Rethinking Attention Reallocation for Multimodal Emotion Recognition

Yazhe Lyu[1], Yixiong Zou[1⋆], Jinghan Hu[2,3], Yuhua Li[1], and Ruixuan Li[1]

[1] School of Computer Science and Technology, Huazhong University of Science and Technology, China
[2] Key Laboratory of Adolescent Cyberpsychology and Behavior (CCNU), Ministry of Education, China
[3] Key Laboratory of Human Development and Mental Health of Hubei Province, School of Psychology, Central China Normal University, China
{yazhelv, yixiongz, idcliyuhua, rxli}@hust.edu.cn, hujinghan@ccnu.edu.cn

Abstract. The emergence of Multimodal Large Language Models has enabled a new generative paradigm for multimodal emotion recognition (MER), where emotional interpretations are produced through multimodal token-level reasoning. In view of multimodal inputs, existing studies commonly observe that reallocating attention from generated tokens to modality tokens during inference consistently improves performance. However, in this work, we reveal a counterintuitive phenomenon: such a strategy only benefits two-modality settings, but degrades performance in three- or more-modality scenarios, where the opposite reallocation direction instead leads to improvements. To understand this phenomenon, through comprehensive analysis, we show that with more modalities, the complex interactions between modalities make the model hard to pay more attention to tokens with more information, especially in deeper layers, leading to improper attention allocation and the observed contradictory behavior. Based on these insights, we propose a training-free attention rectification method that leverages structured shallow-layer attention as a prior to regularize entangled final-layer attention during inference, without introducing additional parameters or modifying the backbone model. Extensive experiments on nine datasets in MER-UniBench demonstrate that our method achieves state-of-the-art performance, consistently outperforming both reallocation directions across diverse multimodal scenarios. Our code is available at https://github.com/yzl77/ReAR.

Keywords: Multimodal Large Language Model · Multimodal Emotion Recognition · Attention Reallocation

1 Introduction

Understanding and accurately perceiving human emotions is fundamental to advancing human–computer interaction, laying the foundation for practical appli-

⋆ Corresponding author.

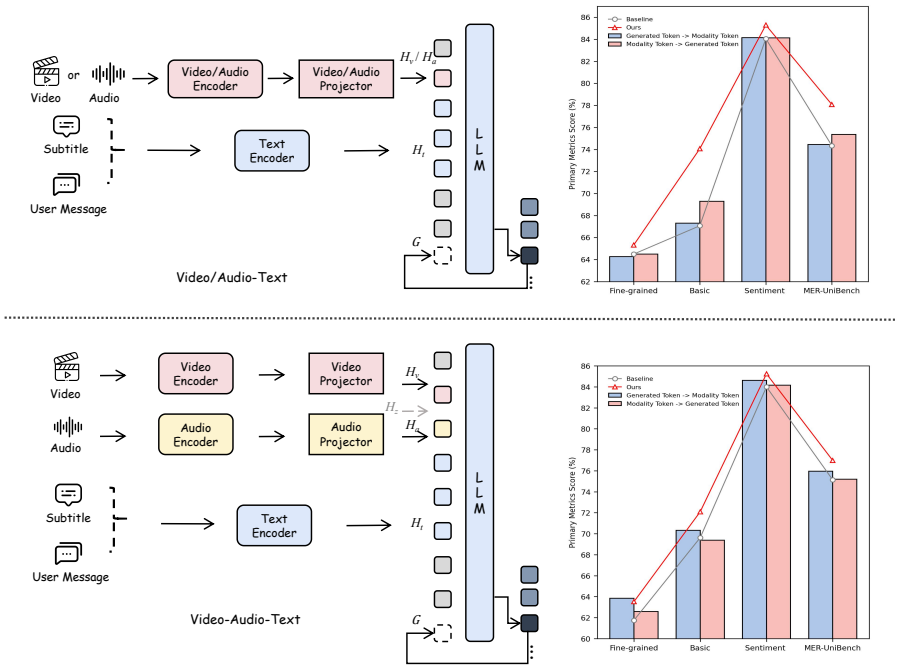

Fig. 1: The contradictory effects of attention reallocation across modality settings. Contrary to common beliefs, we find that reallocating attention from generated tokens to modality tokens **only** helps in two-modality settings, but harms in three-modality settings, where the opposite direction instead yields improvements.

cations such as educational assistance [15] and psychological counseling [2, 14]. Typically, an emotion is described by multimodal signals such as video/facial [16, 19], audio [7, 11], or text. Recent advances in Multimodal Large Language Models (MLLMs) [29, 37, 50] have introduced a new paradigm for Multimodal Emotion Recognition (MER), which takes multimodal signals as inputs and generates descriptive emotion expressions in natural language, enabling more flexible and interpretable affect modeling [3, 10, 24].

