用于上述多模态内容的情绪标签：<愤怒:0，厌恶:1，恐惧:2，高兴:3，平静:4，悲伤:5，惊奇:6>。助手：情绪为<br>*Please select the emotion label applicable to the above multimodal content from: anger:0, disgust:1, fear:2, joy:3, calm:4, sadness:5, surprise:6. Assistant: The emotion is* | Class ID |

For our compact-prompt experiments, valid sentiment values span $[-3, 3]$ at intervals of $0.1$ for CMU-MOSI and CMU-MOSEI and $[-1, 1]$ at intervals of $0.2$ for CH-SIMS v2.0, following prior compact-prompt work (Yang et al., 2025a; Zhao et al., 2026a). The CH-SIMS v2.0 annotations already lie on this $0.2$-spaced grid; no additional projection is applied. During training, MSA targets are represented to one decimal place and serialized with an explicit sign. Zero has the unique representation $+0.0$, and $-0.0$ is not included in the label space. At evaluation, MAE and Corr are computed against the original annotations rather than the serialized targets. Under each evaluated tokenizer, all valid MSA values share the same four-token effective serialization, while ERC labels use one effective token. At inference, TIE and our compact-prompt reproductions use **constrained greedy decoding (CG)**, which can be viewed as token-level multiple-choice decoding: at each step, the model selects the highest-probability next token among continuations that can still form a valid sentiment value or emotion class. The constraint does not alter the model logits or their ordering among valid continuations; it guarantees that every output belongs to the predefined label space, avoids post-processing invalid outputs, and enables deterministic score extraction. We apply the same label space and decoding protocol in our controlled comparisons. **Reproduction difference (CG):** for MSE-Adapter and PaP, CG is the only deliberate change to their original decoding procedures.

We additionally compare CG with unconstrained native greedy decoding on CH-SIMS v2.0 with the ChatGLM3-6B backbone. Table 9 shows that the two protocols yield nearly identical predictive performance. Native decoding nevertheless produces occasional sequences outside the predefined label strings, including some that cannot be parsed as numeric values. Thus, CG primarily enforces a valid and deterministic output space rather than providing a material performance advantage.

### C.3 Baseline Implementations

In our controlled comparisons, the reproduced compact-prompt methods use the same released acoustic and visual features, standard data splits, frozen LLM backbones, task prompts, label representations, and decoding protocol. We retain each baseline's native evidence constructor and constructor-specific objectives; the common protocol changes only the LLM backbone where re-

Table 9: Constrained greedy (CG) versus native greedy decoding on CH-SIMS v2.0 with ChatGLM3-6B. Out counts raw outputs outside the predefined label strings; fail counts outputs with no parsable numeric value or a parsed value outside the candidate range. Counts total five runs ($5 \times 1{,}034$ predictions). Evaluation uses the last numeric value in each output; outputs without one are assigned $0.0$. All samples remain in the metrics without retry.

| Method | Decoding | MAE↓ | Acc-2↑ | Acc-2.w↑ | Corr↑ | Out / Fail↓ |
|---|---|---|---|---|---|---|
| MSE-Adapter | CG | $0.2930_{\pm.0209}$ | $82.94_{\pm2.64}$ | $75.28_{\pm2.75}$ | $0.7316_{\pm.0314}$ | $0/0$ |
| MSE-Adapter | Native | $0.2930_{\pm.0209}$ | $82.94_{\pm2.64}$ | $75.28_{\pm2.75}$ | $0.7316_{\pm.0314}$ | $0/0$ |
| PaP | CG | $0.2610_{\pm.0097}$ | $83.37_{\pm1.87}$ | $73.58_{\pm3.03}$ | $0.7808_{\pm.0122}$ | $0/0$ |
| PaP | Native | $0.2621_{\pm.0099}$ | $83.37_{\pm1.79}$ | $73.66_{\pm2.93}$ | $0.7791_{\pm.0122}$ | $66/10$ |
| TIE | CG | $0.2393_{\pm.0098}$ | $84.82_{\pm1.44}$ | $77.23_{\pm1.44}$ | $0.8024_{\pm.0186}$ | $0/0$ |
| TIE | Native | $0.2393_{\pm.0095}$ | $84.84_{\pm1.49}$ | $77.18_{\pm1.47}$ | $0.8024_{\pm.0184}$ | $10/1$ |

quired and uses the constrained decoding described in Appendix C.2. Reported baseline values in the main tables follow their cited sources.

**MAG-ChatGLM3-6B and CHFN-ChatGLM3-6B.** We use the ChatGLM3-6B adaptations and reported results provided by MSE-Adapter (Yang et al., 2025a). The original MAG formulation requires aligned modality sequences; following the MSE-Adapter implementation, an MLP aligns sequence lengths when aligned features are not provided. CHFN retains its heterogeneous cross-modal fusion architecture while replacing its pretrained text backbone with ChatGLM3-6B (Rahman et al., 2020; Guo et al., 2022).

