TEB: A BANDWIDTH-CONSTRAINED EVIDENCE ALLOCATION FRAMEWORK FOR ENHANCING MULTIMODAL SENTIMENT ANALYSIS AND EMOTION RECOGNITION IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Adapter-based LLM frameworks for multimodal sentiment analysis and emotion recognition in AI conversations are attractive for their lightweight flexibility, but the adapter interface provides only B tokens. Long audio-visual streams must be compressed into a few evidence tokens while preserving conversation-relevant cues. The key challenge is deciding which evidence to retain under this bandwidth constraint. We propose the *Text-Conditioned Evidence Bottleneck* (TEB), a bandwidth-constrained evidence allocation framework. TEB uses transcript embeddings to build B utterance-relevant query states, organizes audio-visual sequences into local segment and latent global memories, and routes evidence per query through soft gating based on audio-text and visual-text consistency discrepancies. Low-rank coefficients and a shared prompt dictionary map each query to one LLM-receivable evidence token. On three LLMs and four public datasets, TEB improves the five-seed mean of all metrics over MSE-Adapter. At $B = 4$, it reduces ChatGLM3-6B's MAE on SIMS-V2 from 0.279 to 0.265 and increases weighted F1 on CHERMA from 72.80 to 74.18. LLaMA2-7B shows significant improvements in Acc-2, F1, weak sentiment accuracy, MAE, and correlation coefficient on SIMS-V2, while Qwen-1.8B also demonstrates consistent gains on MOSEI and CHERMA. These results establish TEB as a lightweight, flexible, and scalable foundation to advance LLMs' multimodal sentiment analysis and emotion recognition capabilities. Our code is available at https://anonymous.4open.science/r/TEBTEB-111.

1 INTRODUCTION

Multimodal sentiment analysis and emotion recognition in conversation are core AI tasks. Unlike isolated sentence-level classification, they identify a speaker's emotional state by integrating text, speech, and vision within conversational context, requiring inter-utterance interaction modeling and speaker-aware context tracking. They have strong potential in social media monitoring, intelligent healthcare, and emotion-aware dialog agents (Wu et al., 2025). The field has shifted from task-specific small models to LLM-driven approaches. Before LLMs, it relied on LSTM- and Transformer-based fusion architectures with cross-modal attention or graph neural networks (Shou et al., 2025). These achieved competitive benchmark results, but generalized poorly due to limited scale and data. LLMs, with strong semantic reasoning, now offer a general foundation for understanding sentiment, driving the unification of multimodal emotion recognition into generative frameworks (Zhang et al., 2026).

Adapters align non-text modalities with LLMs using few trainable parameters, making adapter-based methods popular for their lightweight flexibility. Representative works include Emotion-LLaMA (Cheng et al., 2024), which projects audio-visual features into a unified space via emotion-specific encoders and concatenates them with text tokens for LLaMA; MSE-Adapter (Yang et al., 2025), whose Text-Guide-Mixer adds only about 2.6M parameters to enable 6/7B models for emotion recognition while preserving LLM generality; and Emotion-LLaMAv2 (Peng et al., 2026), whose adapter maps modality sequences into an LLM-compatible shared embedding space. These

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.

methods share a common strategy: freeze the LLM and use a lightweight adapter to convert heterogeneous modality features into LLM-consumable evidence tokens.

However, adapter methods face a key limitation: the adapter interface has extremely limited token bandwidth. When continuous, long-temporal conversational audio-visual information is injected into the LLM, the number of tokens it can carry is much smaller than the original audio-visual token scale. Emotion-LLaMA uses 256 video tokens but only 1 audio token; video information is sparse and noisy due to high dimensionality, forcing inference to favor compact audio cues (Cheng et al., 2024), further highlighting this bottleneck. Thus, under adapter bandwidth constraints, deciding which audio-visual evidence to retain and expose to the LLM is critical for emotion recognition performance. Existing work on this bottleneck includes token compression and selection (Cheng et al., 2024; He et al., 2025; Jaime Bellver-Soler, 2025; Alayrac et al., 2022; Li et al., 2023), evidence token identification and injection (Liu et al., 2026a), and alignment of text-conditioned modality (Yang et al., 2025; Korbar et al., 2024; Liang et al., 2024). However, key gaps remain: (1) evidence selection lacks dynamic text-conditioned guidance, (2) bandwidth-constrained allocation remains static, and (3) evidence competition within limited bandwidth is not systematically studied. In summary, a clear gap exists: how to dynamically and optimally allocate audio-visual evidence tokens within limited bandwidth, conditioned on conversation transcript semantics, to maximally preserve sentiment-relevant cues during compression.

To address these challenges, we propose TEB, a bandwidth-constrained evidence allocation framework to enhance LLM multimodal sentiment analysis and recognition. TEB uses conversation transcript semantics as conditioning signals to dynamically allocate audio-visual evidence tokens under the adapter interface bandwidth constraint B, allowing limited tokens to maximize the preservation of sentiment-relevant audio-visual cues. Unlike existing methods that passively compress or prune tokens, TEB actively and conditionally selects which evidence to retain and expose to the LLM, optimizing evidence allocation under bandwidth constraints. Experiments on three LLMs and four public datasets show that TEB outperforms baselines and the advanced MSE-Adapter. The main contributions are as follows:

- Model Design: We propose TEB, a bandwidth-constrained evidence allocation framework that, under adapter interface bandwidth B, builds transcript-conditioned queries, dynamically allocates audio-visual evidence tokens from local segment and latent global memories, and preserves sentiment-relevant cues to guide LLM inference, offering a reference adapter design for LLM multimodal sentiment analysis and emotion recognition.
- Theoretical Innovation: We formalize the audio-visual memory organization (local segment and learnable latent memories), their independent retrieval and adaptive fusion, and the modeling and learning optimization of evidence allocation.
- Comprehensive Validation: Extensive experiments show TEB effectively enhances LLM multimodal sentiment analysis and emotion recognition across four datasets and three LLMs, even surpassing the advanced MSE-Adapter.

2 Related Work

Multimodal affect modeling. Tensor and low-rank fusion model cross-modal interactions (Zadeh et al., 2017; Liu et al., 2018); modality-specific learning, cross-modal attention and information-maximizing objectives provide other routes (Hazarika et al., 2020; Yu et al., 2021; Tsai et al., 2019; Rahman et al., 2020; Han et al., 2021). CH-SIMS v2.0 emphasizes non-verbal cues (Liu et al., 2022); local–global pathways and text-enhanced fusion also appear in affect recognition (Zhou et al., 2025; Li et al., 2025). Conversational ERC models use speaker and dialogue dependencies (Majumder et al., 2019; Ghosal et al., 2019; Hu et al., 2021); TEB processes each utterance without modeling dialogue history.

Query interfaces and LLM adapters. Flamingo, BLIP-2 and Video-LLaMA use learned query bridges between perceptual encoders and language models (Alayrac et al., 2022; Li et al., 2023; Zhang et al., 2023); TCR and QaP incorporate text conditioning (Korbar et al., 2024; Liang et al., 2024). Our single-memory control tests a related interface choice within one pipeline, not a reproduction of either full architecture. MSE-Adapter, ATGFB-MFF and expert-guided fusion pro-

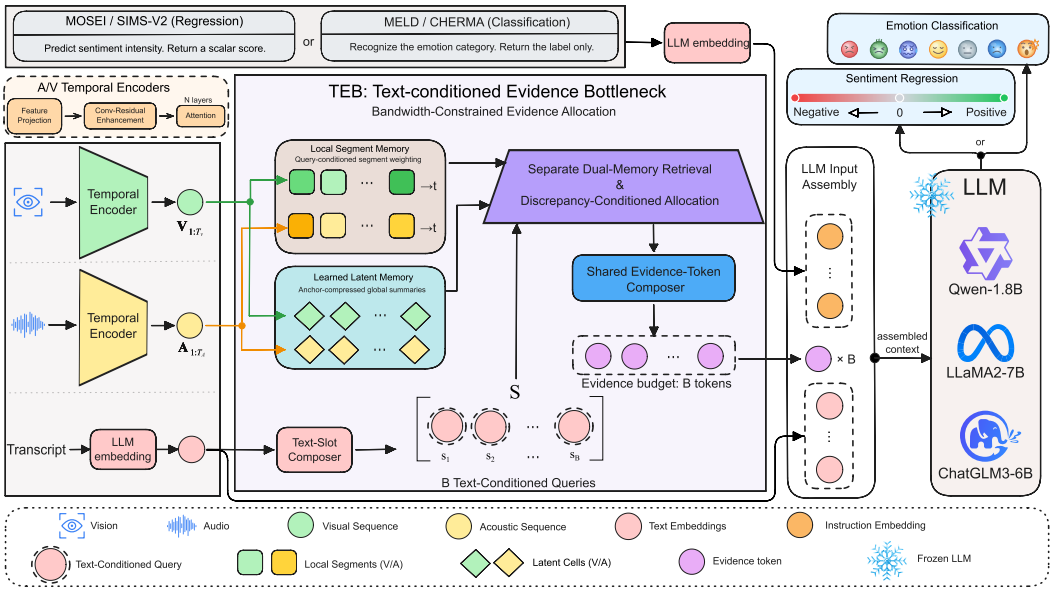

Figure 1: **TEB overview.** The conversation transcript representation serves as both the native input to the LLM and the source of B text-conditioned retrieval queries. Audio and visual information are temporally encoded into sequence representations, which jointly form a segmented local memory and a learnable latent memory. Using the B text conditions, TEB retrieves and fuses these memories to produce exactly B evidence tokens (adapter bandwidth B) for the frozen LLM. The LLM task instruction is emotion regression or classification under autoregressive supervision.

vide complementary multimodal adapters (Yang et al., 2025; Liu et al., 2026b; Qiao et al., 2026). We compare in-house interfaces under a common protocol and use published scores only as non-matched benchmark context.

