Dynamic Modeling of Utterance-Level and Token-Level Modality Imbalance for Multimodal Sentiment Analysis

Anonymous ACL submission

Abstract

Multimodal Sentiment Analysis (MSA) aims to infer human affect from linguistic, acoustic, and visual information. Existing methods often alleviate modality imbalance by adopting text-dominant fusion, but this assumption may be suboptimal when acoustic or visual cues provide more reliable affective evidence. Moreover, modality dominance is not always consistent across different levels: the utterance-level dominant modality may differ from token-level local affective evidence. To address this issue, we propose HD-MoDE, a Hierarchical Dynamic Mixture-of-Dominant-Experts framework for modeling Global-Local Modality Imbalance in MSA. HD-MoDE contains an utterance-level MoDE that estimates the dominant modality of the whole utterance and provides a macro affective prior, and a token-level MoDE that dynamically selects modality-dominant experts to capture fine-grained local modality variations. Experiments on CMU-MOSI and CMU-MOSEI show that HD-MoDE achieves strong performance, especially on fine-grained sentiment recognition. The anonymized code is available at `https://anonymous.4open.science/r/anonymous-MSA-code`.

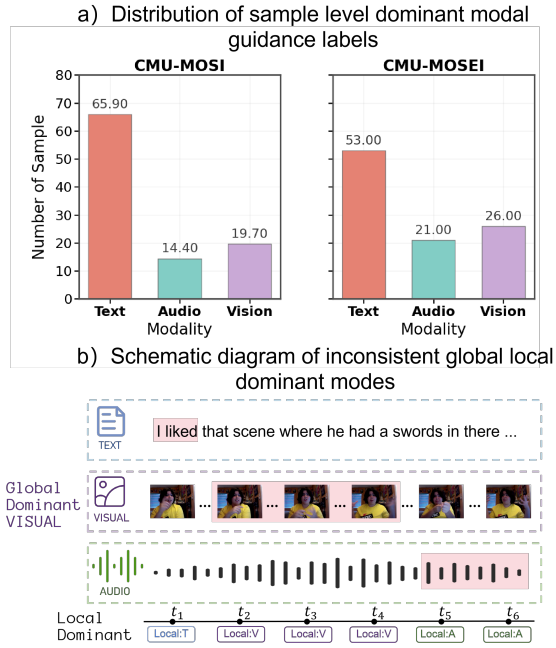

Figure 1: a) the distribution of the utterance-level dominant modal, showing the dynamics of dominant modality. b) the inconsistency between utterance-level and token-level dominant modality patterns in multimodal samples, showing significant temporal heterogeneity and global-local modality imbalance.

1 Introduction

Multimodal Sentiment Analysis (MSA) aims to infer human affect by integrating linguistic, acoustic, and visual information (Zadeh et al., 2017). An effective MSA requires not only extracting information from multiple modalities, but also exploiting their complementarity under diverse emotional contexts (Rahman et al., 2020). However, this multimodal complementarity is often weakened by modality imbalance during joint optimization (Wang et al., 2020; Peng et al., 2022). In practice, some modalities may provide denser semantic information, more stable supervision signals, or easier optimization paths, causing the model to rely disproportionately on them while under-utilizing other informative modalities. Such an imbalance undermines the original goal of multimodal learning: although the model receives text, audio, and visual inputs, its final decision may still be dominated by a limited subset of modalities.

To alleviate modality imbalance, many methods take textual representations as the main anchor (Rahman et al., 2020; Hou et al., 2024) and then inject, align, or attend to acoustic and visual features around the text stream since text usually provides explicit semantic structure and strong task-related signals. For example, DMMF (Wu et al., 2024) propose a text-dominant multistage modality fusion framework, which first enhances textual rep-

resentations with visual and acoustic cues and then strengthens cross-modal interactions through self-attention and cross-attention. TIEMFF (Liu et al., 2024) further build a text-centered multimodal fusion framework, where cross-modal attention and sentiment-aware textual enhancement are used to refine multimodal emotional representations. Nevertheless, although these text-dominant methods have made progress in mitigating modality imbalance, they still face two key challenges.

1) Limitations of the text-dominance assumption. In fact, the dominant modality is not uniformly fixed across samples. As shown in Fig. 1 .a), although text accounts for a large proportion of dominant-modality proposal labels, audio and vision also occupy non-negligible proportions. This suggests that the most reliable affective source can vary from sample to sample. For utterances involving facial expression, acoustic or visual cues may provide stronger information than the text in some contexts. Therefore, a fixed text-dominant fusion paradigm may suppress useful non-verbal information and fail to fully address modality imbalance (Feng et al., 2024).

2) Global–local affective evidence imbalance. Real-world sentiment expression is hierarchical and dynamic. At the global-level, an utterance may exhibit an overall modality tendency; for example, its sentiment may be mainly determined by textual semantics, vocal tone, or visual expressions. At the local-level, however, the source of affective evidence can change across tokens or temporal positions. As shown in Fig. 1 .b), the information dominating sentiment may represent different modalities at different times. This indicates that modality imbalance should not be modeled only at the sample level; instead, it requires a finer global-local perspective.