**MSE-Adapter.** We build upon the authors' released implementation and preserve its Text-Guide-Mixer, multi-scale fusion, four-token interface, and native output projector (Yang et al., 2025a). Frozen-model migration supervision, when enabled, updates only this native projector, whereas the order-aware objective supervises the complete prediction path, including the constructor and projector, for sentiment regression. Neither extension changes the MSE constructor, the number of pseudo tokens, or its inference-time computation.

**PaP.** Because PaP does not provide a complete public implementation or all reproduction settings, we implement its reported formulation within the same experimental infrastructure (Zhao et al., 2026a). For details not specified in the paper, we use a prototype dimension of $512$, a rank-$64$ text projection, eight attention heads, CMD moments up to order five, and no dropout. When enabled, migration supervision updates PaP's output interface, whereas the order-aware objective supervises the complete prediction path for sentiment regression. These extensions leave the prototype-constructor architecture and inference-time prototype sequence unchanged.

### C.4 Implementation and Hyperparameters

All original LLM parameters are frozen, and only the evidence-organization and migration-interface parameters are optimized. ChatGLM3-6B-Base and the causal language-model components of Qwen3.5-9B and Qwen3.5-2B serve as backbones; Qwen3.5's native visual pathway is unused. Audio and visual sequences are encoded by separate single-layer bidirectional LSTMs. Table 10 summarizes the model hyperparameters shared across datasets and backbones.

During each run, early stopping monitors MAE for MSA or weighted F1 for ERC on the validation split. We repeat training with five random seeds; the main results report their mean, while the appendix ablations with five-seed results report mean±standard deviation.

We optimize trainable parameters with AdamW using $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-4}$, and weight decay $0.01$. The learning rate follows a cosine schedule with linear warm-up over the first $10\%$ of optimization steps. Training uses automatic mixed precision and seeds $1111$, $2222$, $3333$, $4444$, and $5555$. Evaluation on the validation split begins after the first epoch, and training stops after $10$ consecutive checks without improvement. Table 11 gives the objective weights used by the full model. A weight is set to zero when its corresponding component is removed in an ablation.

Table 10: Shared TIE model hyperparameters.

| Hyperparameter | Value |
|---|---|
| Evidence dimension $d_E$ | 256 |
| BiLSTM layers | 1 |
| BiLSTM hidden size per direction | 128 |
| Text representation layer | 1 |
| Text-adapter bottleneck dimension $d_R$ | 64 |
| Resampling heads | 4 |
| Dimension per head | 64 |
| Evidence tokens | 4 |
| Role-migration rank $d_M$ | 16 |
| Dropout | 0.0 |

Table 11: Loss weights for MSA and ERC.

| Task | $\lambda_{\text{ref}}$ | $\lambda_{\text{mig}}$ | $\lambda_{\text{ord}}$ |
|---|---|---|---|
| MSA | 1.0 | 0.5 | 5.0 |
| ERC | 1.0 | 0.5 | 0.0 |

For each dataset and frozen LLM backbone, we select hyperparameters on the validation split. Table 12 reports the resulting learning rates and training/evaluation batch sizes. The training horizon defines the cosine-schedule horizon; early stopping may terminate training before it is reached.

Table 12: Dataset- and backbone-specific training hyperparameters.

| Dataset | LR | Train | Eval | Horizon |
|---|---|---|---|---|
| **ChatGLM3-6B** | | | | |
| CMU-MOSI | $5 \times 10^{-4}$ | 20 | 128 | 30 |
| CMU-MOSEI | $1 \times 10^{-4}$ | 20 | 128 | 30 |
| CH-SIMS v2.0 | $2 \times 10^{-4}$ | 36 | 162 | 40 |
| MELD | $5 \times 10^{-4}$ | 10 | 101 | 30 |
| CHERMA | $5 \times 10^{-5}$ | 30 | 180 | 30 |
| **Qwen3.5-2B** | | | | |
| CMU-MOSI | $5 \times 10^{-4}$ | 24 | 686 | 30 |
| CMU-MOSEI | $2 \times 10^{-4}$ | 24 | 512 | 30 |
| CH-SIMS v2.0 | $4 \times 10^{-4}$ | 36 | 647 | 40 |
| MELD | $5 \times 10^{-4}$ | 10 | 555 | 30 |
| CHERMA | $5 \times 10^{-5}$ | 36 | 180 | 30 |
| **Qwen3.5-9B** | | | | |
| CMU-MOSI | $5 \times 10^{-4}$ | 8 | 343 | 30 |
| CMU-MOSEI | $2 \times 10^{-4}$ | 8 | 375 | 30 |
| CH-SIMS v2.0 | $2 \times 10^{-4}$ | 8 | 324 | 40 |
| MELD | $5 \times 10^{-4}$ | 4 | 278 | 30 |
| CHERMA | $5 \times 10^{-5}$ | 16 | 575 | 30 |

### C.5 Efficiency Measurement Protocol

We measure the six configurations in Table 4 on a single AutoDL vGPU-32GB instance in its Beijing B region. The device is identified by `nvidia-smi` as an NVIDIA GeForce RTX 4080 SUPER with 32,760 MiB of reported GPU memory. All measurements use ChatGLM3-6B, CH-SIMS v2.0, batch size 8, automatic mixed precision, and constrained greedy decoding. Each timing run starts from a fresh model with one complete training epoch for warm-up. We then time three complete training epochs and divide their combined, GPU-synchronized elapsed time by the number of processed samples. After training, we warm up inference on the complete test split and time a second pass

over that split. The table reports the mean per-sample time over five independently seeded runs for each configuration; these timing runs are separate from the runs used for the predictive metrics.