3 Text-Conditioned Evidence Bottleneck

We define the bandwidth of the interface, denoted by B, as the number of continuous prompt tokens that a multimodal adapter can provide to a frozen LLM. As shown in Figure 1, the proposed TEB framework takes as input the audio-visual information and transcript of an AI conversation. It organizes audio and visual representations into a segmented local memory and a learnable latent memory, retrieves them via soft attention using B text-conditioned queries, and fuses the results into B evidence tokens that prompt the LLM. Finally, *evidence allocation* modeling is introduced to characterize how audio and visual representations map to appropriate evidence token positions.

Bandwidth-Constrained Evidence Token Generation In TEB, the "bottleneck" specifically refers to the interface bandwidth B of the LLM's multimodal adapter. Under this bandwidth constraint, evidence token generation is defined as follows:

$$\mathbf{P} = g_\theta(\mathbf{x}_a, \mathbf{x}_v; \mathbf{E}_{\text{tr}}) \in \mathbb{R}^{B \times d_{\text{LLM}}}, \qquad \hat{y} = \text{Decode}_{\mathcal{M}}(\Pi(\mathbf{E}_{\text{inst}}, \mathbf{P}, \mathbf{E}_{\text{tr}})), \tag{1}$$

Here, g_θ denotes the trainable multimodal adapter, and $\mathbf{P}$ denotes its output: B evidence tokens supplied to the frozen LLM as continuous prompt tokens. $m \in \{a, v\}$ denotes the audio or visual modality of the conversation, and its representation $\mathbf{x}_m \in \mathbb{R}^{T_m \times d_m}$ is a variable-length sequence. The LLM embedding of the transcript $\mathbf{u}_t$ is $\mathbf{E}_{\text{tr}} = \text{Emb}_{\mathcal{M}}(\mathbf{u}_t)$. $\mathbf{E}_{\text{inst}}$ is the embedding of the task instruction, d_{LLM} is the embedding dimension, and the query dimension d, memory-cell counts and dictionary size R are all independent of B. Π denotes the task prompt template, which assembles the task instruction, $\mathbf{P}$ and $\mathbf{E}_{\text{tr}}$ as input to the LLM, as illustrated in Figure 1. The generated response is mapped to either a sentiment intensity score or an emotion category label. The task instruction defines the prediction objective, while evidence retrieval is conditioned on the transcript. During

training, the frozen LLM computes the autoregressive loss over the target response, and only the parameters of g_θ are optimized.

Transcript-Conditioned Queries A projection of frozen transcript embeddings supplies keys and values for B learnable query anchors. Attention over valid transcript positions, combined with a pooled transcript residual, yields input-dependent states $\mathbf{S} \in \mathbb{R}^{B\times d}$. Anchors are shared across examples, whereas query states are not. Appendix A details their construction.

Two Organizations of Audio-Visual Memory Audio-visual information is encoded by a temporal encoder into contextualized audio-visual features $\mathbf{Z}_m \in \mathbb{R}^{T_m\times d}$ via convolutional refinement and light self-attention, Appendix A provides details. TEB builds two memories from the same features, allowing each query to access both segment-defined and learnable latent aggregations.

Local segment memory. Local memory partitions each encoded audio-visual contextual feature sequence into consecutive S_m segments in temporal order, retaining each segment as a candidate evidence unit. Masked mean pooling and max pooling are computed through a learnable projection block f_L to obtain:

$$\mathbf{z}^L_{m,s} = \mathrm{LN}(f_L([\mathrm{Mean}(\mathbf{Z}_{m,s}); \mathrm{Max}(\mathbf{Z}_{m,s})])) \,. \tag{2}$$

Source embeddings, query-conditioned scalar weighting, and normalization yield the local memory $\mathbf{M}^L \in \mathbb{R}^{(S_a+S_v)\times d}$, as specified in Appendix A. All valid segments are retained, and the process employs soft weighting rather than hard top-k selection.

Learned latent memory. Latent memory offers an aggregation scheme unconstrained by predefined segment boundaries. For each modality's contextual feature sequence, it aggregates using C_m learnable anchors $\mathbf{G}_m \in \mathbb{R}^{C_m\times r_G}$ over the full valid sequence:

$$\mathbf{M}^G_m = \mathrm{LN}\left[\mathrm{softmax}\left(\frac{\mathbf{G}_m(\mathbf{Z}_m\mathbf{W}^G_K)^\top}{\sqrt{r_G}}\right)\mathbf{Z}_m\mathbf{W}^G_V\right]. \tag{3}$$

After incorporating modality source embeddings into the latent units, we concatenate the audio and visual units to obtain $\mathbf{M}^G \in \mathbb{R}^{(C_a+C_v)\times d}$. The two memories differ in how they aggregate encoded positions: local memory units pool over consecutive segments, whereas latent memory units apply learned attention over all valid positions. The attention weights of some latent memory units may still concentrate on local temporal regions. Segment boundaries define pooling regions, not the receptive field of the contextualized encoder. The sizes of both memories are independent of the number of multimodal prompt tokens B.

Independent Retrieval and Adaptive Fusion of Dual Memories Each query retrieves from both memories. Pooled audio-visual states and the difference between their cosine similarities to the mean transcript state form an utterance-level context that modulates the queries and local-memory weighting. This is an embedding discrepancy, not a measured emotional conflict; its formulation is in Appendix A.

Separate reads produce one query-specific summary from each memory before combining the two. A shared cross-attention core performs both reads, followed by path-specific normalization:

$$\mathbf{E}^L = \mathrm{LN}_L(\mathrm{CrossAttn}(\mathbf{S}^c, \mathbf{M}^L)), \qquad \mathbf{E}^G = \mathrm{LN}_G(\mathrm{CrossAttn}(\mathbf{S}^c, \mathbf{M}^G)). \tag{4}$$

Concatenating the same candidate cells for one attention read instead makes their total attention mass compete under a single normalization, where relative key scores and memory-cell counts influence the balance. In TEB, each path first yields its own query-conditioned summary, and mixing takes place after these separate reads. The useful balance between summaries may vary across inputs and queries, so a learned scalar coefficient $\gamma_b \in (0, 1)$ mixes the two paths for query b:

$$\begin{aligned} &\gamma_b = \sigma\big(f_g([\mathbf{s}^c_b; \mathbf{e}^L_b; \mathbf{e}^G_b; \mathbf{W}_m\mathbf{m}])\big) \,, \quad \mathbf{e}_b = \gamma_b\mathbf{e}^L_b + (1-\gamma_b)\mathbf{e}^G_b, \\ &\tilde{\mathbf{s}}_b = \mathrm{LN}(\mathbf{s}^c_b + \rho\mathbf{e}_b), \qquad \hat{\mathbf{s}}_b = \mathrm{LN}(\tilde{\mathbf{s}}_b + \mathrm{FFN}(\tilde{\mathbf{s}}_b)). \end{aligned} \tag{5}$$

