CaM-HG: Causal-Enhanced MoE and Hypergraphs Network for Incomplete Multimodal Emotion Recognition in Conversations

Mingjian Yang Yong Wang Peng Liu Wen Yin*
The Laboratory of Intelligent Collaborative Computing,
University of Electronic Science and Technology of China, Chengdu, China
{ymj0420, yinwen1999}@std.uestc.edu.cn
cla@uestc.edu.cn,lp@uestc.edu.com

Abstract

Multimodal Emotion Recognition in Conversation (MERC) relies on integrating heterogeneous signals, yet real-world modality missingness frequently disrupts these systems. We contend that missingness is not merely a loss of data fidelity but a rupture of the fine-grained inter-modal causal chains essential for reasoning. Existing methods, which primarily focus on statistical reconstruction, often fail to bridge these logical gaps, effectively leaving semantic holes. To address this, we propose the **Causal-Enhanced Mixture-of-Experts and Hypergraph Network (CaM-HG)**, employing a "restore-then-mine" paradigm. First, a Causal-Enhanced MoE module conditions experts on historical context to synthesize missing features that are both realistic and causally consistent, thereby patching the broken topology. Subsequently, an Asymmetric Causal Dynamic Hypergraph mines high-order correlations from the restored graph while enforcing strict temporal causality. Experiments on IEMOCAP, CMU-MOSI, and CMU-MOSEI show consistent improvements in terms of WAF1 and accuracy over strong baselines, e.g., surpassing SOTA benchmarks by 1.43% and 1.25% on IEMOCAP.

1 Introduction

Multimodal Emotion Recognition in Conversation (MERC) empowers machines to decipher complex emotional states through the synergistic fusion of textual, acoustic, and visual signals (Ma et al., 2023; Wang et al., 2025). While prevailing approaches achieve remarkable performance under ideal, complete-data conditions through advanced fusion or consistency-aware learning (Ai et al., 2025; Zhang and Tan, 2025; Lyu et al., 2025; Yin et al., 2026), real-world deployments are frequently plagued by unpredictable modality missingness—ranging from sensor failures and noise to privacy constraints (Wang et al., 2023b; Solanki et al., 2020). In such imperfect environments, the efficacy of established models tends to collapse catastrophically, severely limiting the reliability of MERC systems (Lian et al., 2023; Liu et al., 2024).

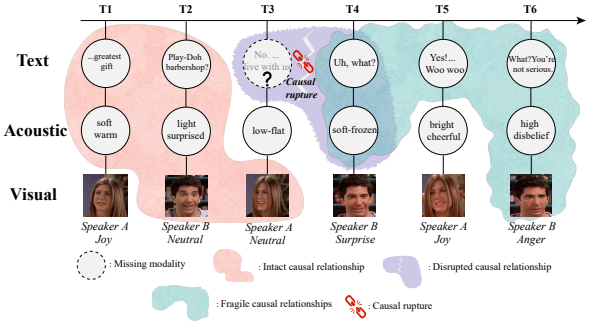

Figure 1: Illustration of "Causal Rupture." The missing visual modality at T3 creates a direct rupture, leaving the reaction at T4 ungrounded. This disruption propagates downstream, rendering subsequent interactions (T5-T6, green area) fragile. Consequently, the historical "storyline" is corrupted, destabilizing future reasoning even if data is complete.

To mitigate modality missingness, existing research focuses on feature reconstruction and robust representation. Generative approaches, spanning from autoencoders (Zhao et al., 2021a) to recent diffusion models (Zhu et al., 2025; Shou et al., 2025), strive to recover missing views with statistical fidelity. In parallel, representation learning frameworks utilize contrastive strategies to extract modality-invariant features resilient to input sparsity (Liu et al., 2024; Li et al., 2025). However, while state-of-the-art methods effectively capture contextual dependencies, they primarily rely on associative correlations rather than causal consistency. Even causal-aware approaches like ECERC (Zhang and Tan, 2025) are restricted to complete modalities and utterance-level interactions. Consequently, existing frameworks fail to address the critical causal ruptures induced by missingness, which fundamen-

* Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 16501–16516

tally disrupts the logical continuity essential for emotional reasoning (Fu et al., 2025).

Crucially, we argue that modality missingness fundamentally invalidates coarse-grained utterance-level modeling, as the true drivers of emotional shifts are often specific, fine-grained inter-modal cues rather than abstract sentence vectors. Consequently, the unavailability of a modality constitutes a *Fine-grained Inter-modal Causal Rupture*, where the specific trigger required for reasoning is severed. As illustrated in Figure 1, the absence of a visual cue at T3 eliminates the causal anchor needed to interpret the interlocutor's reaction at T4, leaving the subsequent emotion ungrounded. This disruption propagates downstream, rendering future interactions (T5-T6) fragile; even if subsequent data is complete, the reasoning foundation remains unstable as the historical storyline is corrupted (Poria et al., 2021; Fu et al., 2025). Motivated by this, a pivotal question arises: *How can we explicitly restore these ruptured fine-grained inter-modal causal relationships?*

To address this, we contend that a robust framework must fulfill two critical imperatives: (1)Structural Causal Restoration. (2)High-order Causal Mining. Instead of simple statistical filling, it is essential to regenerate missing endpoints conditioned on historical context. This ensures topological completeness by eliminating null nodes (You et al., 2020), while aligning restored semantics with the dialogue flow to provide valid causal anchors for reasoning. Building upon this restored topology, we further argue that standard pairwise graphs oversimplify emotional dynamics by restricting interactions to dyadic connections. In contrast, modeling multimodal contexts as high-order sets (e.g., via hypergraphs) allows us to aggregate dispersed historical cues into unified causal units (Fu et al., 2025; Fixelle, 2025). This uniquely captures combinatorial causal patterns, allowing dispersed cues to synergistically trigger an emotion, thereby surpassing the limitations of simple pairwise propagation.

To this end, we propose the Causal-Enhanced Mixture-of-Experts and HyperGraph Network (**CaM-HG**). Adhering to a "Restore-then-Mine" paradigm, the framework operates in two cohesive phases: (1) *Causal-Enhanced MoE Imputation*: To resolve the "one-to-many" ambiguity in emotion restoration, we design a conditional MoE generator. Explicitly guided by historical context, it synthesizes semantically grounded anchors to bridge the logical gaps shown in Figure 1. (2) *Dynamic Causal Hypergraph Fusion*: Upon these anchors, we employ an Asymmetric Causal Dynamic Hypergraph. A novel *Block Causal Mask* is introduced to reconcile multimodal simultaneity with strict temporal constraints, enabling high-order reasoning without future information leakage. Our main contributions are summarized as follows:

- We reconceptualize modality missingness as a topological rupture of fine-grained causal chains, rather than mere data sparsity. This perspective identifies structural restoration as the core bottleneck
- We propose **CaM-HG**, which couples a **Causal-Enhanced MoE** for reliable anchor regeneration with an **Asymmetric Causal Dynamic Hypergraph** to restore and model fine-grained causal dependencies across conversational turns.
- Experiments on IEMOCAP, CMU-MOSI, and CMU-MOSEI demonstrate consistent superiority, notably achieving **1.43% and 1.25% WAF1 gains** over SOTA on IEMOCAP-4/6 with enhanced robustness.

2 Related Work

Paradigms for Incomplete Multimodal Learning Existing studies on incomplete multimodal learning generally fall into three categories: feature reconstruction, robust representation, and pairwise graph fusion. Generative approaches aim to restore data fidelity by synthesizing missing features. Early methods like MMIN (Zhao et al., 2021a) utilize cascaded autoencoders,while recent diffusion-based models such as RMER-DT (Zhu et al., 2025), GSDNet (Shou et al., 2025), and KCDP (Yin et al., 2025) have set new benchmarks by preserving original data distributions. Representation learning frameworks, such as CIF-MMIN (Liu et al., 2024) and UMAP (Li et al., 2025), employ contrastive strategies to learn modality-invariant features that remain stable despite missingness. To capture structural dependencies, pairwise graph methods have also been adopted. For instance, GCNet (Lian et al., 2023) designs a graph completion module to propagate information across temporal and speaker nodes. While these methods excel at recovering statistical fidelity or modeling pairwise interactions, they exhibit limitations. Generative models often neglect the fine-grained causal consistency required for dialogue logic, while standard graph approaches are

restricted to pairwise connections, failing to capture the complex, high-order synergies inherent in multimodal communication.

Hypergraph Learning in Multimodal Analysis To overcome the limitations of pairwise graphs, hypergraphs have emerged as a powerful paradigm for modeling high-order correlations. SDR-GNN (Fu et al., 2025) leverages spectral domain reconstruction on hypergraph topologies to recover missing signals. Recent advancements have further expanded this capability. DIB-HGCN (Chen and Shi, 2025) constructs dynamic bimodal hypergraphs to distinguish between monologic and dialogic interaction patterns, while MATCH (Shi et al., 2025) introduces a modality-calibrated hypergraph network to align emotional pathways. HyperCRM (Lu et al., 2024) models conversations as hypergraphs with speaker-level and sequence-level hyperedges, employing three-stage propagation to capture long-range context while reducing redundant dependencies. In the broader field of computer vision, HgVT (Fixelle, 2025) demonstrates the versatility of hyperedges in modeling complex object relations. However, standard symmetric hypergraphs neglect the temporal causality essential for restoration, risking future leakage. In contrast, our Asymmetric Causal Dynamic Hypergraph employs a Block Causal Mask to enforce strict directionality, ensuring high-order synergies are mined exclusively on a logically valid, restored topology without compromising multimodal simultaneity.