Motivated by these observations, we propose a Hierarchical Dynamic Mixture-of-Dominant-Experts (HD-MoDE) to address the above text-dominant-based limitation and Global-Local Modality Imbalance on multimodal sentiment analysis. HD-MoDE consists of an utterance-level MoDE and a token-level MoDE, aiming to decompose modality dominance into global-level and local-level affective evidence selection. Specifically, an utterance-level router estimates the dominant modality of the whole utterance to provide a macro affective prior, while a token-level router dynamically selects modality-dominant experts for individual tokens to capture fine-grained shifts in affective evidence. Through the collaborative modeling of the utterance-level MoDE and token-level MoDE, our HD-MoDE can simultaneously characterize utterance-level dominant modalities and token-level local modality variations, thereby making fuller use of key affective cues from non-textual modalities such as audio and vision.

Our contributions are summarized as follows:

- We revisit modality imbalance in multimodal sentiment analysis from a global-local perspective, which characterizes the discrepancy between utterance-level modality preference and token-level affective evidence.
- We present a hierarchical dynamic mixture-of-dominant-experts framework, where utterance-level and token-level MoDE are used to adaptively model global-level modality tendency and fine-grained local-level evidence selection.
- We provide comprehensive experiments on CMU-MOSI and CMU-MOSEI dataset, showing that HD-MoDE offers a more flexible way to handle dynamic modality imbalance on the global-local level.

2 Related Work

2.1 Modality Imbalance

Modality imbalance is a common challenge in multimodal sentiment analysis, where different modalities contribute unequally during joint learning. Since language usually provides denser semantic cues and more stable supervision, acoustic and visual modalities may be under-optimized, even when they contain useful affective evidence (Feng et al., 2024). Existing studies mainly address this issue from representation learning and optimization perspectives.

One line of work improves modality-specific representation learning. Self-MM generates unimodal label supervision for each modality, encouraging language, acoustic, and visual branches to learn sentiment-discriminative representations (Yu et al., 2021). DMD decomposes multimodal distillation into shared and modality-specific spaces, thereby enhancing individual modality representations under multimodal heterogeneity (Li et al., 2023). MFON further identifies under-optimized modalities and improves them through modal prompt attention, intra-modal distillation, and inter-modal

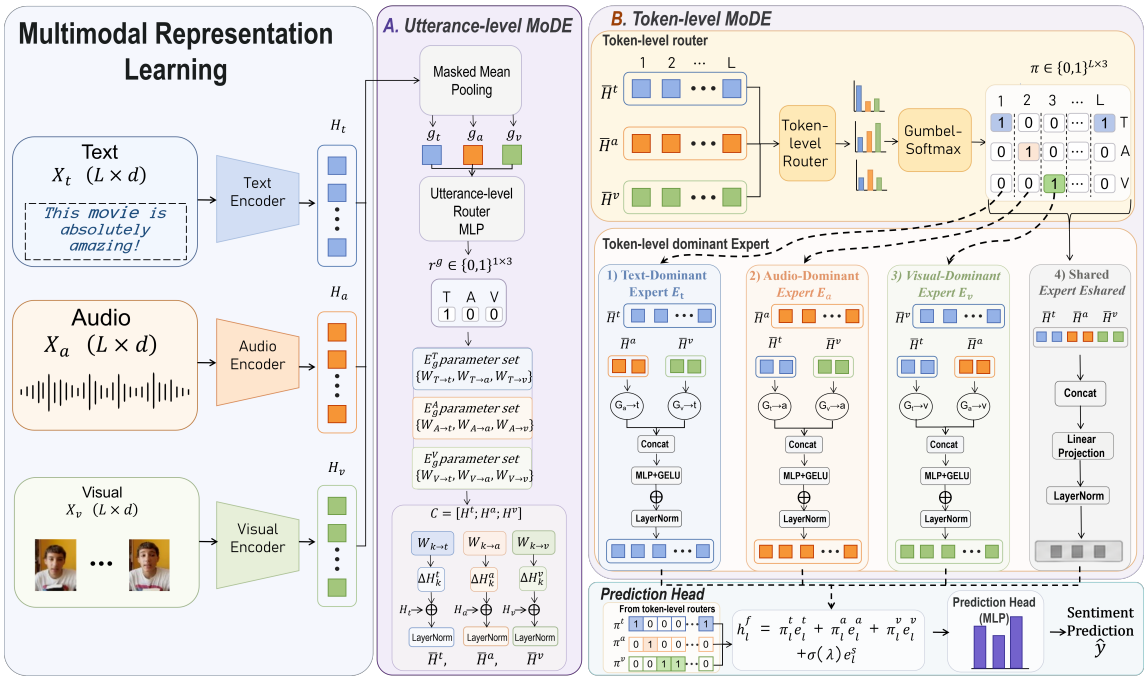

Figure 2: The model uses a global router to modulate multimodal features with macro modality priors, then selects token-level dominant experts and combines them with a shared expert for fusion and sentiment prediction.

contrastive learning (Zhang et al., 2025). These methods reduce modality imbalance by strengthening weaker modality representations.

Another line of work focuses on the optimization process. Gradient Blending analyzes the different overfitting and generalization behaviors of modalities and dynamically adjusts their training signals (Wang et al., 2020). OGM-GE modulates gradients online according to the learning status of each modality, reducing the suppression of weaker modalities (Peng et al., 2022). These optimization-based methods show that modality imbalance is not only caused by representation quality, but also by unequal learning dynamics during multimodal training. However, most existing studies treat imbalance mainly at the modality or sample level, while the inconsistency between utterance-level modality tendency and token-level affective evidence is still insufficiently explored.