The training measurement reads cached shallow text representations for TIE. At inference, TIE recomputes the shallow text representations and lexical routing from each transcript before evidence construction and label generation; this online work is included in its timing. All methods use the same released acoustic and visual features, whose shared offline extraction cost is excluded. Validation and checkpoint I/O are excluded from training time.

# D Additional Experimental Results

## D.1 Temporal Routing and Migration Context

**Temporal routing.** We compare three routing variants under the same TIE training configuration with ChatGLM3-6B and Mig-Short. No Routing assigns each valid modality frame uniformly to all valid lexical positions. Fixed uses a positional Gaussian with zero offset and unit lexical width. Learned uses the modality-conditioned offsets and widths in Appendix A.1. All variants retain reference supervision and the remaining evidence-organization components.

Table 13: Temporal-routing ablation with ChatGLM3-6B. Parenthesized percentages denote relative changes from Learned routing, with negative values indicating degradation.

| Dataset | Routing | MAE↓ | F1↑ | Acc-2↑ | Corr↑ |
|---|---|---|---|---|---|
| CMU-MOSI | None | 0.5677$_{(-3.4\%)\pm.0162}$ | 89.82$_{(-0.6\%)\pm1.27}$ | 89.88$_{(-0.6\%)\pm1.22}$ | 0.8753$_{(-0.9\%)\pm.0056}$ |
| | Fixed | 0.5616$_{(-2.3\%)\pm.0162}$ | 90.24$_{(-0.1\%)\pm1.34}$ | 90.26$_{(-0.1\%)\pm1.31}$ | 0.8776$_{(-0.6\%)\pm.0051}$ |
| | Learned | 0.5490$_{\pm.0131}$ | 90.34$_{\pm0.53}$ | 90.38$_{\pm0.54}$ | 0.8831$_{\pm.0066}$ |
| CH-SIMS v2.0 | None | 0.2426$_{(-1.4\%)\pm.0079}$ | 82.88$_{(-2.4\%)\pm1.15}$ | 82.98$_{(-2.2\%)\pm1.08}$ | 0.7965$_{(-0.7\%)\pm.0154}$ |
| | Fixed | 0.2406$_{(-0.5\%)\pm.0088}$ | 83.74$_{(-1.3\%)\pm1.22}$ | 83.69$_{(-1.3\%)\pm1.22}$ | 0.8019$_{(-0.1\%)\pm.0155}$ |
| | Learned | 0.2393$_{\pm.0098}$ | 84.88$_{\pm1.37}$ | 84.82$_{\pm1.44}$ | 0.8024$_{\pm.0186}$ |

**Migration context.** We compare Mig-Short, the default five-state terminal computation, with Mig-FullCtx, which retains the complete input sequence to the final frozen block. In Mig-FullCtx, non-evidence states are detached, and the auxiliary gradient passes only through the evidence states. Both variants use TIE with ChatGLM3-6B and teacher-forced order-aware scoring on CMU-MOSI and CH-SIMS v2.0 under the same training configuration.

Table 14: Migration-context ablation with ChatGLM3-6B. Parenthesized percentages denote relative changes from Mig-Short, with negative values indicating degradation.

| Dataset | Context | MAE↓ | F1↑ | Acc-2↑ | Corr↑ |
|---|---|---|---|---|---|
| CMU-MOSI | Mig-Short | 0.5490$_{\pm.0131}$ | 90.34$_{\pm0.53}$ | 90.38$_{\pm0.54}$ | 0.8831$_{\pm.0066}$ |
| | Mig-FullCtx | 0.5497$_{(-0.1\%)\pm.0096}$ | 90.26$_{(-0.1\%)\pm1.51}$ | 90.38$_{(0.0\%)\pm1.41}$ | 0.8813$_{(-0.2\%)\pm.0035}$ |
| CH-SIMS v2.0 | Mig-Short | 0.2393$_{\pm.0098}$ | 84.88$_{\pm1.37}$ | 84.82$_{\pm1.44}$ | 0.8024$_{\pm.0186}$ |
| | Mig-FullCtx | 0.2405$_{(-0.5\%)\pm.0097}$ | 84.01$_{(-1.0\%)\pm1.47}$ | 83.98$_{(-1.0\%)\pm1.41}$ | 0.7908$_{(-1.4\%)\pm.0185}$ |

## D.2 Controlled Comparisons with Qwen3.5 Backbones

We compare the three compact-prompt constructors with the same Qwen3.5-9B backbone, released input features, data splits, label space, and decoding protocol on CH-SIMS v2.0. Org retains each constructor's native training configuration, while Full adds the applicable migration interface, frozen-model migration supervision, and order-aware objective without changing its inference-time evidence sequence.

To examine backbone-scale robustness, we additionally compare Org and Full on CMU-MOSI with Qwen3.5-2B. Both variants use the same shared model and training settings reported above; only the Full objectives and role-aware migration interface distinguish the two variants.