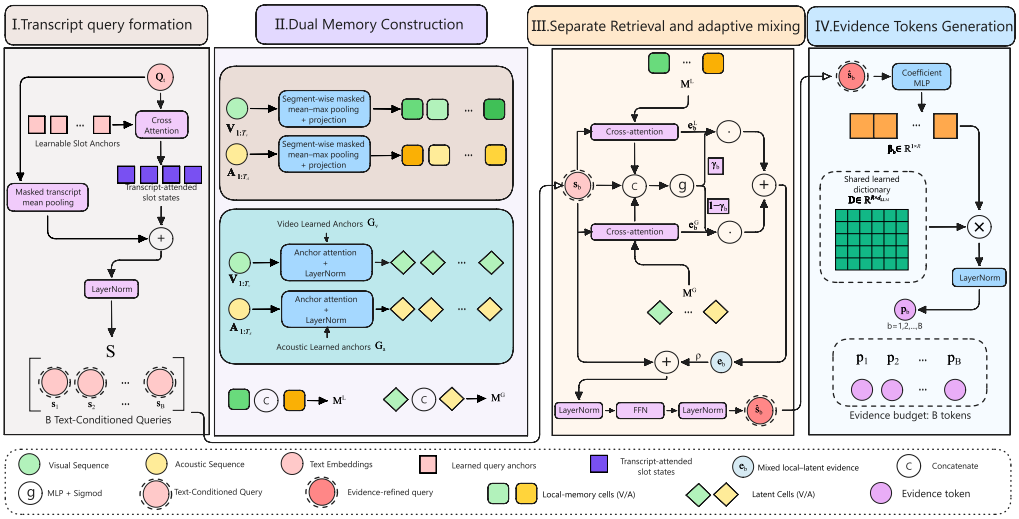

Figure 2: **Memory retrieval and prompt generation.** All B transcript-conditioned queries access the same local and latent memories through soft attention. Local cells pool contiguous segments; latent cells use learned aggregation over valid sequence positions. The retrieved states are fused before each query produces one prompt token. B counts adapter-token positions, not memory cells or semantic categories.

The residual scale ρ is learned. The gate permits a query-dependent local–latent mixture without requiring different queries to read distinct evidence. Residual query states retain transcript information, so the prompt tokens combine text-conditioned states with retrieved audio-visual representations.

Evidence Allocation Modeling and Learning Each refined query predicts coefficients over a shared prompt dictionary and thereby produces one LLM-facing token, so query count and exposed token count are both B. Only the adapter is trained through the frozen LLM's autoregressive task loss. Unlike task-level continuous prefixes and soft prompts (Li & Liang, 2021; Lester et al., 2021), these tokens depend on each example's transcript and audio-visual evidence. This parameterization does not itself establish a resource advantage. Appendix A specifies the composer and objective.

4 Experiments

We evaluate the evidence-allocation design at three levels. First, the benchmark matrix tests whether the complete adapter improves prediction across three backbones and four tasks under our unified protocol, with published references reported separately. Second, five-seed comparisons at the same B test the complete interface against fusion and resampling alternatives, while a retrained text-only control checks whether its retained transcript path reproduces the result. Third, controls using the same local and latent candidates probe the value of memory access and mixing; a TEB budget sweep describes sensitivity to the coupled query and prompt-token count. These comparisons fix the LLM-facing token count where stated, not total adapter capacity or computation.

4.1 Tasks, Comparisons and Reporting

Tasks and backbones. MOSEI (Bagher Zadeh et al., 2018) and CH-SIMS v2.0 (SIMS-V2) (Liu et al., 2022) evaluate multimodal sentiment analysis through intensity prediction and derived classification metrics. MELD (Poria et al., 2019) is a conversational emotion-recognition (ERC) benchmark; CHERMA (Sun et al., 2023) provides multimodal emotion labels for utterance clips. TEB models audio-visual sequences within an utterance; its memories do not represent dialogue history. The benchmark matrix includes Qwen-1.8B (Bai et al., 2023), LLaMA2-7B (Touvron et al., 2023) and ChatGLM3-6B. The equal-budget interface controls use ChatGLM3-6B on SIMS-V2 and CHERMA; same-candidate access is also compared on MELD.

Table 1: Multimodal sentiment analysis results. Published task-specific rows follow MSE-Adapter (Yang et al., 2025); published frozen-LLM rows follow ATGFB-MFF (Liu et al., 2026b) and are marked †. They are not protocol-matched to our five-seed runs. Mean ± sample standard deviation is shown where available; bold compares each in-house MSE/TEB pair. All frozen-LLM adapter rows use $B = 4$ on these tasks.

Method	B	CMU-MOSEI					SIMS-V2				
		Acc-2 ↑	F1 ↑	Acc-7 ↑	MAE ↓	Corr ↑	Acc-2 ↑	F1 ↑	Acc-2 weak ↑	MAE ↓	Corr ↑
Published task-specific multimodal baselines											
TFN	–	78.50	78.96	51.60	0.573	0.714	76.51	76.31	66.27	0.323	0.667
LMF	–	80.54	80.94	51.59	0.576	0.717	77.05	77.02	69.34	0.343	0.638
MulT	–	81.15	81.56	52.84	0.559	0.733	79.50	79.59	69.61	0.317	0.703
MISA	–	83.60	83.80	52.20	0.555	0.756	80.53	80.63	70.50	0.314	0.725
MMIM	–	82.24	82.66	54.24	0.526	0.772	80.95	80.97	72.28	0.316	0.707
CHFN	–	83.70	83.90	54.30	0.525	0.778	–	–	–	–	–
AV-MC	–	–	–	–	–	–	82.50	82.55	74.54	0.297	0.732
UniMSE	–	85.86	85.79	54.39	0.523	0.773	–	–	–	–	–
Paper-reported frozen-LLM adapters (not protocol-matched)											
MSE-Qwen-1.8B†	4	84.88 ± 0.26	84.15 ± 0.28	53.05 ± 0.36	0.563 ± 0.013	0.743 ± 0.010	81.64 ± 0.32	79.41 ± 0.34	73.27 ± 0.39	0.314 ± 0.010	0.660 ± 0.012
ATGFB-Qwen-1.8B†	4	87.75 ± 0.24	87.52 ± 0.26	54.23 ± 0.35	0.531 ± 0.012	0.779 ± 0.011	80.82 ± 0.33	80.67 ± 0.35	74.50 ± 0.38	0.362 ± 0.011	0.692 ± 0.012
MSE-LLaMA2-7B†	4	86.47 ± 0.25	85.71 ± 0.27	55.79 ± 0.34	0.514 ± 0.011	0.766 ± 0.009	75.84 ± 0.36	75.38 ± 0.38	67.91 ± 0.42	0.387 ± 0.013	0.568 ± 0.014
ATGFB-LLaMA2-7B†	4	88.63 ± 0.23	87.84 ± 0.25	57.76 ± 0.33	0.486 ± 0.010	0.798 ± 0.008	77.97 ± 0.35	77.74 ± 0.36	71.59 ± 0.40	0.379 ± 0.012	0.556 ± 0.014
MSE-ChatGLM3-6B†	4	85.78 ± 0.25	85.32 ± 0.27	54.19 ± 0.35	0.541 ± 0.012	0.771 ± 0.009	82.66 ± 0.31	82.34 ± 0.33	75.06 ± 0.38	0.313 ± 0.010	0.708 ± 0.011
ATGFB-ChatGLM3-6B†	4	89.83 ± 0.22	89.49 ± 0.24	58.04 ± 0.32	0.529 ± 0.011	0.803 ± 0.008	86.85 ± 0.30	85.75 ± 0.32	79.43 ± 0.36	0.284 ± 0.009	0.736 ± 0.010
Runs under our unified protocol											
MSE-Qwen-1.8B	4	84.16 ± 0.53	83.68 ± 0.49	52.19 ± 0.46	0.558 ± 0.012	0.732 ± 0.016	80.21 ± 1.06	80.13 ± 1.14	73.42 ± 1.83	0.316 ± 0.012	0.662 ± 0.016
TEB-Qwen-1.8B	4	**85.50** ± 0.40	**84.98** ± 0.54	**54.08** ± 0.98	**0.526** ± 0.011	**0.762** ± 0.007	**80.68** ± 0.41	**80.57** ± 0.35	**74.37** ± 1.40	**0.304** ± 0.007	**0.671** ± 0.008
MSE-LLaMA2-7B	4	86.49 ± 0.30	86.25 ± 0.26	55.08 ± 0.40	0.507 ± 0.007	0.787 ± 0.004	77.00 ± 1.93	76.99 ± 1.91	70.27 ± 2.63	0.370 ± 0.015	0.576 ± 0.026
TEB-LLaMA2-7B	4	**87.09** ± 0.14	**86.79** ± 0.19	**55.89** ± 0.53	**0.495** ± 0.007	**0.799** ± 0.004	**79.50** ± 0.91	**79.40** ± 0.95	**72.50** ± 0.45	**0.311** ± 0.011	**0.655** ± 0.021
MSE-ChatGLM3-6B	4	86.70 ± 0.31	86.29 ± 0.32	54.13 ± 0.71	0.521 ± 0.010	0.785 ± 0.005	83.37 ± 0.21	83.51 ± 0.23	75.21 ± 0.45	0.301 ± 0.005	0.704 ± 0.009
TEB-ChatGLM3-6B	4	**87.36** ± 0.18	**87.45** ± 0.36	**55.27** ± 0.64	**0.505** ± 0.006	**0.794** ± 0.007	**84.60** ± 0.96	**84.50** ± 1.06	**77.60** ± 1.74	**0.265** ± 0.007	**0.761** ± 0.010

Matched protocol. At $B = 4$, the four interfaces share feature inputs, temporal encoders, frozen LLM, prompt composer, training objective, optimizer, random seeds and checkpoint-selection rule. Training uses AdamW, mixed precision, gradient clipping and validation-based early stopping. Regression responses use numeric targets rounded to one decimal during training; classification responses encode the dataset label. The primary controlled metrics are SIMS-V2 mean absolute error (MAE) and CHERMA weighted F1 (WF1), with correlation and accuracy as complementary measures. Results are mean ± sample standard deviation over seeds 1111, 2222, 3333, 4444 and 5555. Standard deviations describe seed variation, not confidence intervals. All in-house experiments are trained and evaluated on a single NVIDIA GeForce RTX 5090 GPU; no multi-GPU or distributed training is used.