3 Methodology

3.1 Task Definition

We study Multimodal Emotion Recognition in Conversation (MERC) under incomplete modality observations. A dialogue is denoted as $\mathcal{D} = \{u_1, \ldots, u_L\}$ with L time-ordered utterances, where each utterance u_t is associated with an emotion label $y_t \in \mathcal{Y}$.

For each utterance u_t, we consider tri-modal features from acoustic (a), visual (v), and linguistic (l) modalities. Let $\mathbf{x}_t = \{\mathbf{x}_t^m \mid m \in \{a, v, l\}\}$ be the complete feature set, where $\mathbf{x}_t^m \in \mathbb{R}^{d_m}$. To model modality missingness, we introduce a binary availability vector $\boldsymbol{\delta}_t = [\delta_t^a, \delta_t^v, \delta_t^l]^\top \in \{0, 1\}^3$. Formally, the observed input is defined as $\tilde{\mathbf{x}}_t^m = \mathbf{x}_t^m$ if $\delta_t^m = 1$, and $\tilde{\mathbf{x}}_t^m = \emptyset$ if $\delta_t^m = 0$. This formulation distinguishes our approach from standard methods that merely treat missing modalities as zero-padded tensors. We define the state $\delta_t^m = 0$ as a Topological Rupture, signifying that the corresponding node is functionally absent from the dialogue graph, which severs the fine-grained causal chains required for reasoning.

We denote the incomplete dialogue observation as $\{(\tilde{\mathbf{x}}_t, \boldsymbol{\delta}_t)\}_{t=1}^L$. The goal is to learn a predictor f_θ that maps this incomplete dialogue to per-utterance emotion predictions, i.e., $f_\theta : \{(\tilde{\mathbf{x}}_t, \boldsymbol{\delta}_t)\}_{t=1}^L \rightarrow \{\hat{y}_t\}_{t=1}^L$. Given the incomplete input, the model f_θ is expected to *restore* the missing semantics in the ruptured topology and leverage the dialogue context to make robust predictions.

3.2 The Proposed CaM-HG

In this section, we elaborate on our proposed **Causal-Enhanced Mixture-of-Experts and Hypergraph Network (CaM-HG)**. As illustrated in Figure 2, the framework comprises three cohesive stages: Input Preprocessing (§3.2.1), Causal-Enhanced MoE Imputation (§3.2.2), and Asymmetric Causal Dynamic Hypergraph (§3.2.3).

3.2.1 Input Preprocessing

To ensure robust representations for causal reasoning, we follow MERC benchmarks (Fu et al., 2025; Qiu et al., 2025) and extract features using pre-trained experts: DeBERTa-large (He et al., 2020) for text ($\mathbf{x}_t^l \in \mathbb{R}^{d_l}$), Wav2Vec 2.0 (Schneider et al., 2019) for audio ($\mathbf{x}_t^a \in \mathbb{R}^{d_a}$), and MANet (Zhao et al., 2021b) for visual signals ($\mathbf{x}_t^v \in \mathbb{R}^{d_v}$).

To mitigate overfitting to dominant modalities (e.g., text), we employ a hybrid masking strategy during training. This comprises independent Bernoulli sampling for each modality and a targeted blindness mechanism that explicitly drops the textual modality with a probability of $p = 0.25$. This constraint forces the model to mine subtle cues from acoustic and visual signals, enhancing robustness against real-world textual missingness.

We map the masked inputs $\tilde{\mathbf{x}}_t^m$ into a unified latent space of dimension d using modality-specific linear projections, leveraging the rich semantics already captured by pre-trained extractors:

$$\mathbf{h}_t^m = \text{LayerNorm}(\mathbf{W}_m \tilde{\mathbf{x}}_t^m + \mathbf{b}_m), \quad (1)$$

where $\mathbf{W}_m$ and $\mathbf{b}_m$ are learnable parameters, $m \in \{a, v, l\}$ indicates the modality. Crucially, we maintain the fine-grained structure of separate nodes rather than performing early concatenation. To distinguish interlocutor and modality identities within this shared space, we inject structural identifiers:

$$\tilde{\mathbf{h}}_t^m = \mathbf{h}_t^m + \mathbf{E}_{spk}(s_t) + \mathbf{P}_{mod}(m), \quad (2)$$

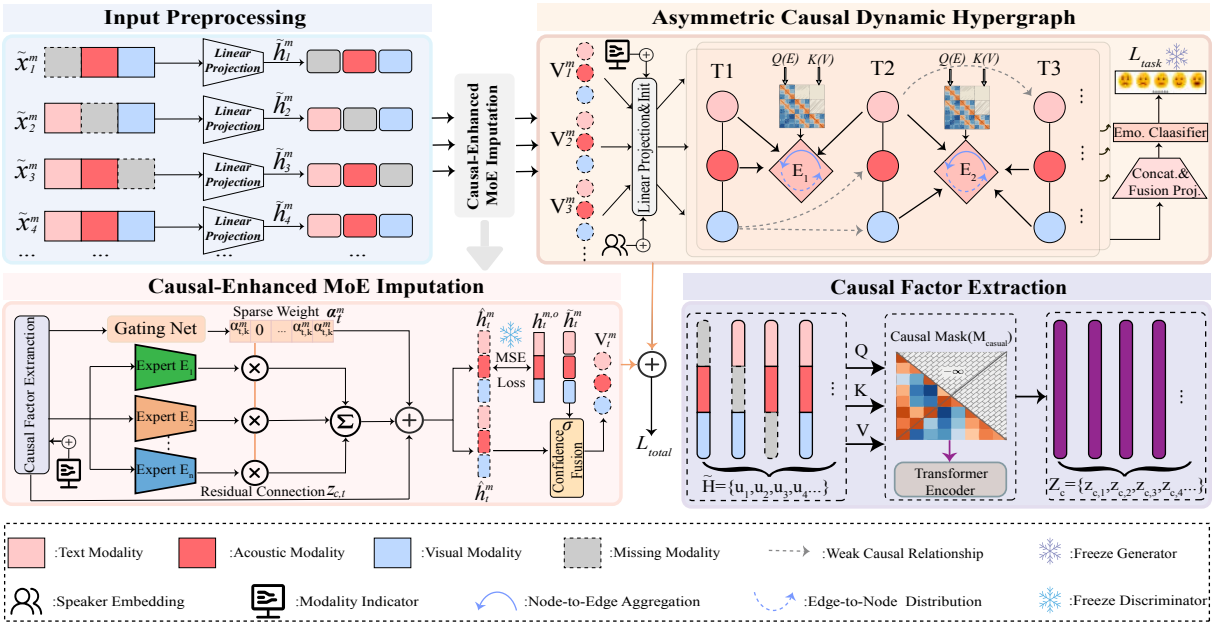

Figure 2: **The overall architecture of CaM-HG.** The framework comprises three cohesive stages: (1) **Input Preprocessing** (§3.2.1), which extracts raw multimodal features and projects them into a unified latent space; (2) **Causal-Enhanced MoE Imputation** (§3.2.2), which internally employs a **Causal Factor Extraction** module to capture historical triggers, guiding the Mixture-of-Experts to restore missing modalities with causal consistency; and (3) **Asymmetric Causal Dynamic Hypergraph** (§3.2.3), which initializes dynamic hyperedges and models high-order correlations via a block causal mask to prevent information leakage.

where $\mathbf{E}_{spk}$ and $\mathbf{P}_{mod}$ denote the learnable embeddings for speaker identity and modality type, respectively. Here, $s_t \in \{A, B\}$ represents the speaker index at turn t. The resulting $\tilde{\mathbf{h}}_t^m$ serves as the initialized node representation. For missing modalities ($\tilde{\mathbf{x}}_t^m = \emptyset$), $\mathbf{h}_t^m$ acts as a zero-initialized learnable placeholder anchor, ready to be restored by the subsequent imputation module.

3.2.2 Causal-Enhanced MoE Imputation

Standard imputation methods typically utilize deterministic generators, which struggle with the inherent "one-to-many" mapping of human emotions (i.e., yielding averaged, smoothed features). To capture diverse emotional variants while strictly adhering to causality, we propose a unified imputation module.

Causal Factor Extraction. Effective imputation requires a global historical perspective. We first construct a unified time-step representation by fusing the multi-view features via a linear projection $\mathbf{W}_{fuse} \in \mathbb{R}^{d \times 3d}$:

$$\mathbf{u}_t = \mathbf{W}_{fuse}[\text{Concat}(\tilde{\mathbf{h}}_t^a, \tilde{\mathbf{h}}_t^v, \tilde{\mathbf{h}}_t^l)] \in \mathbb{R}^d. \quad (3)$$

Let $\tilde{\mathbf{H}} = [\mathbf{u}_1, \ldots, \mathbf{u}_L] \in \mathbb{R}^{L \times d}$ denote the fused dialogue sequence matrix. To extract causal context, we employ a standard Transformer encoder. Crucially, to prohibit future information leakage, we enforce a strict Causal Mask $\mathbf{M}_{causal}$ on the attention mechanism:

$$\mathbf{M}_{causal}[i, j] = \begin{cases} 0, & \text{if } i \geq j \\ -\infty, & \text{otherwise} \end{cases}, \quad (4)$$

The context extraction is thus formulated as $\mathbf{Z}_c = \text{Transformer}(\tilde{\mathbf{H}}, \tilde{\mathbf{H}}, \tilde{\mathbf{H}}, \mathbf{M}_{causal})$. The output $\mathbf{z}_{c,t} \in \mathbb{R}^d$ serves as a "contextual anchor," aggregating the history $\{u_1, \ldots, u_t\}$ to guide the regeneration.