In view of multimodal inputs, current works [31, 44] have observed that the model tends to prioritize newly generated tokens while ignoring multimodal tokens (i.e., modality tokens), and proposed to reallocate attention from generated tokens to modality tokens during inference for better performance. However, in this paper, we find an interesting phenomenon contradicting this common belief: such a reallocation operation helps only under two-modality settings (e.g., video-text or audio-text), but harms the performance under three- or more-modality settings (e.g., video-audio-text, as shown in Fig. 1). Instead, in the three-modality scenario, reallocating attention in the **opposite** direction (i.e., from modality tokens to generated tokens) can inversely improve the perfor-

mance. In our experiments, we observe the wide existence of this phenomenon, but it still remains underexplored in current works.

In this paper, we aim to take a step closer to this contradictory reallocation phenomenon. Instead of the attention that should be reallocated differently, we find that in all scenarios where the attention reallocation (in either direction) can improve performance, the contribution of modality tokens consistently increases, which jointly considers attention and the information of each token. This indicates that under more-modality settings, it becomes harder for the model to pay more attention to tokens that indeed contain more information, which also holds for tokens other than generated or modality ones. Further analysis of the layer-wise attention evolution reveals that attention distributions become increasingly entangled in deeper layers, especially when more modality tokens are involved, as a result of the complex interactions between modalities. Such an ineffectively modeled modality information leads to the incorrect allocation of attention, causing the contradictory reallocation phenomenon.

Based on the above interpretation, we further propose a method to rectify the attention allocation caused by ineffectively modeled modality interactions. Since deeper layers show entangled attention distributions, we also observe that shallow-layer attention, in contrast, remains relatively structured and closer to the initial token-level signals. This motivates us to incorporate shallow-layer attention as a structural prior to regularize the final-layer attention distribution. Specifically, our method consists of two components: (1) a shallow-layer attention extraction module that captures early-stage structural cues, and (2) an attention refinement module that adaptively integrates this guidance into the last-layer attention map. The proposed mechanism is training-free and applied during inference, without introducing additional parameters or modifying the backbone model. We conduct extensive experiments on nine datasets in MER-UniBench [24]. Experimental results show that our method achieves state-of-the-art performance on this benchmark, consistently increasing the contribution of modality tokens and outperforming both reallocation directions under different modality settings (Fig.1).

The main contributions of this paper are summarized as follows:

- We find an interesting attention reallocation phenomenon: contrary to common belief, reallocating attention from generated tokens to modality tokens only benefits two-modality settings, while in three-modality scenarios, the opposite reallocation direction instead improves performance.
- We analyze the mechanism behind this phenomenon: with more modalities, the complex interactions between modalities make the model hard to pay more attention to tokens with more information, especially in deeper layers, leading to this phenomenon.
- Based on the above interpretation, we propose a training-free attention rectification method that leverages shallow-layer attention as a structural prior to regularize the entangled final-layer attention, enabling adaptive refinement during inference without additional parameters or structural modification.
- Extensive experiments on nine datasets in MER-UniBench show that our method consistently achieves state-of-the-art performance across the bench-

mark, while consistently increasing the contribution of modality tokens and outperforming both reallocation directions across diverse multimodal scenarios.

2 Related Works

Multimodal Large Language Models (MLLMs) The rapid development of large language models (LLMs) [42, 43] has led to the emergence of MLLMs [30,46,47], which integrate multimodal information into unified generative frameworks. Benefiting from the strong reasoning and generation capabilities of autoregressive LLMs, representative MLLMs typically combine a pretrained modality encoder [6, 8] with a decoder-only language model [43, 56], enabling cross-modal understanding and instruction-following ability. These models have achieved strong performance across diverse multimodal tasks, including image-language [13, 49], video-language [33], and audio-language [4] scenarios. Despite these advances, effectively understanding and leveraging complex interactions among multiple modalities remains a significant challenge for MLLMs.