Table 2: Multimodal emotion recognition results. Published task-specific rows follow MSE-Adapter (Yang et al., 2025); published frozen-LLM rows follow ATGFB-MFF (Liu et al., 2026b) and are not protocol-matched. In-house results use five seeds; bold compares each MSE/TEB pair. $B_{\mathrm{M/C}}$ denotes the number of continuous prompt tokens provided by the adapter to the LLM on MELD and CHERMA, respectively.

Method	$B_{\mathrm{M/C}}$	MELD		CHERMA	
		Acc ↑	WF1 ↑	Acc ↑	WF1 ↑
Published multimodal baselines					
TFN	–	60.77	57.74	–	68.37
LMF	–	61.15	58.30	–	68.23
MulT	–	–	–	–	69.24
MM-DFN	–	62.49	59.46	–	–
LFMIM	–	–	–	–	70.54
UniMSE	–	65.09	65.51	–	–
Paper-reported frozen-LLM adapters (not protocol-matched)					
MSE-Qwen-1.8B†	2/4	62.73 ± 0.38	60.26 ± 0.40	70.41 ± 0.36	70.52 ± 0.35
ATGFB-Qwen-1.8B†	4/4	65.48 ± 0.36	62.56 ± 0.38	73.99 ± 0.34	73.70 ± 0.33
MSE-LLaMA2-7B†	4/4	65.28 ± 0.36	63.75 ± 0.37	71.59 ± 0.35	71.33 ± 0.34
ATGFB-LLaMA2-7B†	4/4	66.68 ± 0.35	63.55 ± 0.37	74.36 ± 0.33	73.58 ± 0.32
MSE-ChatGLM3-6B†	4/4	66.45 ± 0.35	65.33 ± 0.36	73.26 ± 0.33	72.88 ± 0.32
ATGFB-ChatGLM3-6B†	4/4	69.91 ± 0.33	68.93 ± 0.34	75.79 ± 0.31	75.42 ± 0.31
Runs under our unified protocol					
MSE-Qwen-1.8B	2/4	61.85 ± 0.56	59.69 ± 0.45	70.60 ± 0.24	70.36 ± 0.27
TEB-Qwen-1.8B	2/4	**63.32** ± 0.37	**61.05** ± 0.36	**72.46** ± 0.65	**72.23** ± 0.54
MSE-LLaMA2-7B	4/4	64.93 ± 0.67	63.56 ± 0.80	71.69 ± 0.91	71.55 ± 0.94
TEB-LLaMA2-7B	4/4	**65.92** ± 0.72	**64.16** ± 0.67	**72.69** ± 0.82	**72.70** ± 0.87
MSE-ChatGLM3-6B	4/4	64.49 ± 0.50	63.54 ± 0.55	72.68 ± 0.38	72.78 ± 0.37
TEB-ChatGLM3-6B	4/4	**66.23** ± 0.54	**64.99** ± 0.33	**74.13** ± 0.33	**74.18** ± 0.27

Published references. The frozen-LLM reference rows for both MSE-Adapter and ATGFB-MFF are transcribed from ATGFB-MFF (Liu et al., 2026b) and grouped by backbone. They retain that source's three-seed reporting. Task-specific reference rows follow MSE-Adapter (Yang et al., 2025). These values are not reruns under our protocol. In particular, ATGFB-MFF reports interval-based regression decoding, whereas our runs generate numeric responses. We neither pool their variances with ours nor assign paired significance markers across the two sources.

Benchmark Context Tables 1 and 2 place our matched MSE-Adapter and TEB runs alongside separately marked, non-matched published references. TEB has a better five-seed mean than our MSE runs on every reported metric across three backbones and four benchmarks. This evaluates the complete interface, not the isolated effect of memory access. Published ATGFB-MFF scores are higher in many backbone–dataset settings, including ChatGLM3-6B on MELD (68.93 versus TEB's 64.99 WF1). Its published runs are not protocol-matched; on regression tasks, ATGFB-MFF also uses interval-based decoding whereas ours generates numeric responses. These values therefore provide benchmark context rather than a head-to-head architectural ranking.

4.2 EQUAL-BUDGET INTERFACE COMPARISON

At $B = 4$, we compare four ways to construct the LLM-facing prompt tokens. *Global text-guided fusion* pools each modality and expands one fused state into B outputs. An *unconditional resampler* retrieves from a modality-tagged audio-visual sequence using learned queries shared across samples. A *text-conditioned single-memory resampler* uses transcript-derived queries over that same sequence. TEB instead constructs separate local and latent memories and mixes two retrieved states per query. All four return exactly B tokens through the same prompt composer. Appendix B specifies the interfaces.

Table 3: Controlled comparison of interfaces at $B = 4$. Sequence denotes one modality-tagged audio–visual memory; L/G mix denotes query-wise local–latent mixing. The single-memory resampler uses transcript-conditioned queries without TEB's separate local and latent memories. SIMS-V2 and CHERMA entries use ChatGLM3-6B and report mean ± sample standard deviation over five matched seeds. Bold denotes the best mean within the controlled interfaces. Lower MAE and higher Corr, Acc and WF1 indicate better performance.

(a) Interface configuration

Interface	Query	A/V memory	L/G mix	Discrep.
Global text-guided fusion	Transcript	Pooled	–	–
Unconditional query resampler	Learned	Sequence	–	–
Text-conditioned resampler	Transcript	Sequence	–	–
TEB	Transcript	Local + latent	✓	✓

(b) Equal-budget results

Interface	SIMS-V2		CHERMA	
	MAE ↓	Corr ↑	Acc ↑	WF1 ↑
Global text-guided fusion	0.277 ± 0.006	0.735 ± 0.013	71.56 ± 0.79	71.32 ± 0.78
Unconditional query resampler	0.281 ± 0.009	0.726 ± 0.006	71.85 ± 0.94	71.74 ± 1.17
Text-conditioned resampler	0.279 ± 0.009	0.724 ± 0.006	72.75 ± 0.70	72.80 ± 0.69
TEB	**0.265** ± 0.007	**0.761** ± 0.010	**74.13** ± 0.33	**74.18** ± 0.27

Complete-interface performance. TEB gives the best mean score on all four available metrics in Table 3. Against the single-memory resampler, SIMS-V2 MAE decreases from 0.279 to 0.265 and correlation increases from 0.724 to 0.761. CHERMA WF1 increases from 72.80 to 74.18, a gain of 1.38 percentage points. The corresponding WF1 gains over global fusion and unconditional resampling are 2.86 and 2.44 points. On CHERMA, TEB improves both accuracy and WF1 over the single-memory interface in all five matched seeds; seed-wise WF1 gains range from 0.48 to 2.30 points. These seed-wise differences are descriptive comparisons, not a significance test.

Comparison beyond text conditioning. The single-memory interface already conditions on the paired transcript and exposes the same number of prompt tokens. Its comparison with TEB therefore tests more than the presence of text conditioning or a longer LLM input. Conditioning alone is not uniformly beneficial: relative to unconditional queries, the single-memory interface improves CHERMA WF1 but gives mixed SIMS-V2 changes. The results favor the complete TEB interface in the tested settings. Because this comparison changes memory construction and retrieval together, we examine more specific alternatives below.

Matched Text-Only Control To test whether TEB's performance can be reproduced by its transcript-processing path alone, we retrain a matched text-only variant. It retains the native transcript, the B transcript-conditioned slots, the evidence-token composer, the frozen LLM and the optimization protocol, but bypasses the audio-visual inputs and both evidence memories. This con-

Table 4: Same-candidate memory controls with ChatGLM3-6B at $B = 4$. Values are five-seed mean $\pm$ sample standard deviation; bold marks the best mean.