Mixture-of-Experts (MoE) Generation. To mitigate the mode-collapse issue, we utilize a MoE generator composed of N experts $\{E_k\}_{k=1}^N$, where each expert is implemented as a Multi-Layer Perceptron (MLP). Unlike standard imputation that yields deterministic outputs, our MoE generates modality-specific features conditioned on both the history and the target modality type. Specifically, we define the query vector for the m-th modality at time t as $\mathbf{q}_t^m = \mathbf{z}_{c,t} + \mathbf{P}_{mod}(m)$, where $\mathbf{P}_{mod}(m)$ is the modality embedding defined in Eq. 2. A gating network $\mathbf{W}_g \in \mathbb{R}^{N \times d}$ then computes sparse routing scores to activate the top-K most relevant experts:

$$\begin{aligned} \mathcal{I}_t^m &= \text{TopK}(\mathbf{W}_g \mathbf{q}_t^m, K), \\ \boldsymbol{\alpha}_t^m &= \text{Softmax}(\{(\mathbf{W}_g \mathbf{q}_t^m)_k\}_{k \in \mathcal{I}_t^m}), \end{aligned} \quad (5)$$

where $\mathcal{I}_t^m$ contains the indices of the selected experts, $\boldsymbol{\alpha}_t^m$ denotes the normalized gating weights, and $(\cdot)_k$ represents the k-th element of the vector. The restored feature $\hat{\mathbf{h}}_t^m$ is synthesized by the weighted ensemble of selected experts combined with a residual connection to the context anchor:

$$\hat{\mathbf{h}}_t^m = \sum_{k \in \mathcal{I}_t^m} \alpha_{t,k}^m \cdot E_k(\mathbf{q}_t^m) + \mathbf{z}_{c,t}. \quad (6)$$

By conditioning the experts on $\mathbf{q}_t^m$, the model dynamically adjusts its generation logic, preventing semantic homogenization across modalities.

Confidence-Aware Feature Fusion. To mitigate generation artifacts, we employ a confidence-based fusion strategy. We predict a gate vector $\boldsymbol{\sigma}_t = \sigma(\text{MLP}([\hat{\mathbf{h}}_t^m \oplus \mathbf{z}_{c,t}])) \in (0,1)^d$ conditioned on both the generated candidate and context anchor, where $\sigma(\cdot)$ denotes the Sigmoid activation. The final feature representation is then synthesized via:

$$\mathbf{v}_t^m = \boldsymbol{\sigma}_t \odot \hat{\mathbf{h}}_t^m + (\mathbf{1} - \boldsymbol{\sigma}_t) \odot \tilde{\mathbf{h}}_t^m. \quad (7)$$

This mechanism dynamically calibrates the fusion: it prioritizes the generator ($\boldsymbol{\sigma}_t \to \mathbf{1}$) for missing modalities while preserving raw signals ($\boldsymbol{\sigma}_t \to \mathbf{0}$) when observed, yielding robust node initialization for the subsequent hypergraph.

3.2.3 Asymmetric Causal Dynamic Hypergraph

Existing graph-based MERC methods often rely on static structures or symmetric interactions, which fail to capture the evolving nature of dialogue and inevitably leak future information. To address these limitations, we propose the Asymmetric Causal Dynamic Hypergraph (ACDH).

Hypergraph Construction. We construct a hypergraph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The node set $\mathcal{V} = \{\mathbf{v}_t^m\}_{t=1,m\in\{a,v,l\}}^{L}$ is initialized with the fused features, comprising $3L$ nodes arranged in temporal order. Unlike static approaches with fixed global connections, we employ a content-aware initialization strategy. Specifically, we dynamically project the node representations to form a set of latent hyperedges $\mathcal{E}^{(0)} \in \mathbb{R}^{3L\times d}$ using a 2-layer Multi-Layer Perceptron (MLP) with a GELU activation:

$$\mathcal{E}^{(0)} = \text{MLP}_2(\text{GELU}(\text{MLP}_1(\mathcal{V}))). \quad (8)$$

This formulation ensures each of the $3L$ nodes projects a latent hyperedge encapsulating its modality, temporal, and speaker information. These hyperedges then serve as evolving semantic anchors that aggregate and redistribute information.

Block Causal Mask. A core innovation of ACDH is the Block Causal Mask, designed to reconcile sequential causality with multimodal simultaneity. To avoid erroneously blocking concurrent cross-modal interactions, we define a block-wise mask $\mathbf{M}_{block} \in \mathbb{R}^{3L\times 3L}$ based on time-step indices:

$$\mathbf{M}_{block}[i,j] = \begin{cases} 0, & \text{if } \lfloor i/3 \rfloor \geq \lfloor j/3 \rfloor \\ -\infty, & \text{otherwise} \end{cases}. \quad (9)$$

This enforces a block-triangular structure, enabling bidirectional inter-modal fusion and historical attention, while strictly shielding future blocks.

Dual-Stage Asymmetric Convolution. Hypergraph convolution is performed in two asymmetric stages via a Transformer backbone. We stack this operation for L_{hyp} layers to capture high-order correlations. Crucially, we enforce the Block Mask throughout to maintain strict causality:

$$\begin{aligned} \mathcal{E}^{(l)} &= \text{Transformer}(\mathcal{E}^{(l-1)}, \mathcal{V}^{(l-1)}, \mathcal{V}^{(l-1)}, \mathbf{M}_{block}), \\ \mathcal{V}^{(l)} &= \text{Transformer}(\mathcal{V}^{(l-1)}, \mathcal{E}^{(l)}, \mathcal{E}^{(l)}, \mathbf{M}_{block}). \end{aligned} \quad (10)$$

The process consists of: (1) *Node-to-Edge Aggregation*: Hyperedges update their representations by aggregating information from past and current nodes, effectively compressing complex multi-view interactions into global latent descriptors. (2) *Edge-to-Node Distribution*: Updated hyperedges redistribute high-order semantics back to the nodes, enriching the local time-step representations with structurally aware global context. By consistently applying $\mathbf{M}_{block}$, nodes are restricted to query only historically valid hyperedges. This guarantees strict causal consistency throughout the message passing and prevents future leakage.

Multi-stage Optimization Strategy. Training coupled generative-discriminative frameworks on incomplete data is notoriously unstable due to the competing objectives. To mitigate this, we employ a three-stage Curriculum Learning strategy:

(1) **Discriminative Warm-up**: We first train the Input Projector, Hypergraph, and Classifier solely on the available (observed) modalities. This establishes a meaningful and stationary latent decision boundary before introducing generation.

(2) **Generative Pre-training**: We freeze the pre-trained Input Projector and focus exclusively on optimizing the MoE generator via reconstruction loss. This prevents the "moving target" problem, ensuring the generator learns to reconstruct features

within a stable semantic space defined by the frozen encoder.

(3) **Joint Fine-tuning**: Finally, we optimize the entire framework end-to-end. The total objective $\mathcal{L}_{total}$ combines the task-specific classification loss $\mathcal{L}_{task}$ (Cross-Entropy) and a sparsity-aware dynamic reconstruction loss:

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda(1+\eta_{batch}) \sum_{t,m} \|\hat{\mathbf{h}}_t^m - \mathbf{h}_t^{m,\mathrm{o}}\|^2, \tag{11}$$

where λ is a base hyperparameter and η_{batch} denotes the missing rate of the current batch. This dynamic scaling explicitly imposes a heavier penalty on reconstruction errors in high-sparsity scenarios. Crucially, the target $\mathbf{h}_t^{m,\mathrm{o}}$ is derived from the original unmasked input, creating a self-supervised signal that forces the generator to recover authentic semantics rather than mimicking zero-padded placeholders.

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets

We evaluate our approach using the three widely-used benchmark conversational datasets: IEMOCAP (Busso et al., 2008), CMU-MOSI (Zadeh et al., 2016), and CMU-MOSEI (Zadeh et al., 2018). Statistical details of the datasets are provided in Table 1.

We evaluate on three standard benchmarks. **IEMOCAP** features 151 dyadic conversations across five sessions. Following protocols (Fu et al., 2025), we use Session 5 for testing on both the 6-class (7,433 utterances) and 4-class (merging *excited* to *happy*) settings. For sentiment analysis, we employ **CMU-MOSI**, containing 2,199 video segments annotated on a 7-point scale ($[-3, +3]$). We adopt the standard split of 1,284/229/686 for training/validation/testing. Furthermore, we extend to the large-scale **CMU-MOSEI** dataset, comprising 23,453 utterances from over 1,000 speakers, to assess robustness.

For evaluation, we report binary accuracy (*Acc-2*) and weighted average F1-score (WAF1) for both sentiment datasets (MOSI/MOSEI), while using the weighted average F1-score for IEMOCAP.

4.1.2 Baselines

To verify the effectiveness of our proposed CaM-HG, we compare it against a comprehensive set of state-of-the-art baselines designed for incomplete multimodal learning. Specifically, we select DCCAE (Wang et al., 2015) and CRA (Tran et al., 2017) as representative generative imputation methods that reconstruct missing views via autoencoders or cascaded residuals. For robust representation learning, we include CIF-MMIN (Liu et al., 2024), DiCMor (Wang et al., 2023a), and IMDER (Wang et al., 2023b), which aim to learn modality-invariant features or maximize mutual information to mitigate missingness. Furthermore, we incorporate graph and diffusion-based approaches, including GCNet (Lian et al., 2023) and SDR-GNN (Fu et al., 2025) which utilize graph topology for feature completion, and FedDISC (Qiu et al., 2025) which employs federated diffusion for semantic consistency. These models represent a diverse range of advanced methodologies in the field. In our experiments, we strictly follow their official implementations and evaluation protocols to ensure a fair and rigorous comparison across the IEMOCAP, CMU-MOSI, and CMU-MOSEI datasets.

Table 1: Statistics of the datasets.