Table 15: Controlled compact-prompt comparison on CH-SIMS v2.0 with Qwen3.5-9B. Parenthesized percentages are relative to Org within each constructor; negative values indicate degradation.

| Constructor | Variant | MAE↓ | Acc-2↑ | Acc-2.w↑ | Corr↑ |
|---|---|---|---|---|---|
| MSE-Adapter† | Org | 0.2860±.0075 | 82.86±1.13 | 74.91±1.50 | 0.7297±.0127 |
| MSE-Adapter† | Full | 0.2690(+5.9%)±.0141 | 84.62(+2.1%)±1.60 | 76.81(+2.5%)±1.14 | 0.7627(+4.5%)±.0291 |
| PaP‡ | Org | 0.2618±.0080 | 82.42±0.91 | 75.24±1.76 | 0.7387±.0086 |
| PaP‡ | Full | 0.2473(+5.5%)±.0069 | 84.18(+2.1%)±0.36 | 76.23(+1.3%)±0.63 | 0.7718(+4.5%)±.0089 |
| TIE | Org | 0.2561±.0103 | 84.66±1.51 | 77.31±1.70 | 0.7718±.0144 |
| TIE | Full | 0.2309(+9.8%)±.0104 | 85.90(+1.5%)±1.16 | 78.26(+1.2%)±1.47 | 0.8034(+4.1%)±.0090 |

†Reproduced from the released implementation. ‡Reproduced from the paper because a complete implementation is unavailable.

Table 16: Org/Full comparison on CMU-MOSI with Qwen3.5-2B. Parenthesized percentages are relative to Org; negative values indicate degradation.

| Variant | MAE↓ | F1↑ | Acc-2↑ | Corr↑ |
|---|---|---|---|---|
| Org | 0.6699±.0265 | 84.63±0.91 | 84.69±0.87 | 0.8289±.0035 |
| Full | 0.6313(+5.8%)±.0227 | 86.23(+1.9%)±1.83 | 86.27(+1.9%)±1.71 | 0.8440(+1.8%)±.0065 |

### D.3 Additional Component Ablations

#### D.3.1 Evidence-Source Control

We compare the default evidence sources, the complete modality state $H_m$ and its reference residual $H_m - \mathrm{sg}[\widehat{H}_m]$, with an equal-capacity control in which both resampling branches read $H_m$. The two variants retain four evidence tokens, separate queries, the reference loss, and the same migration and task objectives. We use CH-SIMS v2.0 with ChatGLM3-6B and the same training configuration; both variants have 2.679M trainable parameters. Table 17 reports the results.

Table 17: Evidence-source control on CH-SIMS v2.0 with ChatGLM3-6B. Parenthesized percentages are relative to default evidence sources; negative values indicate degradation.

| Evidence sources | MAE↓ | F1↑ | Acc-2↑ | Corr↑ |
|---|---|---|---|---|
| $H_m$ / $H_m - \mathrm{sg}[\widehat{H}_m]$ | 0.2393±.0098 | 84.88±1.37 | 84.82±1.44 | 0.8024±.0186 |
| $H_m$ / $H_m$ | 0.2689(−12.4%)±.0091 | 80.82(−4.8%)±0.93 | 81.20(−4.3%)±0.81 | 0.7382(−8.0%)±.0168 |

#### D.3.2 Reference-Loss Control

We compare the full TIE configuration with a variant that sets the reference-loss weight to zero. Both use the same $H_m$ / $H_m - \mathrm{sg}[\widehat{H}_m]$ evidence sources, four evidence tokens, learned temporal routing, role-aware migration, frozen-model migration supervision, and teacher-forced order-aware objective. The comparison uses CH-SIMS v2.0 with ChatGLM3-6B, the same training hyperparameters, and 2.679M trainable parameters in each variant. Table 18 reports the mean and standard deviation over five runs.

Table 18: Reference-loss control on CH-SIMS v2.0 with ChatGLM3-6B. Parenthesized percentages are relative to Full; negative values indicate degradation.

| Reference-loss weight | MAE↓ | F1↑ | Acc-2↑ | Corr↑ |
|---|---|---|---|---|
| 1.0 (Full) | $0.2393_{\pm.0098}$ | $84.88_{\pm1.37}$ | $84.82_{\pm1.44}$ | $0.8024_{\pm.0186}$ |
| 0.0 | $0.2647_{(-10.6\%)\pm.0037}$ | $81.51_{(-4.0\%)\pm1.03}$ | $81.70_{(-3.7\%)\pm0.89}$ | $0.7443_{(-7.2\%)\pm.0149}$ |

### D.3.3 Migration-Gradient Scope Control

We examine two auxiliary gradient scopes for frozen-model migration supervision on CH-SIMS v2.0 with ChatGLM3-6B. In the default setting, this gradient updates the migration interface; in the expanded setting, it also reaches the evidence-organization module. The main task objective updates both modules in either setting. Both configurations use the same evidence constructor, loss weights, training hyperparameters, and inference procedure. Table 19 presents the results.