Access configuration	SIMS-V2		CHERMA	
	MAE ↓	Corr ↑	Acc ↑	WF1 ↑
Local only	0.277 $\pm$ 0.008	0.747 $\pm$ 0.015	73.31 $\pm$ 0.88	73.25 $\pm$ 0.89
Latent only	0.275 $\pm$ 0.011	0.741 $\pm$ 0.015	73.55 $\pm$ 0.39	73.42 $\pm$ 0.43
Concatenated single read	0.269 $\pm$ 0.009	0.750 $\pm$ 0.015	73.63 $\pm$ 0.60	73.53 $\pm$ 0.71
Fixed mixture	0.270 $\pm$ 0.012	0.750 $\pm$ 0.023	73.31 $\pm$ 0.81	73.23 $\pm$ 0.73
No discrepancy	0.267 $\pm$ 0.010	0.753 $\pm$ 0.012	73.59 $\pm$ 0.40	73.48 $\pm$ 0.53
TEB	**0.265** $\pm$ 0.007	**0.761** $\pm$ 0.010	**74.13** $\pm$ 0.33	**74.18** $\pm$ 0.27

trol therefore preserves the principal text-side route and the same B LLM-facing adapter positions; it is not parameter- or compute-matched to the complete multimodal model.

The retrained text-only control retains the transcript-conditioned query and token-generation route while removing audio-visual inputs (Appendix Table 8). Relative to it, full TEB reduces SIMS-V2 MAE from 0.334 to 0.265 and raises CHERMA WF1 from 52.92 to 74.18. The audio-visual route materially contributes in these settings, although this control does not isolate audio from vision or identify causal temporal segments.

Sensitivity to the Adapter-Token Budget For TEB, B sets both query count and LLM-facing token count. We vary $B \in \{1, 2, 4, 8\}$ while retaining the memory sizes, hidden width, dictionary rank and training protocol. The ChatGLM3-6B sweep on CHERMA yields mean WF1 values of 73.15, 73.57, 74.18 and 73.49, respectively; Appendix Table 7 gives both metrics and seed standard deviations.

The highest observed mean is at $B = 4$, exceeding $B = 2$ and $B = 8$ by 0.60 and 0.69 WF1 points, respectively; the full range is 1.03 points. This non-monotonic pattern characterizes TEB's sensitivity to the coupled query and prompt-token counts. It does not establish a universal optimum or a comparative low-budget advantage: single-memory results at the other budgets are not reported.

5 Conclusion

TEB constructs local and latent audio-visual memories, lets transcript-conditioned queries read them separately, and maps each query to one of B tokens for a frozen LLM. Its complete interface improves the reported five-seed means over our MSE-Adapter reruns across three backbones and four affect benchmarks. At matched B, it exceeds the tested interfaces on SIMS-V2 and CHERMA and a concatenated single-read control on MELD. The access evidence is task dependent and does not isolate separate reads from active trainable capacity. Candidate organization and access are consequential design choices at a fixed prompt-token count.

Ethics Statement

Affect labels are context-dependent annotations, not direct measurements of a person's internal state. Errors may vary with language, culture, expression style and data collection conditions. This study does not validate the use of its predictions for consequential judgments about individuals. Applications should respect the source datasets' access conditions, participant privacy and applicable consent requirements.

Reproducibility Statement

Section 3 specifies the shared architecture. Section 4 distinguishes matched runs from published reference values and identifies the reported seeds and uncertainty convention. Appendix B details the controlled interfaces and budget results; Appendix C reports completed access controls and their interpretation. Full run manifests and feature-preprocessing settings remain to be supplied with the completed experimental release.

AI Use Statement

AI tools assisted with research-story development, manuscript drafting and language refinement, LaTeX editing, and preliminary figure design. The authors are responsible for verifying all scientific claims, experimental results, citations and visual content before submission.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 23716–23736, 2022.

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 *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, 2018. doi: 10.18653/v1/P18-1208. URL https://aclanthology.org/P18-1208/.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning, 2024. URL https://arxiv.org/abs/2406.11161.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 154–164, 2019. doi: 10.18653/v1/D19-1015. URL https://aclanthology.org/D19-1015/.

Wei Han, Haiyun Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1122–1131, 2020. doi: 10.1145/3394171.3413678. URL https://arxiv.org/abs/2005.03545.

Peng He, Jun Yu, Chengjie Ge, Ye Yu, Wei Xu, Lei Wang, Tianyu Liu, and Zhen Kan. Domain-separated bottleneck attention fusion framework for multimodal emotion recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 21(4), 2025. URL https://doi.org/10.1145/3711865.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5666–5675, 2021. doi: 10.18653/v1/2021.acl-long.440. URL https://aclanthology.org/2021.acl-long.440/.

Ricardo Córdoba Luis Fernando D'Haro Jaime Bellver-Soler, Mario Rodríguez-Cantelar. Cutting through overload: Efficient token dropping for speech emotion recognition in multimodal large language models. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pp. 284–289, 2025.

Bruno Korbar, Yongqin Xian, Alessio Tonioni, Andrew Zisserman, and Federico Tombari. Text-conditioned resampler for long form video understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 15144, pp. 271–288, 2024. doi: 10.1007/978-3-031-73016-0_16.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, 2021. doi: 10.18653/v1/2021.emnlp-main.243. URL https://aclanthology.org/2021.emnlp-main.243/.

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 *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 19730–19742, 2023.

Xiang Li, Xianfu Cheng, Dezhuang Miao, Xiaoming Zhang, and Zhoujun Li. TF-Mamba: Text-enhanced fusion mamba with missing modalities for robust multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 11252–11267, 2025. doi: 10.18653/v1/2025.findings-emnlp.602.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, 2021. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353/.

Tian Liang, Jing Huang, Ming Kong, Luyuan Chen, and Qiang Zhu. Querying as prompt: Parameter-efficient learning for multimodal language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26855–26865, 2024. doi: 10.1109/CVPR52733.2024.02536.

Cheng Liu, Yiping Wang, Quanlong Guan, Chaobo He, Xingyu Zhu, and Liangda Fang. Mitigating evidence suppression: Bi-level active evidence injection for educational video understanding. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1221–1231, 2026a. URL https://doi.org/10.1145/3805712.3809586.

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 International Conference on Multimodal Interaction (ICMI)*, 2022. URL https://arxiv.org/abs/2209.02604.

Zhaowei Liu, Sheng Liu, Weiqing Yan, Peng Song, Yongchao Song, and Rufei Gao. ATGFB-MFF: Adaptive text-guided fiber bundle feature fusion with LLMs for multimodal sentiment analysis and emotion recognition in conversations. In *Proceedings of the ACM Web Conference (WWW)*, pp. 7442–7453, 2026b. doi: 10.1145/3774904.3792577.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2247–2256, 2018. doi: 10.18653/v1/P18-1209. URL https://aclanthology.org/P18-1209/.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. DialogueRNN: An attentive RNN for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 6818–6825, 2019. doi: 10.1609/aaai.v33i01.33016818. URL https://ojs.aaai.org/index.php/AAAI/article/view/4657.

Xiaojiang Peng, Jingyi Chen, Zebang Cheng, Bao Peng, Fengyi Wu, Yifei Dong, Shuyuan Tu, Qiyu Hu, Huiting Huang, Yuxiang Lin, Jun-Yan He, Kai Wang, Zheng Lian, and Zhi-Qi Cheng. Emotion-llamav2 and mmeverse: A new framework and benchmark for multimodal emotion understanding, 2026. URL https://arxiv.org/abs/2601.16449.

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 *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, 2019. doi: 10.18653/v1/P19-1050. URL https://aclanthology.org/P19-1050/.

Jiaqi Qiao, Xinran Li, Yifan Lyu, Xiujuan Xu, et al. Expert-guided multimodal fusion for unified emotion and sentiment analysis. *arXiv preprint arXiv:2601.07565*, 2026. URL https://arxiv.org/abs/2601.07565v2.

Wasif Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 2359–2369, 2020.

Yuntao Shou, Tao Meng, Wei Ai, and Keqin Li. Multimodal large language models meet multimodal emotion recognition and reasoning: A survey, 2025. URL https://arxiv.org/abs/2509.24322.

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 *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 658–670, 2023. doi: 10.18653/v1/2023.acl-long.39. URL https://aclanthology.org/2023.acl-long.39/.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. URL https://arxiv.org/abs/2307.09288.

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 *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 6558–6569, 2019.

ChengYan Wu, Yiqiang Cai, Yang Liu, Pengxu Zhu, Yun Xue, Ziwei Gong, Julia Hirschberg, and Bolei Ma. Multimodal emotion recognition in conversations: A survey of methods, trends, challenges and prospects. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 6257–6274, 2025. doi: 10.18653/v1/2025.findings-emnlp.332. URL https://aclanthology.org/2025.findings-emnlp.332/.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 25642–25650, 2025. doi: 10.1609/AAAI.V39I24.34755.