Dataset	# Conversations			# Utterances			# Cls
	Train	Val	Test	Train	Val	Test	
IEMOCAP-6	120		31	5810		1623	6
IEMOCAP-4	120		31	4290		1241	4
CMU-MOSI	52	10	31	1284	229	686	2
CMU-MOSEI	2249	300	676	16326	1871	4659	2

4.1.3 Settings

Model Configurations. Following our methodology, the unified hidden dimension d is set to 512 for IEMOCAP and 256 for CMU-MOSI and CMU-MOSEI. The *Causal Factor Extractor* consists of a 2-layer Transformer encoder with 8 attention heads. For the *Causal-Enhanced MoE*, we employ $N = 4$ experts and set the top-K gating to $K = 2$. The *Asymmetric Causal Dynamic Hypergraph* utilizes 2 message-passing layers to capture high-order correlations. Dropout rates are set between 0.3 and 0.5 across different datasets to prevent overfitting.

Training and Optimization. We implement the framework in PyTorch using the AdamW optimizer with a batch size of 32. Learning rates are initialized at 2×10^{-4} and reduced to $[2\times10^{-5}, 8\times10^{-5}]$ for joint optimization. To improve generalization, we apply label smoothing with a factor of 0.05. The reconstruction weight λ_{rec} varies from 0.05 to 0.5 depending on the dataset. Notably, instead of

Table 2: Comparison of performance (WAF1 %) with various missing rates on IEMOCAP. The best performance is highlighted in **bold**, and the second best is underlined. * Results come from SDR-GNN (Fu et al., 2025)

Dataset	Method	Missing Rate (η)								Average
		0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	
IEMOCAP (four-class)	DCCAE* (Wang et al., 2015)	63.42	61.66	57.67	54.95	51.08	45.71	39.07	41.42	51.87
	CRA* (Tran et al., 2017)	76.26	71.28	67.34	62.24	57.04	49.86	43.22	38.56	58.23
	MMIN* (Liu et al., 2024)	74.94	71.84	69.36	66.34	63.30	60.54	57.52	55.44	64.91
	GCNet* (Lian et al., 2023)	78.36	77.48	77.34	76.22	75.14	73.80	71.88	71.38	75.20
	SDR-GNN* (Fu et al., 2025)	79.58	78.55	78.08	77.53	77.09	75.84	75.03	74.41	77.01
	FedDISC (Qiu et al., 2025)	78.40	78.12	77.67	77.05	76.43	75.92	75.38	75.04	76.75
	Ours	**80.73**	**79.89**	**79.74**	**79.19**	**77.96**	**77.48**	**76.45**	**76.06**	**78.44**
IEMOCAP (six-class)	DCCAE* (Wang et al., 2015)	46.19	43.77	41.28	37.98	34.58	30.02	26.78	27.66	36.03
	CRA* (Tran et al., 2017)	58.68	53.50	49.76	45.88	39.94	32.88	28.08	26.16	41.86
	MMIN* (Liu et al., 2024)	56.96	53.94	51.46	48.42	45.60	42.82	40.18	37.84	47.15
	GCNet* (Lian et al., 2023)	58.64	58.50	57.64	57.08	56.12	54.40	53.60	53.46	56.18
	SDR-GNN* (Fu et al., 2025)	61.34	60.86	59.83	59.49	59.16	57.38	55.51	55.26	58.60
	FedDISC (Qiu et al., 2025)	**64.70**	61.99	60.91	59.55	59.19	57.48	55.72	55.50	59.38
	Ours	63.68	**62.55**	**61.93**	**60.90**	**60.10**	**59.00**	**58.71**	**58.19**	**60.63**

Table 3: Performance comparison (Acc-2 / WAF1 %) with various missing rates on CMU-MOSI and CMU-MOSEI. Best results are in **bold**, second best are underlined. * Results come from FedDISC (Qiu et al., 2025).

Dataset	Method	Missing Rate (η)								Average
		0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	
MOSI	DCCAE* (Wang et al., 2015)	77.3/77.4	74.5/74.7	71.8/71.9	67.0/66.7	63.6/62.8	62.0/61.3	59.6/58.5	58.1/57.4	66.7/66.3
	GCNet* (Lian et al., 2023)	85.1/85.2	82.3/82.3	79.5/79.4	77.2/77.2	74.4/74.3	69.8/70.0	66.7/67.7	65.4/65.7	75.1/75.2
	DiCMor* (Wang et al., 2023a)	85.6/85.7	83.9/83.9	82.0/82.1	80.2/80.4	77.7/77.9	76.4/76.7	73.0/73.3	70.8/71.1	78.7/78.9
	IMDER* (Wang et al., 2023b)	85.7/85.6	84.9/84.8	**83.5/83.4**	81.2/81.0	78.6/78.5	76.2/75.9	74.7/74.0	71.9/71.2	79.6/79.3
	CIF-MMIN* (Liu et al., 2024)	84.0/83.6	82.5/84.1	82.6/82.7	80.5/80.9	78.4/78.1	77.3/77.2	74.1/74.3	69.8/71.3	78.8/78.9
	SDR-GNN* (Fu et al., 2025)	86.3/86.3	85.0/85.1	81.9/81.9	80.7/80.8	77.9/78.0	76.1/76.2	72.2/72.2	71.1/71.3	78.9/79.0
	FedDISC* (Qiu et al., 2025)	86.3/86.3	85.9/85.8	83.3/83.2	81.1/81.0	**79.7**/79.6	**80.2/80.2**	77.5/77.5	75.7/75.8	81.2/81.2
	Ours	**87.0/87.0**	**86.1/86.0**	83.3/83.3	**81.3/81.4**	**79.7/79.8**	79.4/79.4	**78.4/78.4**	**76.1/76.2**	**81.4/81.4**
MOSEI	DCCAE* (Wang et al., 2015)	81.2/81.2	78.3/78.4	75.4/75.5	72.2/72.3	70.0/70.3	66.4/69.2	63.2/67.6	62.6/66.6	71.2/72.6
	GCNet* (Lian et al., 2023)	85.1/85.2	82.1/82.3	79.9/80.3	76.8/77.5	74.9/76.0	73.2/74.9	72.1/74.1	70.4/73.2	76.8/77.9
	DiCMor* (Wang et al., 2023a)	85.1/85.1	83.5/83.7	81.5/81.8	79.3/79.8	77.4/78.7	75.8/77.7	73.7/76.7	72.2/75.4	78.6/79.9
	IMDER* (Wang et al., 2023b)	85.1/85.1	84.8/84.6	82.7/82.4	81.3/80.7	79.3/78.1	79.0/77.4	78.0/75.5	77.3/74.6	80.9/79.8
	CIF-MMIN* (Liu et al., 2024)	85.8/86.2	85.4/85.5	85.0/85.3	83.1/83.8	82.7/82.5	80.4/81.1	78.5/79.2	77.3/77.4	82.3/82.6
	SDR-GNN* (Fu et al., 2025)	87.3/87.4	86.7/86.8	85.7/85.9	84.7/84.8	83.8/84.0	82.6/82.8	81.3/81.6	80.8/81.0	84.1/84.3
	FedDISC* (Qiu et al., 2025)	86.8/86.4	86.5/68.1	86.4/86.3	85.6/84.7	**84.5/84.3**	82.6/83.2	81.7/82.5	**81.5**/82.2	84.4/84.5
	Ours	**88.0/87.9**	**87.2/87.0**	**86.5/86.4**	**85.9/84.9**	84.3/**84.3**	**82.8/83.6**	**81.9/82.6**	80.9/**82.6**	**84.7/84.9**

a fixed curriculum, we strictly sample the training-time missing rate η from a uniform distribution $\mathcal{U}(0,1)$ for each batch. This dynamic strategy exposes the model to diverse sparsity levels, preventing overfitting to specific missing patterns.

4.2 Model Comparison

Tables 2 and 3 summarize performance under missing rates $\eta \in [0, 0.7]$. On **IEMOCAP**, CaM-HG demonstrates superior robustness in high-missingness scenarios. In the four-class task, it outperforms the strong graph-based competitor SDR-GNN by 1.65% at $\eta = 0.7$ (76.06% vs. 74.41%). For the six-class task, while FedDISC is competitive at $\eta = 0$, CaM-HG exhibits significantly slower decay as sparsity increases, ultimately achieving a 1.50% higher average WAF1 (60.63% vs. 59.13%). This validates that our asymmetric hypergraph preserves semantic consistency amidst turn-shifts, avoiding the sharp declines seen in baselines like CRA. On **CMU-MOSI and CMU-MOSEI**, CaM-HG sets new benchmarks with average scores of 81.4% and 84.9%. Notably, on MOSEI, our model's performance at extreme sparsity ($\eta = 0.7$, 82.6%) surpasses strong baselines operating with complete modalities (e.g., GCNet at 77.9%), confirming that the Causal-Enhanced MoE effectively decouples noise from semantics to ensure reliable reconstruction. This consistent superiority across diverse scales—from dyadic conversations to large-scale monologue opinions—further underscores the generalization capability of our confidence-aware fusion strategy. Furthermore, a detailed quantitative efficiency analysis, demonstrating CaM-HG's ultra-low latency for real-time deployment compared to generative baselines like

Table 4: Ablation study on IEMOCAP (Four-class and Six-class) with varying missing rates. The best performance is highlighted in **bold**, and the performance drop compared to the full model is shown in (green).