Table 19: Auxiliary migration-gradient scope on CH-SIMS v2.0 with ChatGLM3-6B. Parenthesized percentages are relative to migration interface; negative values indicate degradation.

| Gradient scope | MAE↓ | F1↑ | Acc-2↑ | Corr↑ |
|---|---|---|---|---|
| Migration interface | $0.2393_{\pm.0098}$ | $84.88_{\pm1.37}$ | $84.82_{\pm1.44}$ | $0.8024_{\pm.0186}$ |
| Organization + interface | $0.2653_{(-10.9\%)\pm.0081}$ | $82.40_{(-2.9\%)\pm1.28}$ | $82.50_{(-2.7\%)\pm1.17}$ | $0.7444_{(-7.2\%)\pm.0177}$ |

### D.3.4 Interface-Restricted First-Token Control

To distinguish frozen-terminal supervision from an additional polarity target, we compare it with first-token cross-entropy from the ordinary teacher-forced forward. When frozen-terminal migration supervision is disabled, this control retains first-token supervision at the same auxiliary weight. The evidence constructor, target token, and remaining objectives are held fixed, and auxiliary gradients update only the migration interface.

Table 20: Interface-restricted first-token supervision on CH-SIMS v2.0 with ChatGLM3-6B. Parenthesized percentages are relative to full forward; negative values indicate degradation.

| Source | MAE↓ | F1↑ | Acc-2↑ | Corr↑ |
|---|---|---|---|---|
| Full forward | $0.2465_{\pm.0118}$ | $83.74_{\pm1.69}$ | $83.71_{\pm1.65}$ | $0.7890_{\pm.0154}$ |
| Frozen terminal | $0.2393_{(+2.9\%)\pm.0098}$ | $84.88_{(+1.4\%)\pm1.37}$ | $84.82_{(+1.3\%)\pm1.44}$ | $0.8024_{(+1.7\%)\pm.0186}$ |

Table 21: Component ablation of the role-aware interface, order-aware objective, and frozen-model migration supervision on CMU-MOSI with Qwen3.5-9B. Rows without migration supervision retain same-weight first-token cross-entropy. Parenthesized percentages are relative to full configuration; negative values indicate degradation.

| Role | Ord. | Mig. | MAE↓ | Acc-2↑ | Corr↑ |
|---|---|---|---|---|---|
| Yes | Yes | Yes | $0.5258_{\pm.0145}$ | $90.47_{\pm0.61}$ | $0.8954_{\pm.0043}$ |
| No | Yes | Yes | $0.5360_{(-1.9\%)\pm.0134}$ | $90.26_{(-0.2\%)\pm0.54}$ | $0.8931_{(-0.3\%)\pm.0058}$ |
| Yes | No | Yes | $0.5336_{(-1.5\%)\pm.0091}$ | $90.00_{(-0.5\%)\pm0.88}$ | $0.8884_{(-0.8\%)\pm.0030}$ |
| Yes | Yes | No | $0.5303_{(-0.9\%)\pm.0113}$ | $90.09_{(-0.4\%)\pm0.63}$ | $0.8916_{(-0.4\%)\pm.0047}$ |
| No | Yes | No | $0.5338_{(-1.5\%)\pm.0136}$ | $89.48_{(-1.1\%)\pm0.69}$ | $0.8916_{(-0.4\%)\pm.0038}$ |
| Yes | No | No | $0.5388_{(-2.5\%)\pm.0133}$ | $89.74_{(-0.8\%)\pm0.56}$ | $0.8881_{(-0.8\%)\pm.0077}$ |

## E Evidence-Migration Diagnostics

**Scope of the evidence.** Text-prior reinforcement concerns the development of a text-directed preference along the evidence-migration path. Prediction statistics characterize text-directed outputs,

layerwise readouts trace this preference, and source-path interventions test its connection to the final prediction. Evidence replacement tests whether the constructed evidence supports the correct label. Together, these analyses support the coexistence of text-directed contributions and useful evidence; they do not establish that the former suppresses the latter.

### E.1 Text-Directed Behavior

CH-SIMS v2.0 provides a multimodal target $y_i$ and a separately annotated text target $y_{T,i}$. For $y_{T,i} \neq y_i$, we define

$$d_i = |y_{T,i} - y_i|, \qquad B_{T,i} = \frac{\widehat{y}_i - y_i}{y_{T,i} - y_i}. \tag{36}$$

Here, $B_{T,i} = 0$ corresponds to the multimodal annotation and $B_{T,i} = 1$ to the text annotation. Samples are grouped by text–multimodal discrepancy magnitude, with $B_T$ omitted where its denominator is zero. The separately annotated modality targets are used only for this post-hoc analysis and are never supplied during training or inference. A positive trend in $B_T$ characterizes text-directed prediction behavior but does not by itself locate where that behavior arises. The six discrepancy bins contain 315, 361, 172, 116, 36, and 34 test samples, respectively, from smallest to largest discrepancy.

### E.2 Layerwise Text-Directed Readout

We additionally examine the 76 non-neutral CH-SIMS v2.0 test samples for which the text and multimodal annotations have opposite polarities. At the label-decision position of each frozen-LLM layer, we apply the frozen terminal normalization and language-model head to obtain a positive-versus-negative sign-token logit margin and orient it toward the text-label polarity. Positive values indicate a text-directed preference under this diagnostic readout. Because the terminal head is applied to intermediate states, the values trace representational changes rather than calibrated predictions at each layer.