2.2 Mixture of Experts

Mixture of Experts (MoE) uses a routing mechanism to assign different inputs to different expert networks, enabling conditional computation and adaptive representation learning. Early MoE models introduce a gating network to combine multiple local experts (Jacobs et al., 1991), while sparsely-gated MoE scales this idea to deep networks by activating only a small subset of experts for each input (Shazeer et al., 2017). Switch Transformer further simplifies token-level expert routing by assigning each token to a single expert, greatly improving model scalability (Fedus et al., 2022).

Recent studies have introduced MoE into multimodal emotion and sentiment analysis to model modality heterogeneity and sample-dependent modality importance. EMOE designs modality-specific enhanced emotion experts and dynamically selects experts for multimodal emotion recognition (Fang et al., 2025). AGS-SMoE combines adaptive gradient scaling with sparse expert routing to alleviate modal preemption in MSA (Chen et al., 2025). GLoMo explores global-local modal fusion for multimodal sentiment analysis, showing the benefit of modeling modality interactions from multiple levels (Zhuang et al., 2024). These studies suggest that MoE provides a flexible framework for dynamic multimodal fusion. However, existing MoE-based methods still face challenges in stable routing and fine-grained modality dominance modeling, especially when global modality tendency and local affective evidence are inconsistent.

3 Method

3.1 Overview

The overall framework is built upon a hierarchical mixture-of-experts. First, the **Utterance-level**

MoDE starts from the global multimodal representation of the entire utterance, estimates the utterance-level modality dominance, and selects the corresponding global dominant expert to perform macro-level enhancement over the three modality sequences. Then, the **Token-level MoDE** further conducts fine-grained dynamic routing on the enhanced sequences, selecting the corresponding modality-dominant expert for each token position and performing multimodal fusion.

3.2 Utterance-level MoDE

Utterance-level MoDE is the first dynamic expert module, which models global modality dominance from an utterance-level perspective. Given the encoded trimodal sequences $H^t, H^a, H^v \in \mathbb{R}^{L\times d}$, the model first applies masked mean pooling to each modality and obtains modality summaries over the whole utterance $g^m = \frac{\sum_{l=1}^{L} M_l h_l^m}{\sum_{l=1}^{L} M_l}$, where M_l indicates whether the l-th position is a valid token, so padding positions are excluded from the utterance-level summary.

Then, the three modality summaries are concatenated and fed into the utterance-level dominant modality router:

$$z^g = f_g([g^t; g^a; g^v]). \tag{1}$$

where, $z^g \in \mathbb{R}^3$ denotes the routing logits for the three utterance-level dominant experts, corresponding to text, audio, and vision. During training, the model obtains an approximately discrete routing vector $r^g \in \{0,1\}^3$ through Gumbel-Softmax $r^g = \mathrm{GumbelSoftmax}(z^g, \tau)$. During inference, it directly uses $\arg\max(z^g)$ to obtain the one-hot routing vector and select the utterance-level dominant expert. Here, each element r_k^g indicates whether the k-th global dominant expert is selected.

Utterance-level MoDE contains three modality-dominant experts $\{E_t^g, E_a^g, E_v^g\}$. Each expert receives the complete trimodal sequences and reconstructs the three modality representations $\tilde{H}_k^m$ conditioned on a specific dominant modality:

$$\begin{aligned}\tilde{H}_k^m &= E_k^g(H^t, H^a, H^v; \Delta_k^m),\\ &= \mathrm{LN}\Big(H^m + \Delta_k^m([H^t; H^a; H^v])\Big)\end{aligned} \tag{2}$$

where, $\Delta_k^m(C) = CW_{k\to m} + b_{k\to m}$ is a learnable projection branch in expert k, which maps the concatenated trimodal context C to the residual update for modality m. Thus, each dominant expert contains three modality-specific projection branches, denoted as $W_{k\to t}$, $W_{k\to a}$, and $W_{k\to v}$. In other words, it produces the residual term added to H^m before layer normalization.

Finally, the model selects the expert outputs according to the utterance-level routing vector:

$$\bar{H}^m = \sum_{k\in\{t,a,v\}} r_k^g \tilde{H}_k^m. \tag{3}$$

In this way, utterance-level MoDE provides an utterance-level modality prior for each sample and outputs the residual-calibrated trimodal sequences, which are then used as the input to token-level MoDE.

3.3 Token-level MoDE

Token-level MoDE selects the corresponding modality-dominant expert for each token position. Notably, Token-level MoDE does not directly inject the logits of the utterance-level dominant modality router. Instead, it performs token-level dominant modality routing in the representation space where the utterance-level modality prior has already been incorporated.

The token-level dominant modality router first uses independent bidirectional GRUs to contextually encode the three modalities:

$$U^m = \mathrm{BiGRU}_m(\bar{H}^m) \tag{4}$$

For the l-th position, the three token-level contextual representations are concatenated and fed into the routing network:

$$z_l^r = f_r([u_l^t; u_l^a; u_l^v]), \quad z_l^r \in \mathbb{R}^3. \tag{5}$$

where, u_l^m denotes the l-th contextual token in U^m. z_l^r represents the routing logits of three token-level dominant experts. During training, Gumbel-Softmax is used to generate the routing vector $\pi_l = \mathrm{GumbelSoftmax}(z_l^r, \tau)$, while $\pi_l = \arg\max(z_l^r)$ is adopted for hard selection during inference.