Dataset	Method	Missing Rate (η)							
		0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7
IEMOCAP (four-class)	**CaM-HG (Full)**	**80.73**	**79.89**	**79.74**	**79.19**	**77.96**	**77.48**	**76.45**	**76.06**
	CaM-HG$_{w/o\ \mathcal{H}}$	78.36 (-2.37)	72.80 (-7.09)	74.65 (-5.09)	74.99 (-4.20)	74.47 (-3.49)	72.98 (-4.50)	73.68 (-2.77)	73.83 (-2.23)
	CaM-HG$_{w/o\ \mathcal{I}}$	78.05 (-2.68)	74.25 (-5.64)	73.91 (-5.83)	74.53 (-4.66)	71.17 (-6.79)	71.48 (-6.00)	71.84 (-4.61)	72.79 (-3.27)
	CaM-HG$_{w/o\ \text{MoE}}$	78.08 (-2.65)	72.81 (-7.08)	72.44 (-7.30)	73.20 (-5.99)	73.93 (-4.03)	72.91 (-4.57)	72.88 (-3.57)	72.55 (-3.51)
	CaM-HG$_{w/o\ \mathcal{C}}$	77.85 (-2.88)	72.84 (-7.05)	73.68 (-6.06)	72.29 (-6.90)	72.87 (-5.09)	71.76 (-5.72)	70.37 (-6.08)	72.61 (-3.45)
	CaM-HG$_{w/o\ \mathcal{G}}$	78.02 (-2.71)	75.25 (-4.64)	72.46 (-7.28)	74.40 (-4.79)	72.61 (-5.35)	72.85 (-4.63)	73.35 (-3.10)	71.40 (-4.66)
IEMOCAP (six-class)	**CaM-HG (Full)**	**63.68**	**62.55**	**61.93**	**60.90**	**60.10**	**59.00**	**58.71**	**58.19**
	CaM-HG$_{w/o\ \mathcal{H}}$	58.69 (-4.99)	52.66 (-9.89)	53.82 (-8.11)	54.59 (-6.31)	54.51 (-5.59)	55.27 (-3.73)	54.82 (-3.89)	55.83 (-2.36)
	CaM-HG$_{w/o\ \mathcal{I}}$	60.04 (-3.64)	54.32 (-8.23)	54.49 (-7.44)	53.73 (-7.17)	52.81 (-7.29)	51.59 (-7.41)	51.28 (-7.43)	51.78 (-6.41)
	CaM-HG$_{w/o\ \text{MoE}}$	59.62 (-4.06)	55.36 (-7.19)	55.31 (-6.62)	54.50 (-6.40)	53.38 (-6.72)	53.37 (-5.63)	53.78 (-4.93)	53.10 (-5.09)
	CaM-HG$_{w/o\ \mathcal{C}}$	59.74 (-3.94)	52.63 (-9.92)	51.76 (-10.17)	53.37 (-7.53)	52.30 (-7.80)	50.96 (-8.04)	51.67 (-7.04)	52.88 (-5.31)
	CaM-HG$_{w/o\ \mathcal{G}}$	58.61 (-5.07)	52.69 (-9.86)	52.27 (-9.66)	54.07 (-6.83)	53.13 (-6.97)	52.73 (-6.27)	53.57 (-5.14)	53.43 (-4.76)

Table 5: Prediction results on incomplete conversational data from the IEMOCAP dataset. The conversation layout aligns speaker utterances to the left (Speaker A) and right (Speaker B) to visualize the interaction flow. Red circles in the Mods columns indicate missing modalities.

Trn	Conversation		Mods			GT	Predictions				
	Speaker A	*Speaker B*	A	V	L	Emo	CRA	MMIN	GCNet	SDR	CaM-HG
1	Excuse me? I've been waiting here for over 45 minutes.		•	•	•	Fru	Fru	Fru	Fru	Fru	Fru
2		Yeah, well, sometimes it takes time to unload the plane.	○	○	•	Neu	Fru	Neu	Neu	Neu	Neu
3	Everyone else has left! It's just me standing here!		•	○	•	Fru	Fru	Fru	Neu	Fru	Fru
4		Ma'am, I don't know what to tell you.	○	○	○	Fru	Neu	Neu	Fru	Fru	Fru
5	Don't tell me you lost it. My medication is in that bag!		•	•	○	Ang	Fru	Ang	Ang	Ang	Ang
6		Well, did you check with the lost and found?	○	•	•	Fru	Neu	Fru	Fru	Fru	Fru

FedDISC, is provided in Appendix C.5.

4.3 Ablation Study

To verify the contribution of each component, we evaluate five variants on IEMOCAP: **w/o** $\mathcal{H}$ (replaces the hypergraph with a standard Transformer encoder to test topological effectiveness), w/o $\mathcal{I}$ (no imputation), w/o MoE (single MLP), w/o $\mathcal{C}$ (no causal factor), and w/o $\mathcal{G}$ (no gating). The results in Table 4 confirm the indispensability of each module. Quantitative results demonstrate significant contributions from all components. Specifically, the **Causal-Enhanced MoE** is pivotal for restoration; removing the Causal Factor Extractor (w/o $\mathcal{C}$) causes a sharp drop (e.g., 59.00% to 50.96% at $\eta = 0.5$), indicating that decoupling causal semantics from noise is vital. The **Asymmetric Hypergraph** ensures semantic consistency, evidenced by the performance decline from 63.68% to 58.69% at $\eta = 0.0$ when replaced by a Transformer (w/o $\mathcal{H}$).This confirms our high-order topology captures complex patterns missed by pairwise attention. Moreover, removing **Imputation** causes sharp drops (51.78% at $\eta = 0.7$), verifying robustness under extreme sparsity.

4.4 Case Study

Table 5 analyzes a representative dialogue to assess robustness. In Turn 6, where the acoustic modality is missing, the deceptive textual politeness leads shallow generative methods (e.g., CRA) to incorrectly predict *Neutral*. In contrast, CaM-HG successfully recovers the latent *Frustration* by leveraging the causal trigger from Turn 5. This validates that causal modeling is essential for bridging semantic ruptures in incomplete data.

5 Conclusion

In this work, we redefine modality missingness in MERC as a rupture of dialogue causal chains that severs reasoning pathways, rather than mere data sparsity. We propose CaM-HG, a novel framework adhering to a "Restore-then-Mine" paradigm. By integrating a Causal-Enhanced Mixture-of-Experts

to regenerate missing modalities as valid semantic anchors conditioned on historical context, and employing an Asymmetric Causal Dynamic Hypergraph to capture high-order combinatorial synergies, our approach effectively bridges semantic gaps while strictly preserving temporal logic. Extensive experiments on IEMOCAP, CMU-MOSI, and CMU-MOSEI confirm CaM-HG's state-of-the-art robustness, validating the necessity of causal consistency for incomplete multimodal data.

Limitations

We identify two aspects of our framework that present avenues for future research. First, regarding extreme data sparsity, our Causal-Enhanced MoE is designed to regenerate features by leveraging historical triggers. While CaM-HG achieves state-of-the-art robustness, recovering fine-grained semantics under conditions of *continuous* extreme missingness (e.g., prolonged sensor failure) remains an information-theoretic challenge inherent to the MERC task. Future work could address this by integrating external commonsense knowledge to hallucinate plausible contexts in the absence of history. Second, regarding the scope of causality, our framework prioritizes temporal causality and history-dependent generation to ensure stability in sequence modeling. While this effectively addresses the "causal rupture" problem, it differs from structural causal discovery methods that aim to disentangle complex confounders in static graphs. Extending our temporal framework to explicitly incorporate structural causal discovery algorithms could further enhance interpretability.

Acknowledgments

This work was supported by the Sichuan Science and Technology Program under Grant Number 2024ZDZX0011. The authors would also like to thank the anonymous reviewers and the meta-reviewer for their constructive feedback, which significantly improved the quality of this paper.

References

Wei Ai, Fuchen Zhang, Yuntao Shou, Tao Meng, Haowen Chen, and Keqin Li. 2025. Revisiting multimodal emotion recognition in conversation from the perspective of graph spectrum. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 11418–11426.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Xuping Chen and Wuzhen Shi. 2025. Dynamic interactive bimodal hypergraph networks for emotion recognition in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1256–1264.

Joshua Fixelle. 2025. Hypergraph vision transformers: Images are more than nodes, more than edges. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9751–9761.

Fangze Fu, Wei Ai, Fan Yang, Yuntao Shou, Tao Meng, and Keqin Li. 2025. Sdr-gnn: spectral domain reconstruction graph neural network for incomplete multimodal learning in conversational emotion recognition. *Knowledge-Based Systems*, 309:112825.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Ziyi Li, Wei-Long Zheng, and Bao-Liang Lu. 2025. Multimodal emotion recognition with missing modality via a unified multi-task pre-training framework. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 5717–5725.

Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. 2023. Gcnet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on pattern analysis and machine intelligence*, 45(7):8419–8432.

Rui Liu, Haolin Zuo, Zheng Lian, Björn W Schuller, and Haizhou Li. 2024. Contrastive learning based modality-invariant feature acquisition for robust multimodal emotion recognition with missing modalities. *IEEE Transactions on Affective Computing*, 15(4):1856–1873.

Nannan Lu, Zhiyuan Han, and Zhen Tan. 2024. A hypergraph based contextual relationship modeling method for multimodal emotion recognition in conversation. *IEEE Transactions on Multimedia*.

Xiaosen Lyu, Jiayu Xiong, Yuren Chen, Wanlong Wang, Xiaoqing Dai, and Jing Wang. 2025. Cross-space synergy: A unified framework for multimodal emotion recognition in conversation. *arXiv preprint arXiv:2512.03521*.

Hui Ma, Jian Wang, Hongfei Lin, Bo Zhang, Yijia Zhang, and Bo Xu. 2023. A transformer-based model with self-distillation for multimodal emotion recognition in conversations. *IEEE Transactions on Multimedia*, 26:776–788.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, and 1 others. 2021. Recognizing emotion cause in conversations. *Cognitive Computation*, 13(5):1317–1332.