Figure 7 reports the raw signed margins without subtracting the first-layer value. TIE-Org and TIE-Full begin with nearly identical margins ($-0.599$ and $-0.597$) and both develop a positive text-directed margin through the middle frozen blocks, consistent with reinforcement of a text-directed preference under this diagnostic readout. Their final-layer margins are $0.158$ and $-0.250$, respectively. The curves characterize an aggregate tendency rather than monotonic behavior for every sample.

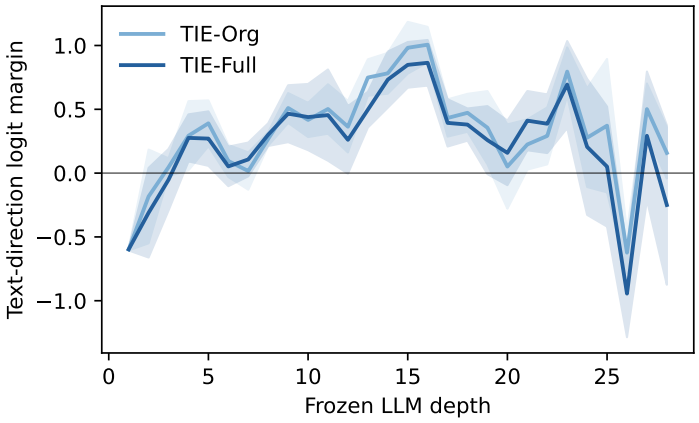



Figure 7: Raw, uncentered text-directed sign-token logit margin across frozen-LLM layers on CH-SIMS v2.0 polarity-conflict samples. Positive values favor the text-label polarity; negative values favor the multimodal-label polarity. Lines and shading show the mean and standard deviation across seeds.

### E.3 Conditional Evidence Propagation and Gold-Label Support

For each checkpoint, we compute a separate reference for each of the four evidence-token positions by averaging its migrated representation over the training split. We replace a selected group of these tokens with their corresponding references while leaving the transcript and task prompt unchanged. We consider all evidence, acoustic evidence, visual evidence, text-referenced roles, and text-residual roles. For group $g$, layer $\ell$, and sequence position $j$, the resulting hidden-state displacement is

$$D_g^{(\ell,j)} = \mathrm{RMS}\left(h_{\mathrm{clean}}^{(\ell,j)} - h_{\mathrm{replace}(g)}^{(\ell,j)}\right). \tag{37}$$

Layer-by-position maps retain evidence, transcript, task-prompt, decision, and label positions. For the displayed retention summaries, we average $D_g^{(\ell,j)}$ over downstream positions or layers, average the final quarter of the resulting profile, and divide by that profile's maximum. The deep–far statistic averages the final layer quarter over the final downstream-position quarter and divides by the 99th-percentile displacement on the downstream grid. Each quantity is normalized within the same model, run, sample, and intervention; it measures persistence relative to that response's own peak rather than absolute displacement across models. The case heatmaps preserve the true layer–position structure, while the curves summarize responses across seeds.

Figure 4 also reports teacher-forced gold-label probabilities. Each annotation gives the clean probability followed in parentheses by the clean-minus-replacement probability difference. A positive difference indicates that evidence replacement reduces support for the gold label.

The displayed case is an illustrative conditional trace rather than a population estimate. The replacement maps show where an evidence change remains observable, while the probability differences quantify its effect on gold-label support.

Figures 8 and 9 provide two further CH-SIMS v2.0 cases under the same replacement and plotting protocol. They illustrate variation in the layer–position response and in how the evidence change affects gold-label support; neither case is a population estimate.

Figure 10 shows a corresponding case for PaP on the same dataset. It illustrates that conditional propagation and changes in gold-label support can also be examined with a different evidence constructor. Figure 11 extends the conditional trace to CHERMA emotion classification. The gold-class probability is shown once because each class label occupies one effective token.

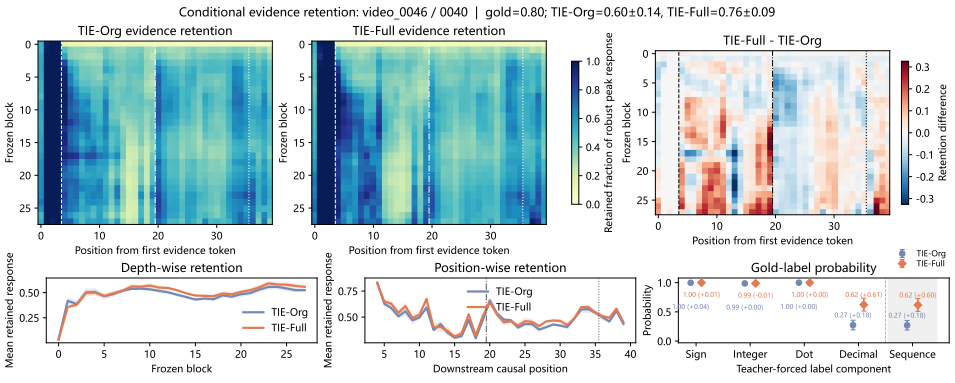



Figure 8: Additional conditional propagation case on CH-SIMS v2.0 (video 0046/0040). Conventions follow Figure 4; parenthesized probabilities show the clean-minus-replacement difference.