Subsequently, each local dominant expert uses one modality as the backbone and absorbs information from the other modalities through gating. Taking the text-dominant expert as an example:

$$\begin{aligned}e_l^t = \mathrm{LN}\Big(&\underbrace{\bar{h}_l^t}_{\text{text}} + \underbrace{\phi_{a\to t}([\bar{h}_l^a; \bar{h}_l^t]) \odot W_{a\to t}\bar{h}_l^a}_{\text{audio-to-text update}}\\ &+ \underbrace{\phi_{v\to t}([\bar{h}_l^v; \bar{h}_l^t]) \odot W_{v\to t}\bar{h}_l^v}_{\text{vision-to-text update}}\Big).\end{aligned} \tag{6}$$

where $\phi_{a\to t}(\cdot)$ and $\phi_{v\to t}(\cdot)$ are gating functions. $W_{a\to t}$ and $W_{v\to t}$ are learnable projection parameters that map audio and vision features into the text-dominant fusion space, respectively.

The audio-dominant and vision-dominant experts adopt symmetric structures, using audio or vision as the backbone and absorbing information from the remaining modalities. Given the three candidate expert outputs e_l^t, e_l^a, e_l^v, Token-level MoDE selects the token-level dominant path according to the routing vector: $h_l^{dom} = \sum_{m\in\{a,v,t\}} \pi_l^m e_l^m$.

Additionally, we propose a lightweight shared compensation path to preserve necessary cross-modal common information:

$$e_l^s = E_s([\bar{h}_l^t; \bar{h}_l^a; \bar{h}_l^v]), \tag{7}$$

$$h_l^f = h_l^{dom} + \sigma(\lambda) e_l^s. \tag{8}$$

where λ is a learnable scalar. Finally, the final fused sequence $H^f = \{h_l^f\}_{l=1}^L$ is passed through attention pooling and a unified prediction head to output the sentiment prediction.

3.4 Training Objective and Optimization Strategy

The model is trained with a two-stage routing strategy. During the first E_{pre} epochs, the MoDE modules use random logits for Gumbel-Softmax routing, so that different modality-dominant experts can be activated in the early training stage. Afterwards, the model switches to learnable routing, where the utterance-level dominant modality router and the token-level dominant modality router generate global and local routing logits, respectively. For Gumbel-Softmax routing, the temperature τ controls the sharpness of the routing distribution. We use an epoch-wise annealing schedule:

$$\tau_E = \max(\tau_{\min}, \tau_0 \gamma^E), \tag{9}$$

where τ_0, γ, and $\tau_{\min}$ denote the initial temperature, decay factor, and lower bound, respectively. This gradually changes routing from smooth exploration to near-hard expert selection. Proposal labels are generated from training-set unimodal prediction errors after warm-up; see Appendix A. To alleviate modality bias during training, the model further constructs dynamic perturbations based on the modality distribution p^g produced by the utterance-level dominant modality router. For modality $m \in \{t, a, v\}$, its masking probability is defined as:

$$q_m = \min(\eta_m p_m^g, \bar{q}_m), \tag{10}$$

where η_m is a scaling coefficient and $\bar{q}_m$ is the upper bound of the masking probability. This strategy imposes stronger perturbations on modalities with higher utterance-level routing weights during training, thereby encouraging token-level MoDE to exploit more cues from non-dominant modalities.

The joint optimization objective is defined as:

$$\mathcal{L} = \mathcal{L}_{task} + \alpha\mathcal{L}_{aux} + \beta\mathcal{L}_{router} - \gamma\mathcal{L}_{ent} \tag{11}$$

where $\mathcal{L}_{task}$ is the MSE loss of the final MSA prediction. $\mathcal{L}_{aux}$ is the average MSE loss of the three unimodal auxiliary prediction heads, which is used to maintain the discriminative ability of each modality branch. $\mathcal{L}_{router}$ is the cross-entropy loss between the utterance-level router logits and the proposal labels, and its weight β gradually decays during training. $\mathcal{L}_{ent}$ denotes the instance-level entropy of the utterance-level modality distribution, which is maximized in the overall objective to prevent premature collapse of the utterance-level routing distribution.

4 Experiments and Analysis

4.1 Experimental Setup

Datasets. Experiments are conducted on two standard MSA benchmark datasets, CMU-MOSI (Zadeh et al., 2016) and CMU-MOSEI (Bagher Zadeh et al., 2018). The CMU-MOSI dataset contains 2,199 utterance-level video segments, each annotated by human experts with a continuous sentiment intensity score in the range of $[-3, +3]$. We split this dataset into 1,284 training samples, 229 validation samples, and 686 test samples. CMU-MOSEI is a large-scale extension of CMU-MOSI, and it is split into 16,326 training samples, 1,871 validation samples, and 4,659 test samples.

Evaluation Metrics. We follow the standard evaluation protocol, report Mean Absolute Error (MAE) for the regression task, and binary accuracy (Acc-2), seven-class accuracy (Acc-7) and F1-score for classification.

Baselines. To verify the effectiveness of the proposed method, we select two categories of representative methods as baselines. The first category includes general fusion and interaction modeling methods, including TFN (Zadeh et al., 2017), LMF (Liu et al., 2018), MFN (Zadeh et al., 2018), Graph-MFN (Bagher Zadeh et al., 2018), MulT (Tsai et al., 2019), PS-Mixer (Lin et al., 2023), CMHFM (Wang et al., 2024), and

Table 1: Comparison results on CMU-MOSI dataset.