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. doi: 10.1609/aaai.v35i12.17289. URL https://ojs.aaai.org/index.php/AAAI/article/view/17289.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1103–1114, 2017. doi: 10.18653/v1/D17-1115. URL https://aclanthology.org/D17-1115/.

Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 543–553, 2023. doi: 10.18653/v1/2023.emnlp-demo.49. URL https://aclanthology.org/2023.emnlp-demo.49/.

Hongrui Zhang, Daiqing Wu, Yangyang Li, Kuien Liu, Yuhui Wang, Yu Zhou, and Sicheng Zhao. Multimodal emotion recognition with large language models, 2026. URL https://arxiv.org/abs/2605.21239.

Miao Zhou, Lina Yang, Thomas Wu, Dongnan Yang, and Xinru Zhang. Dual-path dynamic fusion with learnable query for multimodal sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 11355–11365, 2025. doi: 10.18653/v1/2025.emnlp-main.571.

A Implementation Details

A.1 Temporal Feature Enhancement

Audio and visual feature sequences are processed separately. A lightweight convolutional residual block refines each stream, followed by projection into a shared d-dimensional space, sinusoidal positions, and self-attention/feed-forward blocks. The resulting $\mathbf{Z}_a$ and $\mathbf{Z}_v$ retain a temporal index for memory construction. Pooled encoder states supply $\mathbf{h}_a$ and $\mathbf{h}_v$ in Eq. (9). Padding masks are propagated through sequence encoding, pooling and memory attention.

The same temporal front end supplies contextualized observations to every matched interface, so it is held fixed in the equal-budget comparison.

A.2 Local-Memory Construction

Let $\mathbf{z}^L_{m,s}$ be the pooled segment state in Eq. (2), including the nonlinear projection and normalization. The mean context-conditioned query $\bar{\mathbf{s}}^c = B^{-1}\sum_b \mathbf{s}^c_b$ enters a segment-weighting network:

$$w_{m,s} = \sigma\left(\frac{f_{\mathrm{sal}}(\mathbf{z}^L_{m,s}) + \langle \mathbf{W}_z\mathbf{z}^L_{m,s}, \mathbf{W}_s\bar{\mathbf{s}}^c\rangle/\sqrt{r_L}}{\tau_L}\right). \tag{6}$$

With modality embedding $\mathbf{e}_m$ and local-memory embedding $\mathbf{e}_L$, the memory cells are

$$\mathbf{M}^L = \mathrm{LN}_{\mathrm{mem}}\big(\mathrm{Concat}_{m,s}\left[w_{m,s}(\mathbf{z}^L_{m,s} + \mathbf{e}_m + \mathbf{e}_L)\right]\big). \tag{7}$$

Only valid segments participate in downstream attention. All valid cells remain in memory; weighting does not change cardinality. The scalar weights are intermediate activations, not calibrated relevance scores. In particular, the normalization after weighting can attenuate the effect of scalar rescaling. Attention or weight visualizations alone therefore cannot establish that a segment caused the prediction.

A.3 Prompt Parameterization

The dictionary in Eq. (10) is learned jointly with the coefficient network. Its size R bounds the rank of the pre-normalization factorization, independently of B. Orthogonal initialization specifies the starting dictionary, not its geometry throughout training. Optional token-wise soft scaling changes amplitudes without selecting or removing tokens. Its setting must be held fixed within a controlled comparison and identified in the run specification.

A.4 Transcript Query Construction

The paired transcript conditions retrieval before the audio-visual representations enter the fixed-size prompt interface. This allows the adapter's reads to depend on verbal context even though the LLM also receives the transcript directly. A projection maps transcript token embeddings to $\mathbf{Q}_t \in \mathbb{R}^{T_t\times d}$. With B learned anchors $\mathbf{A} \in \mathbb{R}^{B\times r_t}$, we form

$$\mathbf{S} = \mathrm{LN}\left[\mathrm{softmax}\left(\frac{\mathbf{A}(\mathbf{Q}_t\mathbf{W}_K)^\top}{\sqrt{r_t}}\right)\mathbf{Q}_t\mathbf{W}_V + \phi(\bar{\mathbf{Q}}_t)\right] \in \mathbb{R}^{B\times d}. \tag{8}$$

Padding positions are masked, and $\bar{\mathbf{Q}}_t$ is the masked transcript mean. The anchors are shared across examples; the pooled transcript values make the resulting query states input dependent.

A.5 Utterance Context

Each query reads both memories. We provide an utterance-level context as an additional conditioning signal for these reads and their subsequent mixing. Pooled audio-visual states and their relative cosine similarity to the transcript first modulate the queries:

$$c = \tfrac{1}{2}\left|\cos(\mathbf{h}_a, \bar{\mathbf{Q}}_t) - \cos(\mathbf{h}_v, \bar{\mathbf{Q}}_t)\right|, \qquad \mathbf{m} = \psi([\mathbf{h}_a; \mathbf{h}_v; c]), \qquad \mathbf{S}^c = \mathbf{S} + \mathbf{W}_c\mathbf{m}, \tag{9}$$

Here, $\mathbf{h}_a, \mathbf{h}_v$ are pooled encoder states, and c is half the absolute difference between their cosine similarities to the mean transcript representation. The vector $\mathbf{W}_c\mathbf{m}$ is broadcast across query positions. The resulting states $\mathbf{S}^c$ also condition local-memory weighting (Appendix A). This context describes the current utterance, not dialogue history; c measures embedding discrepancy rather than labeled emotional conflict.

A.6 Prompt Composition and Objective

A coefficient network maps each refined state to R coefficients over a shared dictionary $\mathbf{D} \in \mathbb{R}^{R\times d_{\mathrm{LLM}}}$:

$$\boldsymbol{\beta}_b = f_\theta(\hat{\mathbf{s}}_b), \qquad \mathbf{p}_b = \mathrm{LN}(\boldsymbol{\beta}_b\mathbf{D}), \qquad b = 1, \ldots, B. \tag{10}$$

Stacking the outputs gives $\mathbf{P}$ in Eq. (1). Mapping each refined query directly to one token avoids an additional pooling step across queries and ties the query count to the exposed token count B. The shared dictionary is a factorized prompt parameterization; its net resource benefit depends on the chosen widths and is not established by the factorization alone. Appendix A describes its initialization.

For a target response $\mathbf{y} = (y_1, \ldots, y_L)$, the shared autoregressive task objective is

$$\mathcal{L}_{\mathrm{LLM}} = -\sum_{\ell=1}^{L} \log p_{\mathcal{M}}(y_\ell \mid y_{<\ell}, \Pi(\mathbf{E}_{\mathrm{inst}}, \mathbf{P}, \mathbf{E}_{\mathrm{tr}})). \tag{11}$$

The backbone weights remain frozen, while gradients through the LLM train the adapter. All in-house SIMS-V2 comparisons in the main paper use this base objective and feed the uncorrected evidence tokens $\mathbf{P}$ to the LLM. Appendix D evaluates structured calibration (SC) as a separate, optional SIMS-V2 extension.

A.7 Trainable-Parameter Counts

Table 5 reports TEB's trainable (non-frozen) adapter parameters, excluding the frozen LLM backbone and the offline audio-visual feature extractors; the count is identical across seeds. The optional SIMS-V2 structured-calibration extension is excluded here.

Table 5: Trainable adapter parameters (millions) of TEB. Each backbone–dataset configuration is counted once; B follows the settings of Tables 1 and 2.

Backbone	MOSEI	SIMS-V2	MELD	CHERMA
Qwen-1.8B	2.05	1.84	1.93	2.31
LLaMA2-7B	2.69	1.99	2.34	2.51
ChatGLM3-6B	2.84	2.37	2.60	2.94

B Controlled Evaluation Details

B.1 Interface Definitions

Table 6 specifies the four interfaces. They share the encoder, frozen LLM, output dictionary rank, target format, optimizer, seeds and checkpoint rule. They differ in the construction of the pre-LLM states. Matching the adapter-token count does not match trainable parameters, since memory construction and mixing introduce their own parameters. The comparisons therefore evaluate complete interfaces at a fixed output count.

B.2 Token-Budget Sweep

For every tested B, query count and prompt-token count are both B. Internal segment counts, latent-cell counts, hidden width, dictionary rank and decoding settings are held fixed. The $B = 4$ entries in Table 7 are shared with the equal-budget study, not independent repetitions. Every filled entry is computed from the five individual seed rows; no values are imputed for the unreported single-memory runs.

Table 6: Definitions of the equal-budget interfaces. Every variant returns exactly B multimodal prompt tokens.