Xihang Qiu, Jiarong Cheng, Yuhao Fang, Wanpeng Zhang, Yao Lu, Ye Zhang, and Chun Li. 2025. Federated dialogue-semantic diffusion for emotion recognition under incomplete modalities. *arXiv preprint arXiv:2511.00344*.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Jiandong Shi, Ming Li, Lu Bai, Feilong Cao, Ke Lu, and Jiye Liang. 2025. Match: Modality-calibrated hypergraph fusion network for conversational emotion recognition. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 6164–6172.

Yuntao Shou, Jun Yao, Tao Meng, Wei Ai, Cen Chen, and Keqin Li. 2025. Gsdnet: Revisiting incomplete multimodal-diffusion from graph spectrum perspective for conversation emotion recognition. *arXiv preprint arXiv:2506.12325*.

Darshan Solanki, Hsia-Ming Hsu, Olivia Zhao, Renyue Zhang, Weihao Bi, and Raman Kannan. 2020. The way we think about ourselves. In *International Conference on Human-Computer Interaction*, pages 276–285. Springer.

Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. 2017. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1405–1414.

Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. 2015. On deep multi-view representation learning. In *International conference on machine learning*, pages 1083–1092. PMLR.

Yuanzhi Wang, Zhen Cui, and Yong Li. 2023a. Distribution-consistent modal recovering for incomplete multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22025–22034.

Yuanzhi Wang, Yong Li, and Zhen Cui. 2023b. Incomplete multimodality-diffused emotion recognition. *Advances in Neural Information Processing Systems*, 36:17117–17128.

Yusong Wang, Xuanye Fang, Huifeng Yin, Dongyuan Li, Guoqi Li, Qi Xu, Yi Xu, Shuai Zhong, and Mingkun Xu. 2025. Big-fusion: Brain-inspired global-local context fusion framework for multimodal emotion recognition in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1574–1582.

Wen Yin, Yong Wang, Guiduo Duan, Dongyang Zhang, Xin Hu, Yuan-Fang Li, and Tao He. 2025. Knowledge-aligned counterfactual-enhancement diffusion perception for unsupervised cross-domain visual emotion recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3888–3898.

Wen Yin, Siyu Zhan, Cencen Liu, Xin Hu, Guiduo Duan, Xiurui Xie, Yuan-Fang Li, and Tao He. 2026. Tical: Typicality-based consistency-aware learning for multimodal emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 17948–17956.

Jiaxuan You, Xiaobai Ma, Yi Ding, Mykel J Kochenderfer, and Jure Leskovec. 2020. Handling missing data with graph representation learning. *Advances in Neural Information Processing Systems*, 33:19075–19087.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. 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)*, pages 2236–2246.

Tao Zhang and Zhenhua Tan. 2025. Ecerc: evidence-cause attention network for multi-modal emotion recognition in conversation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2064–2077.

Jinming Zhao, Ruichen Li, and Qin Jin. 2021a. Missing modality imagination network for emotion recognition with uncertain missing modalities. 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)*, pages 2608–2618.

Zengqun Zhao, Qingshan Liu, and Shanmin Wang. 2021b. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Transactions on Image Processing*, 30:6544–6556.

Xianxun Zhu, Yaoyang Wang, Erik Cambria, Imad Rida, José Santamaría López, Lin Cui, and Rui Wang. 2025. Rmer-dt: Robust multimodal emotion recognition in conversational contexts based on diffusion and transformers. *Information Fusion*, page 103268.

A Implementation Details

To facilitate reproducibility and provide deeper insights into the training stability of CaM-HG, we detail the network architecture, curriculum learning strategy, and critical initialization techniques. We also provide a theoretical justification for our hyperparameter choices, specifically regarding the Mixture-of-Experts (MoE) configuration.

A.1 Detailed Architecture Configuration

All experiments were conducted on a single NVIDIA RTX 4090 GPU using PyTorch 2.3.0.

Feature Projection. We project raw features from DeBERTa-large ($d_l = 1024$), Wav2Vec 2.0 ($d_a = 1024$), and MANet ($d_v = 512$) into a unified latent space with dimension D.

- **IEMOCAP:** We set $D = 512$ and Dropout $p \in [0.45, 0.5]$ to accommodate the rich semantic shifts in 4-way and 6-way emotion classification.
- **CMU-MOSI/MOSEI:** We set $D = 256$ with a lower dropout ($p = 0.2$) as these datasets focus on sentiment polarity (regression), requiring less capacity than fine-grained emotion recognition.

A.2 Expert Configuration Analysis: Why $N = 4$?

A critical design choice in our Causal-Enhanced MoE module is the number of experts (N) and the sparsity factor (K). We empirically and theoretically determined that $N = 4$ with Top-$K = 2$ is optimal for this task. As illustrated in Figure 3, the performance on the IEMOCAP (4-class) validation set follows an inverted U-shaped trend, peaking at $N = 4$ and degrading as N increases further.

The rationale behind this configuration is threefold:

- **Representation Diversity ($N > 2$):** Emotional expressions in conversation are highly complex and multimodal (e.g., "Frustration" vs. "Anger" vs. "Excitement"). Setting $N = 2$ would limit the latent space to a binary manifold (e.g., merely capturing Positive vs. Negative), which is insufficient for resolving the "one-to-many" mapping inherent in emotion restoration.

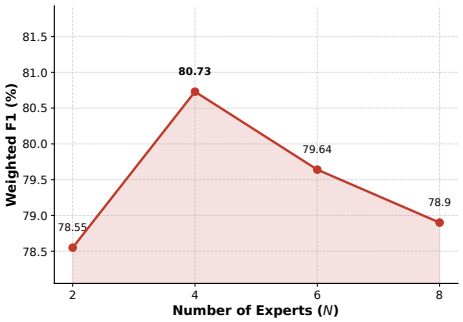

Figure 3: Parameter sensitivity analysis on the number of experts (N) on the IEMOCAP(4-class) dataset. The curve demonstrates a clear performance peak at $N = 4$, validating our choice of balancing representation diversity and parameter efficiency.

- **Parameter Efficiency ($N < 8$):** While increasing experts theoretically boosts capacity, datasets like IEMOCAP (∼7k utterances) are data-scarce compared to large-scale LLM training. Setting $N = 8$ or higher leads to severe overfitting and the "Expert Collapse" problem, where only 1-2 experts remain active while others receive zero gradients.
- **The Sweet Spot ($N = 4, K = 2$):** $N = 4$ allows the model to learn distinct "emotional prototypes" (e.g., approximating the 4 quadrants of the Valence-Arousal space). With Top-$K = 2$, we enable $\binom{4}{2} = 6$ unique expert combinations, providing rich combinatorial diversity to reconstruct subtle emotional nuances without inflating the parameter count.

A.3 Training Strategy and Hyperparameter Configuration

We employ a three-phase curriculum. Notably, in **Phase 3 (Joint Fine-tuning)**, we adopt a **Differential Learning Rate (Diff-LR)** strategy to stabilize the adversarial-like interplay between the generator and classifier. As detailed in Table 6, our hyperparameter configuration reveals three critical strategic insights:

Stabilizing Generator-Gate Dynamics (Diff-LR): The *Confidence Gate* generally requires a significantly higher learning rate ($\sim 10^{-3}$) compared to the *Generator* ($\sim 10^{-5}$ to 10^{-4}). This aggressive update allows the gate to rapidly learn the switching policy—trusting the generator only when necessary—while the generator fine-tunes its output at a more conservative pace to avoid disrupting the pre-trained manifold.

Algorithm 1 Training Procedure for CaM-HG

Input: Dataset $\mathcal{D}$, Max Epochs K_1, K_2, K_3, Diff-LRs $\eta_{base}, \eta_{gen}, \eta_{gate}$.
Initialize: Projectors $\{\mathbf{W}_m\}$, MoE Experts $\{E_k\}$, Gating Network $\mathbf{W}_g$, Hypergraph ACDH, Classifier f_θ.
Define: $\mathcal{L}_{task}$ (Cross-Entropy) and $\mathcal{L}_{rec}$ (MSE).
 # Phase 1: Discriminative Warm-up (Topology Learning)
for epoch $= 1$ to K_1 **do**
  for batch $\{\mathbf{x}_t\}_{t=1}^{L}$ in $\mathcal{D}$ **do**
    Generate availability $\boldsymbol{\delta}_t$ via Bernoulli sampling & Text Blindness ($p = 0.25$).
    Obtain masked observations $\tilde{\mathbf{x}}_t^m$ and project to latent $\tilde{\mathbf{h}}_t^m$ via Eq. (2).
    Forward ACDH: Pass $\tilde{\mathbf{h}}_t^m$ (treating missing as zero placeholders) into Hypergraph.
    Compute $\mathcal{L}_{task}$ and **Update** $\{\mathbf{W}_m\}$, ACDH, f_θ.
  end for
end for
 # Phase 2: Generative Pre-training (Causal Restoration)
Freeze Projectors $\{\mathbf{W}_m\}$. **Init** MoE Generator.
for epoch $= 1$ to K_2 **do**
  for batch $\{\mathbf{x}_t\}_{t=1}^{L}$ in $\mathcal{D}$ **do**
    Compute Target $\mathbf{h}_t^{m,\mathrm{o}}$ (Unmasked) and Input $\tilde{\mathbf{h}}_t^m$ (Masked).
    Context Extraction: Fuse history $\mathbf{u}_t$ via Eq. (3) and extract $\mathbf{z}_{c,t} = \text{Transformer}(\ldots, \mathbf{M}_{causal})$.
    Query Construction: $\mathbf{q}_t^m = \mathbf{z}_{c,t} + \mathbf{P}_{mod}(m)$.
    Generate: $\hat{\mathbf{h}}_t^m = \sum_{k \in \mathcal{I}_t^m} \alpha_{t,k}^m \cdot E_k(\mathbf{q}_t^m) + \mathbf{z}_{c,t}$ via Eq. (6).
    Compute $\mathcal{L}_{rec} = \sum_{t,m} \|\hat{\mathbf{h}}_t^m - \mathbf{h}_t^{m,\mathrm{o}}\|^2$ and **Update** Generator.
  end for
end for
 # Phase 3: Joint Fine-tuning (Restore-then-Mine)
Unfreeze All. **Init** Confidence Gate bias $\mathbf{b}_{gate} \leftarrow -2.0$.
for epoch $= 1$ to K_3 **do**
  for batch $\{\mathbf{x}_t\}_{t=1}^{L}$ in $\mathcal{D}$ **do**
    Restore: Generate $\hat{\mathbf{h}}_t^m$ via MoE (as in Phase 2).
    Fuse: Compute gate $\boldsymbol{\sigma}_t$ and fuse $\mathbf{v}_t^m = \boldsymbol{\sigma}_t \odot \hat{\mathbf{h}}_t^m + (\mathbf{1} - \boldsymbol{\sigma}_t) \odot \tilde{\mathbf{h}}_t^m$ via Eq. (7).
    ACDH Mining: Update hyperedges $\mathcal{E}^{(l)}$ and nodes $\mathcal{V}^{(l)}$ with $\mathbf{M}_{block}$ via Eq. (10).
    Predict: $\hat{y}_t = f_\theta(\mathcal{V}^{(L)})$.
    Loss: Compute $\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda(1 + \eta_{batch})\mathcal{L}_{rec}$ via Eq. (11).
    Backpropagate and Update with Diff-LRs ($\eta_{gate}, \eta_{gen}, \eta_{base}$).
  end for
end for