Method	MAE↓	Corr↑	F1↑	Acc-2↑	Acc-7↑
General Fusion / Interaction					
TFN (2017)	0.901	0.698	0.8007	0.8008	0.3490
LMF (2018)	0.917	0.695	0.8240	0.8250	0.3320
MFN (2018)	0.927	0.670	0.7890	0.7887	0.3583
Graph-MFN (2018)	0.956	0.649	0.7835	0.7835	0.3464
MulT (2019)	0.871	0.698	0.8200	0.8300	0.4000
PS-Mixer (2023)	0.748	0.794	0.8030	0.8030	0.4430
CMHFM (2024)	0.822	0.723	0.7930	0.7945	0.4213
GLoMo(2024)	0.7257	0.7895	0.8382	0.8382	0.4665
Modality-Aware / Imbalance-Oriented					
MISA (2020)	0.783	0.776	0.8170	0.8180	0.4230
MAG-BERT (2020)	0.790	0.766	0.8112	0.8213	0.4143
Self-MM (2021)	0.712	0.795	0.8268	0.8268	0.4580
DMMF (2024)	0.7367	0.8059	0.8493	0.8489	0.4190
DLF (2025)	0.731	0.781	0.8504	0.8506	0.4708
Ours	**0.6737**	**0.8215**	**0.8631**	**0.8641**	**0.4818**

Table 2: Comparison results on CMU-MOSEI dataset.

Method	MAE↓	Corr↑	F1↑	Acc-2↑	Acc-7↑
General Fusion / Interaction					
TFN (2017)	0.593	0.700	0.8210	0.8250	0.5020
LMF (2018)	0.623	0.677	0.8210	0.8200	0.4800
MFN (2018)	0.575	0.718	0.8285	0.8285	0.5134
Graph-MFN (2018)	0.575	0.713	0.8343	0.8348	0.5137
MulT (2019)	0.580	0.703	0.8230	0.8250	0.5180
PS-Mixer (2023)	0.537	0.765	0.8310	0.8310	0.5300
CMHFM (2024)	0.548	0.747	0.8418	0.8431	0.5261
GLoMo(2024)	0.5412	0.7507	0.8370	0.8343	0.5299
Modality-Aware / Imbalance-Oriented					
MISA (2020)	0.555	0.756	0.8380	0.8360	0.5220
MAG-BERT (2020)	0.543	0.755	0.8277	0.8251	0.5267
Self-MM (2021)	**0.529**	**0.767**	0.8295	0.8268	0.5346
DMMF (2024)	0.5503	0.7617	0.8529	0.8529	0.5191
DLF (2025)	0.536	0.764	0.8527	0.8542	0.5390
Ours	0.5354	0.7595	**0.8552**	**0.8564**	**0.5542**

GLoMo (Zhuang et al., 2024). The second category includes modality-aware and imbalance modeling methods, including MISA (Hazarika et al., 2020), MAG-BERT (Rahman et al., 2020), Self-MM (Yu et al., 2021), DMMF (Wu et al., 2024), and DLF (Wang et al., 2025).

Implementation Details. The proposed model is implemented based on the PyTorch framework with dual Intel Xeon Gold 6248R processors and 10 NVIDIA A30 GPUs with 24 GB memory each. The system environment is based on CUDA 12.5. We optimize the model with AdamW and a cosine annealing scheduler. The learning rates are set to 1×10^{-5} for XLNet, 2×10^{-3} for acoustic/visual branches, and 5×10^{-3} for the shared coefficient, with a weight decay of 0.01. The training batch size is 32, and the validation/test batch size is 128. All experiments are trained for at most 120 epochs with an early stopping patience of 25.

4.2 Comparison with State-of-the-Art

CMU-MOSI. On CMU-MOSI, the proposed method achieves the best results across all metrics, as shown in the Table 1. Compared with the strongest baseline, MAE decreases from 0.712 to 0.6737, and Corr improves from 0.8059 to 0.8215. For classification metrics, F1 and Acc-2 reach 0.8631 and 0.8641, respectively, while Acc-7 is also improved to 0.4818. These results indicate that the proposed method not only improves overall regression performance, but also distinguishes fine-grained sentiment intensity more effectively.

CMU-MOSEI. On CMU-MOSEI, the proposed method achieves the best results on F1, Acc-2, and Acc-7, as shown in the Table 2. Although Self-MM performs slightly better on MAE and Corr, our method reaches an Acc-7 of 0.5542, which is clearly higher than 0.5390 achieved by DLF. This demonstrates that the proposed method has stronger discriminative capability for sentiment polarity and fine-grained category boundaries in large-scale and complex scenarios.

Results Analysis. From the overall trend, general fusion methods such as TFN, LMF, MFN, and MulT mainly model cross-modal interactions, but they usually adopt a fixed fusion paradigm and thus struggle to adapt to modality dominance differences across samples. Text-enhanced or text-guided methods can improve stability by leveraging strong language representations, but they may also implicitly assume text dominance. When key affective evidence comes from speech or vision, their ability to exploit non-verbal modalities is still limited. In contrast, the proposed method models global modality dominance through utterance-level MoDE and further captures local modality variations through token-level MoDE, thereby addressing modality imbalance at both global and local levels. In particular, the consistent improvement in Acc-7 on both datasets demonstrates that the utterance-token-level MoDE structure can make fuller use of local non-textual cues and is more effective for fine-grained sentiment recognition.