Interface	Query source	Audio-visual memory	Operation before prompt composition
Global text-guided fusion	Pooled transcript state	Pooled audio and visual states	A text-conditioned modality gate forms one global state, which is expanded with B learned output anchors.
Unconditional query resampler	B learned queries shared across samples	One modality-tagged sequence memory	Queries cross-attend to the concatenated audio and visual sequences. No transcript-derived state enters retrieval.
Text-conditioned resampler	B transcript-conditioned queries	The same modality-tagged sequence memory	Transcript-conditioned queries cross-attend to the same memory used by the unconditional variant.
TEB	B transcript-conditioned queries	Local segment and compressed latent memories	Each query reads both memories, mixes the retrieved states and produces one prompt token.

Table 7: Available prompt-budget results on CHERMA with ChatGLM3-6B. Each entry is the mean and sample standard deviation over seeds 1111, 2222, 3333, 4444 and 5555. Query count and exposed token count both equal B. The $B = 4$ entries are shared with Table 3; bold denotes the highest observed mean for each metric.

Interface	Metric	$B = 1$	$B = 2$	$B = 4$	$B = 8$
TEB	Acc ↑	73.47 ± 0.64	73.67 ± 0.57	**74.13** ± 0.33	73.38 ± 0.87
TEB	WF1 ↑	73.15 ± 0.58	73.57 ± 0.78	**74.18** ± 0.27	73.49 ± 0.80

Table 8: Matched text-only control with ChatGLM3-6B at $B = 4$. Every entry is the mean and sample standard deviation over the same five seeds. The text-only variant retains the transcript-conditioned query and prompt-generation path while removing audio-visual evidence. Bold denotes the better mean.

Evidence input	SIMS-V2					CHERMA	
	Acc-2 ↑	F1 ↑	Acc-2 weak ↑	MAE ↓	Corr ↑	Acc ↑	WF1 ↑
Text only	78.34 ± 0.90	78.02 ± 0.97	70.39 ± 1.31	0.334 ± 0.009	0.660 ± 0.010	53.67 ± 1.19	52.92 ± 1.65
Multimodal TEB	**84.60** ± 0.96	**84.50** ± 1.06	**77.60** ± 1.74	**0.265** ± 0.007	**0.761** ± 0.010	**74.13** ± 0.33	**74.18** ± 0.27

C Component Ablations and Interpretation

This appendix reports the completed SIMS-V2, CHERMA and MELD access controls and the scope of their interpretation. Every numeric entry is a five-seed measurement.

C.1 Memory Access and Adaptive Mixing

Local-only and latent-only variants restrict retrieval to one memory. The concatenated single-read control retains both sets of constructed cells but places them in one attention normalization. A fixed mixture instead reads the memories separately and combines the retrieved states with constant local-memory weight γ_0. The no-discrepancy variant removes only c from Eq. (9); $\mathbf{h}_a$ and $\mathbf{h}_v$, both memories, separate reads and learned mixing remain. The reported controls use ChatGLM3-6B,

$B = 4$ and the same five seeds as their full-TEB references. The fixed-weight result represents the evaluated setting, not an exhaustive search over constant weights.

Table 9: Memory-access, mixing and discrepancy controls with ChatGLM3-6B at $B = 4$. Entries are mean ± sample standard deviation over five seeds. The concatenated variant exposes the same local and latent memory candidates through one attention read; γ_0 denotes the tested fixed local-memory weight. Bold denotes the best available mean in each metric.

Variant	Local	Latent	Access/mixing	SIMS-V2		CHERMA	
				MAE ↓	Corr ↑	Acc ↑	WF1 ↑
Local only	✓	–	–	0.277 ± 0.008	0.747 ± 0.015	73.31 ± 0.88	73.25 ± 0.89
Latent only	–	✓	–	0.275 ± 0.011	0.741 ± 0.015	73.55 ± 0.39	73.42 ± 0.43
Concatenated single read	✓	✓	Joint	0.269 ± 0.009	0.750 ± 0.015	73.63 ± 0.60	73.53 ± 0.71
Fixed mixture	✓	✓	γ_0	0.270 ± 0.012	0.750 ± 0.023	73.31 ± 0.81	73.23 ± 0.73
No discrepancy	✓	✓	Learned	0.267 ± 0.010	0.753 ± 0.012	73.59 ± 0.40	73.48 ± 0.53
TEB	✓	✓	Learned	**0.265** ± 0.007	**0.761** ± 0.010	**74.13** ± 0.33	**74.18** ± 0.27

On CHERMA, full TEB exceeds latent-only by 0.75 WF1 points and fixed mixing by 0.95 points; both differences are positive in every matched seed. Its mean gains over local-only and the concatenated single read are 0.93 and 0.64 points, respectively, with four wins and one loss in each comparison. Removing c reduces mean WF1 by 0.70 points. TEB also has the smallest observed WF1 standard deviation among the six completed variants (0.27), although five runs do not establish a general stability guarantee. These observations favor the complete design in this setting.

On SIMS-V2, TEB reduces mean MAE relative to local-only, latent-only, concatenated single-read and fixed-mixture access by 0.01180, 0.00994, 0.00444 and 0.00502, respectively, using unrounded means. Removing c increases mean MAE by 0.00244 and lowers mean correlation by 0.00788. Its mean correlation gain over the concatenated control is 0.01124. TEB improves MAE in three of five matched seeds against concatenated access and in four of five against the no-discrepancy variant. These seed-wise differences are descriptive and do not establish benefits of separate retrieval or learned mixing in isolation. Unlike CHERMA, fixed mixing and concatenated access both improve mean MAE over either local-only or latent-only retrieval.

C.2 MELD Concatenated-Access Control

The MELD control uses ChatGLM3-6B, $B = 4$ and seeds 1111, 2222, 3333, 4444 and 5555. It reads the same local and latent candidate memories as TEB through one attention normalization. The single-read branch bypasses the learned mixing network, so this is a comparison of complete access paths rather than an active-parameter-matched isolation of separate reads.

Table 10: MELD concatenated-access control with ChatGLM3-6B and $B = 4$. Entries are five-seed mean ± sample standard deviation. The TEB row repeats the in-house reference in Table 2; higher values are better.

Access configuration	Acc ↑	WF1 ↑
Concatenated single read	65.48 ± 0.60	64.52 ± 0.29
TEB	**66.23** ± 0.54	**64.99** ± 0.33

TEB exceeds the concatenated read by 0.75 Acc and 0.47 WF1 points in the five-seed means. These differences do not establish a general benefit from separate normalization, because the two access paths use different active parameters and computations.

C.3 Limits of Input Attribution

Table 8 reports a retrained matched text-only control. Because it retains the native transcript, transcript-conditioned slots and evidence-token composer while removing audio-visual inputs, its large deficit shows that TEB's complete performance is not reproduced by the retained text-processing path alone. It does not establish how much audio and vision contribute individually, whether both are necessary for a given example, or which temporal observations cause a prediction.

No query-text substitution, audio-visual shuffling or separately trained unimodal controls are reported. Adapter-side text interventions would need to keep the native LLM transcript fixed and recompute all quantities derived from the substituted text. Inference-time substitutions introduce distribution shift and therefore address a different question from the retrained control. Attention maps, local-memory weights and discrepancy-conditioned gate statistics describe model behavior but do not establish causal evidence selection.

C.4 Prompt Parameterization and Resources

The factorized dictionary is shared by the controlled interfaces. The current results do not separately evaluate its contribution against a dense projection, nor the effect of using one query per output position against repeating a single generated token. Such a repeated-token variant would change retrieval capacity as well as output diversity. We therefore treat these choices as part of the adapter parameterization rather than independently demonstrated sources of improvement. Equal token counts also do not establish an advantage in trainable parameters (Appendix Table 5), peak memory, training time or inference latency.

C.5 Statistical Interpretation

The completed comparisons report five-seed means and sample standard deviations. Seed-wise differences are descriptive; no statistical significance tests or test-population confidence intervals are reported. Repeated training seeds on the same test set measure variation across runs, not independent draws of the test population. External paper-reported values are not included in paired comparisons.

Aggregate scores do not establish improved recognition of weak affect, correction of text-based errors or conversational reasoning. Those interpretations would require corresponding input controls or predefined subgroup analyses.

D SIMS-V2 Structured Calibration

The main-paper SIMS-V2 results use the base autoregressive objective and uncorrected evidence tokens. Here we evaluate structured calibration (SC) only as an optional extension of TEB with ChatGLM3-6B at $B = 4$. The two rows in Table 11 use the same five seeds; the base row repeats the main-paper result. SC combines structured auxiliary supervision with a residual correction to the prompt embeddings. It is not used in the other main-paper interface or memory controls, nor on MOSEI, MELD or CHERMA.