Task-Specific Reconstruction Weights (λ_{rec}): We observe a distinct divergence in the optimal reconstruction weight λ_{rec} across datasets. For fine-grained emotion recognition (IEMOCAP), a lower weight ($\lambda_{rec} \approx 0.108$) is preferred to prevent pixel-level reconstruction losses from overriding subtle emotional semantics. In contrast, for sentiment analysis (CMU-MOSI/MOSEI), the model tolerates a higher weight ($\lambda_{rec} \approx 0.4$) due to the coarser granularity of sentiment labels, allowing for stronger regularization via reconstruction.

Adaptive Regularization for Data Scarcity: To mitigate overfitting on the smaller IEMOCAP dataset (~7k utterances), we employ stricter regularization techniques, including **R-Drop** ($\alpha = 1.0$) and **Label Smoothing** (0.05), coupled with a *ReduceLROnPlateau* scheduler. Conversely, for the larger-scale CMU-MOSI/MOSEI datasets, we find that standard *Cosine Annealing* without additional regularization (R-Drop $\alpha = 0.0$) yields the best generalization.

Critical Initialization. Consistent with our "Restore-then-Mine" philosophy, we initialize the bias of the Confidence Gating layer to **-2.0**. This ensures the model starts by trusting the raw (fused) features ($\sigma \approx 0.12$), preventing the "Cold-Start" noise of the generator from degrading the classifier's performance in early epochs.

B Training Algorithm and Loss Flow

To further elucidate the multi-stage optimization process described in Section A, we provide the detailed pseudo-code of the CaM-HG training procedure in Algorithm 1. This algorithm explicitly illustrates the curriculum learning flow, the freezing/unfreezing of modules, and the dynamic application of the differential learning rates and reconstruction weights.

C Supplementary Experimental Analysis

In this section, we provide further visualizations and detailed analyses to substantiate the robustness

Table 6: Best Parameter Configurations. Note the specific settings for each dataset and the differential learning rates applied in Phase 3.

Hyperparameter	**IEMOCAP (6-way)**	**IEMOCAP (4-way)**	**CMU-MOSI**	**CMU-MOSEI**
Model Structure				
Hidden Dim (D)	512	512	256	256
Dropout	0.45	0.50	0.40	0.40
Num. Experts (N)	4	4	4	4
Top-K	2	2	2	2
Phase 1 (Warm-up)				
Learning Rate ($P1\_LR$)	1.78×10^{-4}	1.85×10^{-4}	2.83×10^{-4}	1.31×10^{-4}
Weight Decay	1×10^{-3}	1×10^{-3}	1×10^{-4}	1×10^{-4}
Phase 3 (Fine-tuning)				
Gate LR ($P3\_Gate\_LR$)	3.11×10^{-3}	2.64×10^{-3}	1.00×10^{-3}	1.00×10^{-3}
Gen LR ($P3\_Gen\_LR$)	5.78×10^{-5}	6.41×10^{-5}	5.05×10^{-5}	1.91×10^{-4}
Rec. Weight (λ_{rec})	0.106	0.108	0.407	0.421
Regularization				
R-Drop Alpha	1.0	1.0	0.0	0.0
Label Smoothing	0.05	0.0	0.0	0.0
Scheduler	Plateau	Plateau	Cosine	Cosine

and effectiveness of CaM-HG under various experimental settings.

C.1 Robustness Analysis Across Modality Subsets

To rigorously evaluate the model's resilience against unpredictable sensor failures, we tested CaM-HG on all possible subsets of available modalities (i.e., $\{l\}, \{v\}, \{a\}, \{l, v\}, \dots$). The detailed comparison with state-of-the-art methods on IEMOCAP-4 and IEMOCAP-6 is presented in Table 7.

Dominance in IEMOCAP-4. On the four-class emotion recognition task, CaM-HG demonstrates absolute dominance, achieving the best performance (bold) across all 7 modality combinations. Notably, in the extreme case where only the visual modality is available ($\{v\}$), our model achieves an Accuracy/F1 of **63.8/63.5**, outperforming the strongest baseline (FedDISC) by a significant margin. This suggests that our *Causal-Enhanced MoE* effectively leverages the visual context to hallucinate plausible acoustic or linguistic semantics, compensating for the missing signals.

Superiority in Partial Modalities (IEMOCAP-6). For the more challenging six-class task, while FedDISC shows strong performance on complete data ($\{l, v, a\}$), CaM-HG excels in most incomplete scenarios, particularly in uni-modal and text-audio settings, which is the core focus of this work. Specifically:

- **Single Modality Robustness:** In the uni-modal settings ($\{l\}$, $\{v\}$, $\{a\}$), our model consistently ranks first or second. For instance, with only audio input ($\{a\}$), CaM-HG achieves **61.2/60.8**, surpassing both SDR-GNN and GCNet.
- **Key Modality Synergy:** In the crucial $\{l, a\}$ combination (Text + Audio), which is often considered the most informative pair for emotion recognition, CaM-HG achieves a new state-of-the-art result of **60.8/60.5**.

Overall, these results confirm that CaM-HG does not merely rely on the fusion of all modalities but constructs a robust internal representation that remains discriminative even when the input topology is severely ruptured.

C.2 Robustness Analysis via Confusion Matrices

To provide a granular view of the model's predictive behavior under varying degrees of sparsity, we visualize the confusion matrices for both IEMOCAP-4 and IEMOCAP-6 tasks in Figure 4. The evaluation covers a broad spectrum of missing rates $\eta \in \{0.1, 0.3, 0.5, 0.7\}$, ranging from mi-

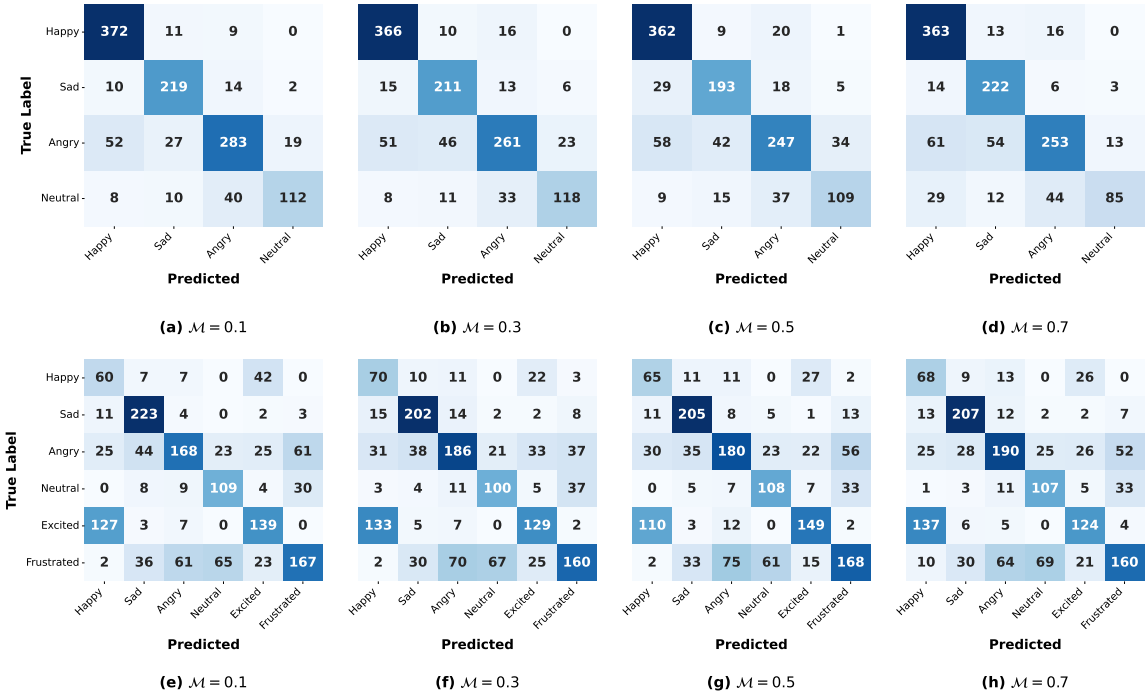

Figure 4: **Confusion Matrices on IEMOCAP Test Set across Varying Missing Rates.** Top row: IEMOCAP-4 (4-class); Bottom row: IEMOCAP-6 (6-class). The dominance of diagonal elements even at $\eta = 0.7$ (Columns d, h) demonstrates the model's capability to maintain semantic consistency despite severe modality loss.

nor noise to extreme structural rupture. **Diagonal Stability and Resistance to Collapse.** The most striking observation is the remarkable stability of the diagonal elements (True Positives). As the missing rate η increases from 0.1 to 0.7, we do not observe the catastrophic performance collapse typical of baseline methods. For instance, in the 4-class task (Top Row), the correct predictions for the *Happy* class decrease only marginally from 372 ($\eta = 0.1$) to 363 ($\eta = 0.7$). Similarly, the *Angry* class maintains high recall (283 $\rightarrow$ 253). This evidence strongly supports our hypothesis that the *Causal-Enhanced MoE* successfully hallucinates semantically consistent features rather than reverting to uniform noise, effectively "locking" the model's attention on the correct emotional manifold even when 70% of the input is masked.