4.3 Ablation Studies

To further verify the effectiveness and necessity of the core design components in the proposed HD-MoDE architecture, we design and evaluate four ablation variants on the MOSI and MOSEI datasets. The variants are defined as follows:

- **w/o Utterance-level MoDE**: removes the utterance-level MoDE module, feeding three

Table 3: Ablation results on MOSI and MOSEI.

Model	MAE↓	Corr↑	F1↑	Acc-2↑	Acc-7↑
			CMU-MOSI		
Full Model	**0.6737**	**0.8215**	**0.8631**	**0.8641**	**0.4818**
w/o Utterance-level MoDE	0.6916	0.8169	0.8599	0.8611	0.4613
w/o Token-level MoDE	0.9038	0.8194	0.8566	0.8565	0.3168
w/o MoE	0.6973	0.8166	0.8592	0.8595	0.4803
Text-dominance	0.7073	0.8213	0.8519	0.8519	0.4526
Audio-dominance	0.7053	0.8185	0.8607	0.8611	0.4365
Vision-dominance	0.7062	0.8049	0.8534	0.8534	0.4453
			CMU-MOSEI		
Full Model	**0.5354**	**0.7595**	**0.8552**	**0.8564**	**0.5542**
w/o Utterance-level MoDE	0.5646	0.7571	0.8519	0.8558	0.5204
w/o Token-level MoDE	0.6628	0.7581	0.8542	0.8558	0.4620
w/o MoE	0.5581	0.7589	0.8496	0.8517	0.5150
Text-dominance	0.5495	0.7543	0.8465	0.8461	0.5281
Audio-dominance	0.5449	0.7566	0.8501	0.8525	0.5288
Vision-dominance	0.5704	0.7522	0.8226	0.8207	0.5275

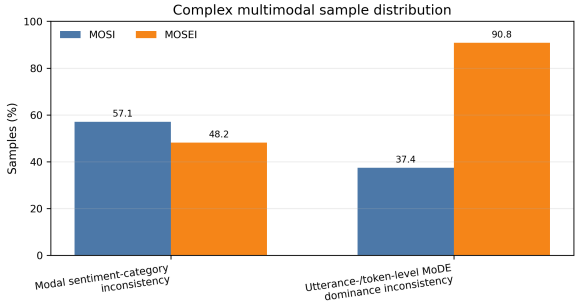

Figure 3: Statistics of complex samples on MOSI and MOSEI. The results show that modality-wise sentiment inconsistency and global-local dominance inconsistency are common in MSA.

modalities' features directly into the subsequent local fusion module.

- **w/o Token-level MoDE**: removes the token-level MoDE module, degenerating trimodal feature fusion into static fusion.
- **w/o MoDE**: removes both the utterance-level MoDE and token-level MoDE , reducing the model to a standard static fusion scheme.
- **w/ Unimodal-dominate**: fix the dominant modality in both the utterance-level and token-level routers to a single modality, replacing adaptive routing with unimodal dominance.

As shown in the table 3, the full model achieves the best results on both datasets, indicating that each module contributes to performance improvement. Among them, removing token-level MoDE leads to the most significant degradation. In particular, Acc-7 drops from 0.4818 to 0.3168 on MOSI and from 0.5542 to 0.4620 on MOSEI, demonstrating that token-level modality-dominance selection is crucial for fine-grained sentiment recognition.

After removing Utterance-level MoDE, the model shows declines in both MAE and Acc-7, indicating that the global modality prior can provide effective guidance for subsequent local routing. In contrast, **w/o MoE** degenerates into static fusion, further limiting overall performance and verifying the necessity of the dynamic MoE structure for modeling changes in modality dominance. Overall, the ablation results verify the effectiveness of the proposed global-local design.

To verify the limitations of fixed unimodal dominance, we fix the dominant modality routers in both utterance-level MoDE and token-level MoDE to a single modality. The full model consistently outperforms these fixed unimodal-dominant variants on both datasets, showing that the optimal affective source is not static. By adaptively modeling utterance-level and token-level dominant modalities, our routers better exploit complementary multimodal cues and alleviate the bias caused by fixed modality dominance.

4.4 In-depth Analysis of HD-MoDE

Given the effectiveness of HD-MoDE, we further conduct an in-depth analysis of its hierarchical routing behavior. Specifically, we discuss three questions: 1) whether global-local modality imbalance exists in MSA; 2) how HD-MoDE coordinates utterance-level and token-level modality dominance; and 3) whether dynamic routing is more reliable than fixed routing.

Q1: Does global-local affective evidence imbalance really exist? Yes. To verify this, we analyze two types of complex samples. The first type is *samples with inconsistent modality-wise sentiment categories*, where the auxiliary branches of text, audio, and vision are mapped to Acc-7 categories and produce inconsistent predictions. The second type is *samples with global–local affective evidence imbalance*, where the dominant modality selected by the utterance-level router is inconsistent with the majority modality (route switching rate > 0.35).