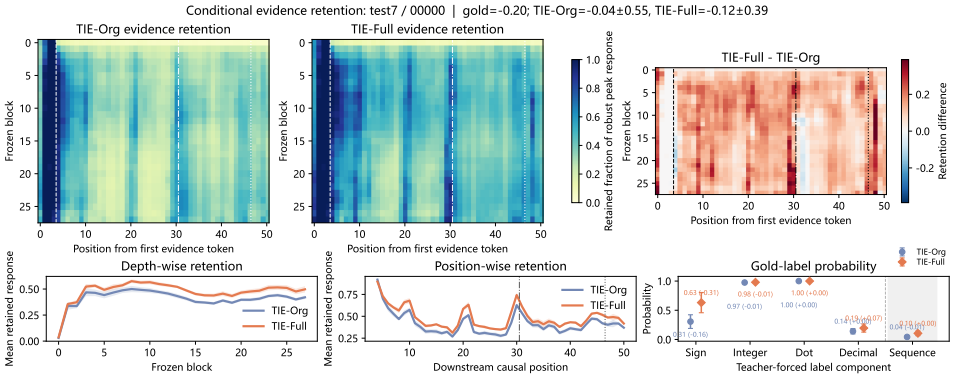



Figure 9: Additional conditional propagation case on CH-SIMS v2.0 (test7/00000), using the same conventions as Figure 4.

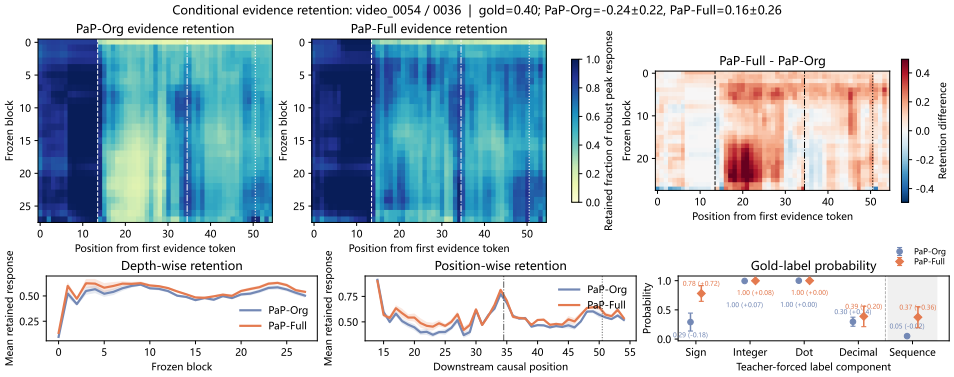



Figure 10: Conditional propagation case for PaP on CH-SIMS v2.0 (video 0054/0036). Org and Full denote the constructor before and after the combined training configuration; other conventions follow Figure 4.

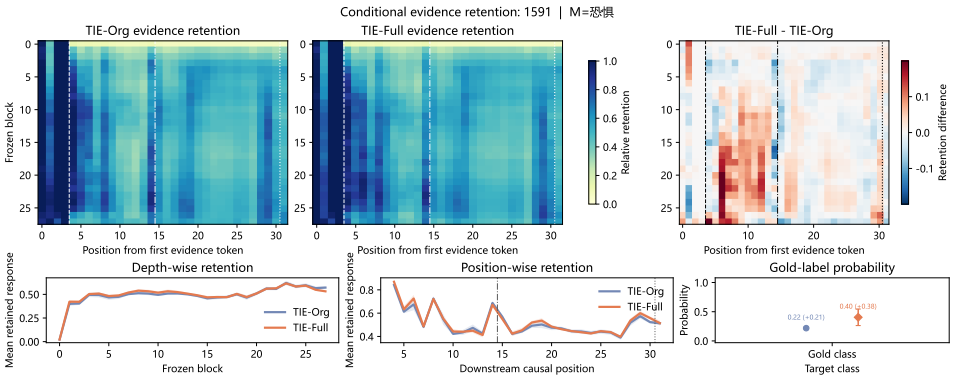



Figure 11: Conditional propagation case for TIE on CHERMA (sample 1591). The lower-right panel shows the teacher-forced gold-class probability, with the clean-minus-replacement change in parentheses.