Semantic Error Patterns. Instead of degenerating into random guessing, the error patterns (off-diagonal elements) reflect genuine semantic ambiguities inherent in human communication. In IEMOCAP-4, the *Neutral* class is the primary source of confusion, often misclassified as *Angry* or *Sad*, which is expected as neutral utterances lack strong activation cues. In IEMOCAP-6 (Bottom Row), we observe a persistent confusion between *Excited* and *Happy* (e.g., in Figure 4(h), 137 *Excited* samples are predicted as *Happy*). Crucially, this error pattern is structurally consistent with the full-modality performance. Since both emotions share high valence and high arousal, the fact that CaM-HG preserves this specific confusion pattern—rather than distributing errors randomly—indicates that it has learned the true underlying topology of the emotion space.

C.3 Impact of Hypergraph Depth (L)

We investigate the impact of the number of layers L in the Asymmetric Causal Dynamic Hypergraph, varying L from 1 to 6. As shown in Figure 5, the model achieves optimal performance on IEMOCAP(4-class) at $L = 2$.

- **Underfitting ($L = 1$):** With a single layer, the hypergraph can only capture immediate correlations, failing to model the complex, multi-hop causal dependencies inherent in long dialogues.
- **Optimal Range ($L = 2$):** At $L = 2$, the model effectively aggregates high-order interactions without introducing excessive noise, representing the "sweet spot" for reasoning depth.
- **Over-smoothing ($L \geq 4$):** Consistent with

Table 7: Comparison results on IEMOCAP4 and IEMOCAP6 datasets across different available modalities. The best result in each row is highlighted in **bold**, and the second best is underlined. Results of baseline methods marked with * are cited from previous work (Qiu et al., 2025).

Dataset	Available	Baseline*	**Ours**	FedDISC (P)*	GCNet*	CIF-MMIN*	SDR-GNN*	IMDer*	DiCMor*
IEMOCAP4	$\{l\}$	59.6/59.4	**70.5/70.8**	68.0/68.2	66.7/65.9	58.0/59.3	66.2/66.2	60.4/60.4	28.6/28.5
	$\{v\}$	58.1/57.3	**63.8/63.5**	62.5/60.3	55.7/55.7	53.3/51.3	56.3/55.6	56.2/54.7	21.9/15.6
	$\{a\}$	53.0/50.4	**64.5/64.2**	61.2/60.6	56.8/56.3	56.3/58.4	57.8/57.3	50.8/50.7	34.0/27.2
	$\{l, v\}$	67.2/67.3	**77.2/77.5**	75.8/75.8	64.9/65.0	73.2/74.0	68.1/68.0	67.3/67.3	37.0/25.6
	$\{l, a\}$	66.5/65.8	**79.1/79.3**	78.0/78.1	68.7/68.3	74.3/75.6	73.0/72.1	65.9/66.2	32.5/32.3
	$\{v, a\}$	60.6/60.3	**70.8/71.2**	68.7/68.0	60.4/60.8	65.7/66.9	60.2/60.3	65.0/64.8	47.1/37.8
	$\{l, v, a\}$	78.6/78.4	**80.6/80.7**	78.6/78.4	78.4/78.3	78.3/78.5	78.5/78.1	78.1/78.3	78.2/78.3
IEMOCAP6	$\{l\}$	46.4/45.8	**59.2/59.1**	56.8/56.5	50.8/50.0	54.7/54.3	58.8/58.9	44.6/44.9	34.5/33.8
	$\{v\}$	36.2/36.4	**56.2/55.9**	56.0/55.5	55.7/55.7	39.7/35.2	41.8/41.0	39.3/35.1	33.6/32.4
	$\{a\}$	36.8/29.9	**61.2/60.8**	60.6/59.9	56.8/56.3	56.3/58.4	51.6/50.7	39.2/37.5	38.6/38.7
	$\{l, v\}$	49.4/49.9	**61.5/61.3**	57.4/57.6	49.3/47.8	51.3/52.1	60.6/60.3	49.6/49.0	34.4/34.8
	$\{l, a\}$	51.3/50.6	**60.8/60.5**	59.4/59.3	51.9/51.3	53.8/50.1	60.3/60.4	51.5/50.4	37.1/37.0
	$\{v, a\}$	41.1/36.3	60.4/60.6	**66.4/66.3**	44.0/43.4	56.5/56.9	60.0/50.7	43.8/41.2	36.5/35.9
	$\{l, v, a\}$	64.3/64.7	63.5/63.7	**64.3/64.7**	58.6/59.1	61.3/60.2	61.3/61.2	58.7/58.4	55.9/56.2

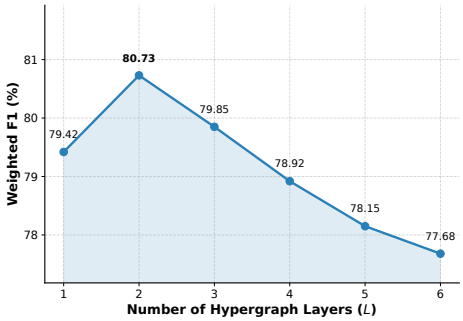

Figure 5: Sensitivity analysis of the number of hypergraph layers (L) on IEMOCAP(4-class). The performance drops when $L > 3$, indicating the issue of over-smoothing in high-order message passing.

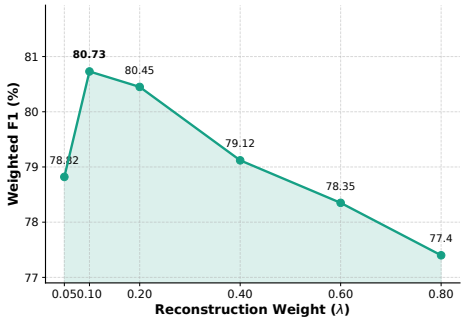

Figure 6: Impact of the reconstruction loss weight (λ) on model performance on the IEMOCAP (4-class) dataset. A moderate weight ($\lambda \approx 0.1$) yields the best results, while excessive weights degrade the primary classification task.

findings in GNN literature, stacking too many layers leads to the over-smoothing problem, where node representations become indistinguishable, causing a steady decline in F1 score.

C.4 Sensitivity to Reconstruction Weight (λ)

The reconstruction loss weight λ balances the primary classification task and the auxiliary imputation task. We evaluate $\lambda \in [0.05, 0.8]$. The results on the IEMOCAP(4-class) dataset are visualized in Figure 6.

- **Insufficient Supervision ($\lambda < 0.1$):** When λ is too small, the generator receives weak gradient signals, leading to poor reconstruction of missing modalities and subsequently noisy inputs for the classifier.
- **Task Interference ($\lambda > 0.4$):** As λ increases beyond 0.4, the reconstruction objective begins to dominate the optimization landscape. The model allocates excessive capacity to pixel-level or feature-level restoration at the expense of learning discriminative emotion semantics, leading to a performance drop.
- **Optimal Balance:** We observe that $\lambda \in [0.1, 0.2]$ provides the best stability, ensuring that the restored features are semantically consistent without distracting from the main MERC task.

C.5 Computational Efficiency Analysis

To comprehensively evaluate the architectural efficiency of our proposed framework, we compare the inference latency and throughput of CaM-HG against the recent generative baseline, FedDISC (Qiu et al., 2025). It is worth noting that while FedDISC relies on a heavy

pipeline—including R-GCNs, extensive attention mechanisms, and an iterative diffusion module—CaM-HG is highly lightweight. Given that our pre-trained backbones function purely as offline feature extractors with frozen weights, our trainable core consists solely of simple projectors, a sparse MoE, and a shallow hypergraph.

As summarized in Table 8, based on the standard IEMOCAP test set scale (1,312 utterances), CaM-HG demonstrates a stark efficiency advantage. Utilizing a single consumer-grade RTX 4090 GPU and a 1-pass MoE routing strategy, CaM-HG achieves a massive throughput of 2613.83 utterances per second with an ultra-low latency of 0.38 ms per utterance. This is over 11 times faster than the accelerated DDIM version of FedDISC (50 iterative steps) running on dual enterprise-grade NVIDIA L40S GPUs. These results explicitly validate that our sparse routing (activating 2 of 4 experts) and 2-layer hypergraph incur negligible computational overhead, making CaM-HG exceptionally well-suited for real-time deployment scenarios.

Table 8: Inference efficiency comparison on the IEMOCAP test set.

Model (Generation Mechanism)	Hardware Setup	Est. Throughput	Est. Latency
FedDISC (DDPM, 1000 steps)	2× NVIDIA L40S	10.27 utt/s	97.40 ms
FedDISC (DDIM, 50 steps)	2× NVIDIA L40S	230.18 utt/s	4.34 ms
CaM-HG (Ours, 1-pass MoE)	**1× RTX 4090**	**2613.83 utt/s**	**0.38 ms**