As shown in Fig. 3, 57.1% and 48.2% of samples in MOSI and MOSEI, exhibit inconsistent modality-wise sentiment categories, indicating that different modalities often carry diverse affective evidence. Meanwhile, the proportions of global-local dominance inconsistency reach 37.4% and 90.8% on MOSI and MOSEI. This suggests that the sample-level dominant modality is not always

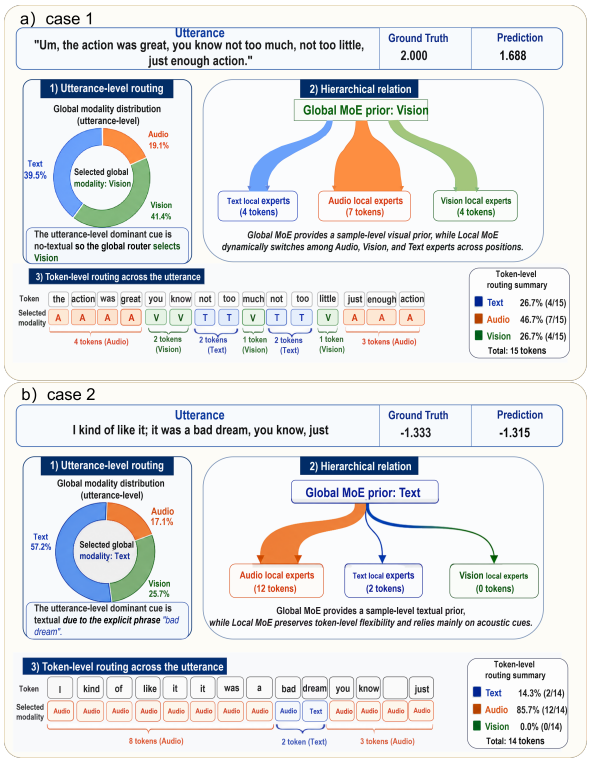

Figure 4: Case analysis of hierarchical routing behavior. The utterance-level router captures the global dominant modality, while the token-level router dynamically selects local experts according to fine-grained affective evidence.

consistent with the source of local affective evidence. These results directly support the necessity of modeling modality dominance at both the utterance level and the token level.

Q2: How does HD-MoDE handle inconsistency between global dominance and local evidence? HD-MoDE decouples global judgment and local selection through a two-level MoDE design. The utterance-level MoDE first determines the dominant modality of the whole utterance and provides a sample-level affective prior. The token-level MoDE then reselects the Text, Audio, or Vision expert at each local position, thereby preserving fine-grained adjustment ability.

As shown in Fig. 4, the two samples show different utterance-level dominance. In Case 1, the utterance-level router selects Vision as the global dominant modality, while the token-level router still switches among Text, Audio, and Vision experts. In Case 2, the utterance-level router selects Text, but the token-level router mainly shifts to Audio experts. These results show that HD-MoDE can capture global dominance while preserving local routing flexibility.

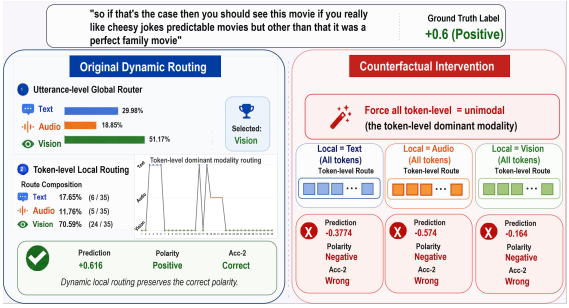

Figure 5: Counterfactual analysis of fixed routing. Replacing the original dynamic hierarchical routing with fixed routing paths leads to incorrect predictions, demonstrating the necessity of adaptive global-local routing.

Q3: Is dynamic hierarchical global-local routing better than fixed routing? Yes. We conduct a counterfactual intervention on a sample that is originally predicted correctly. As shown in Fig. 5, the utterance-level router selects Vision, while the token-level router dynamically assigns Text, Audio, and Vision experts in the original HD-MoDE, finally producing the correct positive prediction. Instead, when the global dominant modality is kept on Vision, and all token-level routes are fixed to uni-Text, Audio, or Vision, respectively, all three settings incorrectly predict the sample as negative. This indicates that the key advantage of HD-MoDE lies in maintaining the adaptiveness of hierarchical global-local routing. Once this dynamic combination is disrupted, even enumerating fixed local modality paths can hardly recover the correct prediction, further verifying the effectiveness of dynamic hierarchical routing in complex samples.

5 Conclusion

In this paper, we investigated Global-Local Modality Imbalance in multimodal sentiment analysis and proposed HD-MoDE, a Hierarchical Dynamic Mixture-of-Dominant-Experts framework. HD-MoDE models modality dominance at both the utterance level and the token level, allowing the model to capture global modality tendencies while dynamically adjusting to local affective evidence. Experiments on CMU-MOSI and CMU-MOSEI demonstrate the effectiveness of the proposed method, especially for fine-grained multimodal sentiment recognition.

Limitations

This work has several limitations. First, our experiments are mainly conducted on CMU-MOSI and

CMU-MOSEI, so the generalization of HD-MoDE to other multimodal tasks and more diverse real-world scenarios remains to be further investigated. Second, the hierarchical routing design introduces additional computation compared with static fusion methods, which may limit its efficiency in resource-constrained settings. Third, although routing analysis helps interpret global-local modality dominance, the learned routing paths are still not guaranteed to fully align with human judgments of affective evidence. Future work will explore more efficient and interpretable routing mechanisms and evaluate the proposed framework on broader multimodal understanding tasks.

References

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*, pages 2236–2246.