D.1 Effect on SIMS-V2

Table 11: Optional structured calibration for TEB on SIMS-V2 with ChatGLM3-6B and $B = 4$. Entries are mean $\pm$ sample standard deviation over matched seeds 1111, 2222, 3333, 4444 and 5555. The base row is the TEB result used throughout the main paper. Lower MAE and higher scores elsewhere are better.

Setting	Acc-2 ↑	F1 ↑	Acc-2^{weak} ↑	MAE ↓	Corr ↑
Base TEB	84.60 $\pm$ 0.96	84.50 $\pm$ 1.06	77.60 $\pm$ 1.74	0.265 $\pm$ 0.007	0.761 $\pm$ 0.010
TEB + SC	85.36 $\pm$ 0.79	85.33 $\pm$ 0.84	78.55 $\pm$ 1.12	0.264 $\pm$ 0.009	0.762 $\pm$ 0.013

Using unrounded five-seed means, SC raises Acc-2, F1 and weak Acc-2 by 0.75, 0.83 and 0.95 percentage points, respectively. It reduces MAE by 0.00088 and raises Corr by 0.00126; only two of five matched seeds improve on MAE and three of five on Corr. Thus, the regression changes are small and seed-dependent, while the classification means increase by about one point or less on the reported measures. These descriptive comparisons are not significance tests. The main-paper architectural comparisons use the base TEB setting.

D.2 Ordinal–Residual Structured Readout

Let $\bar{\mathbf{e}} = B^{-1}\sum_{b=1}^{B}\hat{\mathbf{s}}_b$ denote the mean refined query state. We discretize the SIMS-V2 score interval into K_o ordered centers $\{v_k\}_{k=1}^{K_o}$. A triangular or Gaussian kernel converts the continuous target y into a soft ordinal distribution

$$q_k(y) = \frac{\kappa(y, v_k)}{\sum_j \kappa(y, v_j)}. \tag{12}$$

An ordinal head predicts $p_k = \mathrm{softmax}(f_o(\bar{\mathbf{e}}))_k$ and is trained by

$$\mathcal{L}_{\mathrm{ord}} = -\sum_{k=1}^{K_o} q_k(y)\log p_k. \tag{13}$$

The ordinal expectation $\mu = \sum_k p_k v_k$ provides a coarse score estimate. A residual head adds a correction bounded in magnitude by $\Delta/2$,

$$\tilde{y} = \mu + \frac{\Delta}{2}\tanh(f_r(\bar{\mathbf{e}})), \qquad \mathcal{L}_{\mathrm{res}} = \mathrm{SmoothL1}(\tilde{y}, y), \tag{14}$$

where Δ is the spacing between adjacent ordinal centers. An optional cumulative-distribution term compares the predicted and target ordinal distributions through their cumulative sums. These heads provide training-time structure and do not replace the frozen LLM prediction.

D.3 Sign- and Boundary-Sensitive Objectives

Near-zero scores are both important and ambiguous in SIMS-V2. The polarity objective therefore ignores a narrow dead zone and increases supervision smoothly with label magnitude. With thresholds $\epsilon_0 < \epsilon_1$, the sample weight is

$$w_{\mathrm{pol}}(y) = \mathrm{clip}\left(\frac{|y| - \epsilon_0}{\epsilon_1 - \epsilon_0}, 0, 1\right), \tag{15}$$

and $\mathcal{L}_{\mathrm{pol}}$ is the corresponding weighted binary cross-entropy for the target $\mathbb{I}[y > 0]$. A weak-sign head operates only inside a prescribed near-neutral sentiment band, excluding the dead zone. It uses the soft target $\sigma(y/\tau_s)$ and gives greater weight to samples closer to the center of that band.

Boundary-weighted pairwise ranking complements per-sample supervision. For a positive sample i and a negative sample j outside the dead zone, define

$$u(y) = \sqrt{\mathrm{clip}\left(\frac{|y| - \epsilon_0}{\epsilon_f - \epsilon_0}, 0, 1\right)\mathrm{clip}\left(1 - \frac{|y|}{\epsilon_f}, 0, 1\right)} \tag{16}$$

and $w_{ij} = \max(u(y_i), u(y_j))$. The pairwise objective is

$$\mathcal{L}_{\mathrm{pair}} = \frac{\sum_{i,j} w_{ij}\tau_b \log\left[1 + \exp\left(\frac{m_{ij} - (\tilde{y}_i - \tilde{y}_j)}{\tau_b}\right)\right]}{\sum_{i,j} w_{ij}}, \tag{17}$$

where $m_{ij} = m_0 + m_1 \min(|y_i|, |y_j|, \epsilon_f)$. This term emphasizes sign ordering for weak but non-ambiguous positive–negative pairs.

D.4 Structured-to-Prompt Correction

When enabled, the structured readout also produces a residual correction to the multimodal prompt tokens. Its context concatenates $\bar{\mathbf{e}}$ with the structured score, polarity and weak-sign logits, an entropy-derived concentration score, the discrepancy score and mean local–latent gate. The concentration score is $u_o = 1 - H(\mathbf{p})/\log K_o$, with $H(\mathbf{p}) = -\sum_k p_k \log p_k$; it is not a calibrated probability of correctness. A coefficient network and a separate dictionary generate $\Delta\mathbf{P} \in \mathbb{R}^{B\times d_{\mathrm{LLM}}}$. The corrected prompt is

$$\mathbf{P}' = \mathbf{P} + \lambda_c \, \mathbf{G}_{\mathrm{tok}} \odot \mathbf{G}_{\mathrm{conf}} \odot \Delta\mathbf{P}, \tag{18}$$

where $\mathbf{G}_{\mathrm{tok}}$ is a learned token-wise gate and $\mathbf{G}_{\mathrm{conf}}$ is derived from u_o. The prompt-correction scale is bounded, and the structured context can be detached before the correction network. Optional prompt-level regression, sign, weak-band and ordered-anchor terms regularize the corrected token set.

D.5 Objective and Isolation Protocol

When SC is enabled, the SIMS-V2 training objective can be written as

$$\mathcal{L} = \mathcal{L}_{\mathrm{LLM}} + \lambda_{\mathrm{aux}}\mathcal{L}_{\mathrm{readout}} + \mathcal{L}_{\mathrm{prompt}} + \lambda_{\mathrm{pair}}\mathcal{L}_{\mathrm{pair}}, \tag{19}$$

where $\mathcal{L}_{\mathrm{readout}}$ is a weighted combination of the active ordinal, residual, distribution, polarity and weak-sign terms, and $\mathcal{L}_{\mathrm{prompt}}$ collects the active prompt-level terms. A coefficient of zero disables the corresponding component.

D.6 Memory Access and Adaptive Mixing

At $B = 4$, the completed variants share the local and latent memory candidates where applicable. Concatenated access reads both candidate sets under one attention normalization; fixed mixing retains separate reads but uses a constant weight. The no-discrepancy variant omits only the scalar c from the utterance context (Appendix Eq. (9)), retaining pooled audio-visual states, both memory reads and learned mixing. Table 4 reports both metrics per dataset; Appendix Table 9 additionally identifies the candidate sets and access/mixing rules. Apart from no-discrepancy, these controls change multiple operations and cannot uniquely isolate separate reads or the adaptive gate.

On CHERMA, TEB exceeds concatenated access by 0.50 Acc and 0.64 WF1 points using unrounded means, winning four of five matched WF1 seeds; it exceeds the tested fixed mixture by 0.95 WF1 points. Removing c lowers mean WF1 from 74.18 to 73.48. On SIMS-V2, TEB reduces mean MAE by 0.00444 and increases mean Corr by 0.01124 relative to concatenated access, using unrounded five-seed means; it has lower MAE in three of five matched seeds. Removing c increases mean MAE from 0.265 to 0.267. The complete configuration therefore has clearer same-candidate support on CHERMA. On MELD, TEB reaches 64.99 WF1 versus 64.52 for a concatenated single read (five-seed means; Appendix Table 10). This 0.47-point mean difference compares complete access paths and does not isolate separate reads at matched active parameter capacity.

E Discussion and Limitations

Broad benchmarks and equal-B comparisons support the complete interface within our protocol. Against the concatenated single-read control, TEB improves CHERMA WF1 in four of five matched seeds and SIMS-V2 MAE in three of five; MELD shows a 0.47-point mean WF1 advantage, with only aggregate five-seed results reported. Removing c increases mean SIMS-V2 MAE and decreases mean CHERMA WF1; the SIMS-V2 difference is small relative to seed variation.

The targeted controls use one backbone. Equal B matches LLM-facing tokens, not adapter parameters or computation. Concatenation bypasses the learned mixer and changes attention normalization, so active capacity differs. The retrained text-only control supports the audio-visual route without separating modalities or identifying causal segments; removing c tests only one context input. Active-capacity matching, modality interventions, resource use, token specialization, dialogue history and missing-modality robustness remain untested.