### E.4 Source-Path Intervention and Predictive Influence

We examine all 76 non-neutral CH-SIMS v2.0 test samples whose text and multimodal annotations have opposite polarities, together with 76 non-neutral same-polarity controls selected by a fixed sample-ID rule. Let $D$ be the last prompt position, which predicts the first sign token, and let $S$ contain either the transcript or the four evidence-token positions. At selected frozen layers, we attenuate the post-softmax contribution from $S$ to the attention output at $D$:

$$c_{\ell,S} = W_O^\ell \operatorname{Concat}_h \left( \sum_{j \in S} a_{h,D,j}^\ell v_{h,j}^\ell \right), \qquad o'_{\ell,D} = o_{\ell,D} - \epsilon c_{\ell,S}. \tag{38}$$

The output projection bias and other attention contributions remain unchanged. For the 28 ChatGLM3 transformer blocks, the fixed early, middle, and late groups contain zero-based layers 0–8, 9–17, and 18–27, respectively. We use attenuation factors $\epsilon \in \{0.1, 0.25\}$. At the same decision position, controls apply a random-direction perturbation or attenuate the whole attention output, matching the source intervention's actual perturbation norm. A zero-intervention forward checks that the sign logits are unchanged.

For each sample, we orient the final positive-versus-negative sign-logit margin toward the multimodal annotation and report its change from the clean forward. Table 22 reports the middle-layer results for the conflict subset; positive changes favor the multimodal polarity. The table compares interventions within each trained model and does not treat Org and Full checkpoints as paired by seed.

Table 22: Change in the correct-direction sign-token margin on 76 CH-SIMS v2.0 polarity-conflict samples. Middle-layer group; $\epsilon = 0.25$. Controls match the norm of the intervention on the same source and checkpoint.

| Variant | Source | Attenuation | Whole control | Random control |
|---|---|---|---|---|
| Org | Transcript | +0.1171$_{\pm 0.0472}$ | +0.0188$_{\pm 0.0304}$ | +0.0285$_{\pm 0.0153}$ |
| Full | Transcript | +0.2009$_{\pm 0.0744}$ | +0.0519$_{\pm 0.0544}$ | +0.0426$_{\pm 0.0188}$ |
| Org | Evidence | -0.0241$_{\pm 0.0955}$ | +0.0668$_{\pm 0.0499}$ | -0.0093$_{\pm 0.0231}$ |
| Full | Evidence | -0.1573$_{\pm 0.1331}$ | +0.0591$_{\pm 0.0853}$ | -0.0196$_{\pm 0.0392}$ |

Table 23: Source-position attenuation across layer groups and polarity subsets at $\epsilon = 0.25$. Positive changes favor the multimodal-label polarity. Entries are mean±standard deviation across analyzed model instances.

| Variant | Subset | Source | Early | Middle | Late |
|---|---|---|---|---|---|
| Org | Conflict | Transcript | -0.0051$_{\pm 0.0034}$ | +0.1171$_{\pm 0.0472}$ | +0.0040$_{\pm 0.0040}$ |
| Full | Conflict | Transcript | -0.0125$_{\pm 0.0032}$ | +0.2009$_{\pm 0.0744}$ | +0.0010$_{\pm 0.0079}$ |
| Org | Conflict | Evidence | +0.0403$_{\pm 0.0226}$ | -0.0241$_{\pm 0.0955}$ | -0.1325$_{\pm 0.0944}$ |
| Full | Conflict | Evidence | +0.0349$_{\pm 0.0511}$ | -0.1573$_{\pm 0.1331}$ | -0.1902$_{\pm 0.0759}$ |
| Org | Same polarity | Transcript | +0.0132$_{\pm 0.0111}$ | -0.2902$_{\pm 0.0128}$ | +0.0060$_{\pm 0.0071}$ |
| Full | Same polarity | Transcript | +0.0158$_{\pm 0.0165}$ | -0.1915$_{\pm 0.0861}$ | +0.0079$_{\pm 0.0068}$ |
| Org | Same polarity | Evidence | -0.0227$_{\pm 0.0841}$ | -0.2729$_{\pm 0.1579}$ | +0.0588$_{\pm 0.1212}$ |
| Full | Same polarity | Evidence | +0.0673$_{\pm 0.1196}$ | -0.2643$_{\pm 0.1366}$ | -0.0577$_{\pm 0.0881}$ |

In Full, attenuating the transcript contribution favors the multimodal polarity on conflict samples in the middle layers, whereas attenuating the evidence contribution has the opposite effect. On same-polarity samples, attenuating either source in the middle layers reduces the correct-direction margin. Early-layer effects are smaller, while late-layer evidence attenuation is less distinct from the whole-output control. The middle-layer evidence effect is also present at $\epsilon = 0.1$. These margin changes rarely flip the predicted sign. Transcript positions can already carry multimodal information, and the evidence tokens are text-conditioned, so the interventions identify source-position contributions rather than pure linguistic and non-linguistic information.

The replacement analysis provides an output-level check on the same conflict subset. With the transcript fixed, replacing all constructed evidence by its training-split reference lowers the teacher-forced gold-sign probability by $0.1811$ for Org and $0.2370$ for Full, averaged over the analyzed checkpoints. The corresponding gains on the full $1{,}034$-sample test set are $0.0222$ and $0.0250$. Thus, the larger conflict-set response should not be generalized to every sample; it shows that the constructed evidence influences correct-label support even where the text and multimodal annotations disagree.

## F AI USE STATEMENT

We used generative AI tools to assist with manuscript editing and to identify potentially relevant literature. The authors reviewed AI-assisted revisions for factual accuracy and consistency with the intended meaning, and checked suggested references against their original sources. The authors take full responsibility for the final manuscript, including its claims, citations, and conclusions.