Jili Chen, Qionghao Huang, Changqin Huang, and Xiaodi Huang. 2025. Actual cause-guided adaptive gradient scaling for balanced multimodal sentiment analysis. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 21(6):1–24.

Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, and Mang Ye. 2025. EMOE: Modality-specific enhanced dynamic emotion experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14314–14324.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Xinyu Feng, Yuming Lin, Lihua He, You Li, Liang Chang, and Ya Zhou. 2024. Knowledge-guided dynamic modality attention fusion framework for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14755–14766.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1122–1131.

Jingming Hou, Nazlia Omar, Sabrina Tiun, Saidah Saad, and Qian He. 2024. TCHFN: Multimodal sentiment analysis based on text-centric hierarchical fusion network. *Knowledge-Based Systems*, 300:112220.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. 1991. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87.

Yong Li, Yuanzhi Wang, and Zhen Cui. 2023. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6631–6640.

Han Lin, Pinglu Zhang, Jiading Ling, Zhenguo Yang, Lap Kei Lee, and Wenyin Liu. 2023. PS-Mixer: A polar-vector and strength-vector mixer model for multimodal sentiment analysis. *Information Processing & Management*, 60(2):103229.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2247–2256.

Zijun Liu, Li Cai, Wenjie Yang, and Junhui Liu. 2024. Sentiment analysis based on text information enhancement and multimodal feature fusion. *Pattern Recognition*, 156:110847.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. 2022. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8238–8247.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569.

Lan Wang, Junjie Peng, Cangzhi Zheng, Tong Zhao, and Li'an Zhu. 2024. A cross modal hierarchical fusion multimodal sentiment analysis method based on multi-task learning. *Information Processing & Management*, 61(3):103675.

Pan Wang, Qiang Zhou, Yawen Wu, Tianlong Chen, and Jingtong Hu. 2025. DLF: Disentangled-language-focused multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21180–21188.

Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12695–12705.

Jun Wu, Jiangpeng Wang, Shilong Jing, Jinyu Liu, Tianfeng Zhang, Min Han, Pengfei Zhan, and Gan Zuo. 2024. Text-dominant strategy for multistage optimized modality fusion in multimodal sentiment analysis. *Multimedia Systems*, 30(6):353.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. 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, pages 10790–10797.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 5634–5641.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

Xiangmin Zhang, Wei Wei, and Shihao Zou. 2025. Modal feature optimization network with prompt for multimodal sentiment analysis. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4611–4621.

Yan Zhuang, Yanru Zhang, Zheng Hu, Xiaoyue Zhang, Jiawen Deng, and Fuji Ren. 2024. GLoMo: Global-local modal fusion for multimodal sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1800–1809.

A Construction of Dominant-Modality Proposal Labels

To provide sample-level supervision for the utterance-level dominant modality router, we construct dominant-modality proposal labels before the main training stage. Since the text branch usually benefits from pretrained language models such as XLNet and may converge faster than acoustic and visual branches, directly deriving proposal labels from late-stage unimodal predictions can make the labels overly biased toward text and reduce the exploration space of non-textual modalities.

Therefore, we adopt an independent modality-expert warm-up strategy. During warm-up, no proposal labels or routing supervision are used. Instead, the text, acoustic, and visual auxiliary prediction heads are optimized independently to fit the ground-truth sentiment labels. The warm-up objective is the sum of three unimodal MSE losses:

$$\mathcal{L}_{warm} = \mathcal{L}_t^{mse} + \mathcal{L}_a^{mse} + \mathcal{L}_v^{mse}. \tag{12}$$

After warm-up, the model is switched to evaluation mode, and the sample-level MSE of the three unimodal auxiliary prediction heads is computed on the training set. For each training sample, the modality with the smallest prediction error is selected as its proposal dominant modality:

$$y_i^p = \arg \min_{m \in \{t,a,v\}} (\hat{y}_i^m - y_i)^2. \tag{13}$$

If the text, acoustic, or visual branch obtains the minimum error, the sample is assigned to the Text, Audio, or Vision proposal label, respectively.

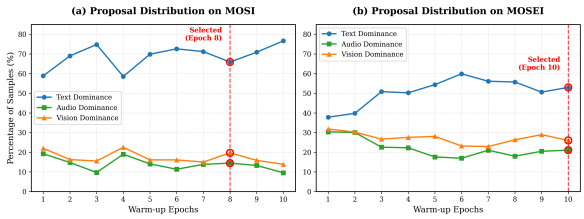

Figure 6: Proposal label distributions under different warm-up epochs on CMU-MOSI and CMU-MOSEI. The selected warm-up epochs are marked by red dashed lines.

All three unimodal losses are assigned the same weight of 1.0. Thus, the proposal labels are determined solely by sample-level unimodal prediction errors, without introducing additional manual modality bias. To avoid premature randomness and late-stage text over-dominance, we monitor the proposal distribution during warm-up and select the endpoint when the unimodal predictors have obtained basic discriminative ability while non-textual proposal labels are still preserved. As shown in Figure 6, the selected warm-up epoch is 8 for CMU-MOSI, where the proposal distribution is 65.89% Text, 14.44% Audio, and 19.67% Vision. For CMU-MOSEI, the selected warm-up epoch is 10, where the distribution is 52.97% Text, 21.04% Audio, and 25.99% Vision.

The constructed proposal labels are attached only to the training set and are used solely for supervising the utterance-level router. No proposal labels are used for the validation or test sets.