Multimodal Emotion Recognition and Reasoning Early studies mainly address emotion understanding through video captioning [39] and fixed-label multimodal emotion classification [25]. With the rapid development of MLLMs, recent efforts have shifted toward open-ended multimodal emotion reasoning, where models generate both predictions and textual explanations [3, 24]. Several benchmarks and models have been proposed to support this direction by integrating instruction tuning and multimodal alignment [10, 51]. However, in MER tasks, how model architectures and internal mechanisms interact with different modalities, and how such interactions affect emotion understanding and reasoning, remains insufficiently studied.

Attention Reallocation Recent efforts have explored attention-based strategies [31,44] to improve the inference behavior of multimodal large language models. These methods typically adjust the distribution of attention weights across token groups, aiming to better emphasize task-relevant information while suppressing spurious correlations, thereby alleviating issues such as hallucination. However, how such attention reallocation strategies behave in MER, especially under varying numbers of modalities, remains insufficiently studied. In this work, we analyze attention reallocation behavior in MER under different modality scenarios and propose a training-free attention refinement method.

3 Rethinking Attention Reallocation between Generated and Modality Tokens

3.1 Preliminaries

Task Formulation and Model Overview Multimodal Emotion Recognition (MER) aims to infer human emotional states by integrating multimodal signals such as visual, acoustic, and textual inputs. Given a multimodal input

sample, the objective is to predict its corresponding emotion representation, which can be formulated as categorical labels, affective dimensions, or descriptive emotional interpretations. With the rapid advancement of MLLMs [42, 50], MER has increasingly shifted toward a generative paradigm. In this setting, tokenized multimodal inputs are processed through cross-modal reasoning, and the model produces natural language descriptions of emotional states rather than fixed categorical labels. In this work, we adopt AffectGPT [24] as our baseline model, a representative framework for MLLM-based MER. Following its modeling paradigm, we describe the multimodal processing pipeline as follows.

Formally, for each sample, the visual, audio, and textual inputs are denoted as X_v, X_a, and X_t, respectively. Visual and acoustic features are first extracted using pretrained encoders and then projected into latent token representations:

$$H_v = G_v(E_v(X_v)), \quad H_a = G_a(E_a(X_a)), \tag{1}$$

where E_v and E_a denote pretrained encoders, and G_v and G_a are learnable projection layers.

The textual input X_t is concatenated with the instruction Q to form a prompt and encoded into token embeddings H_t via the language model's embedding layer. In the full three-modality setting (video–audio–text), the visual and acoustic tokens are first processed through a pre-fusion module, and the resulting fused representation is projected to obtain an additional multimodal token H_z before being fed into the LLM decoder. The decoder therefore receives H_t, H_v, H_a, and H_z as input. In partial-modality settings (e.g., video–text or audio–text), the available modality tokens together with H_t are directly fed into the LLM decoder, and the fused token H_z is not constructed. This tokenized multimodal input is then processed by the LLM decoder for contextual reasoning and response generation.

Autoregressive Generation The model generates responses in an autoregressive manner. At each decoding step, the hidden representations are processed through N stacked transformer layers to produce contextualized outputs.

To characterize the decoding process, the architecture of the transformer layers is detailed below. Let $h_i = \{h_i^1, \ldots, h_i^T\}$ denote the hidden states of the i-th transformer layer, where T is the sequence length. Each layer performs self-attention to enable information exchange across all tokens. The attention mechanism is defined as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V, \tag{2}$$

where Q, K, and V are linear projections of the input tokens, and d_k denotes the feature dimension. Through self-attention, information from different modalities is adaptively integrated.

Finally, the hidden state of the last token in the final layer, h_N^T, is projected to the vocabulary space via an affine transformation $\phi(\cdot)$ to predict the next token:

$$P(r_T \mid r_{<T}) = \text{softmax}(\phi(h_N^T)), \quad r_T \in \mathcal{V}. \tag{3}$$

3.2 A Preliminary Exploration of Attention Reallocation

Recent studies [31,44] have observed that MLLMs tend to assign excessive attention to newly generated tokens during decoding, while underutilizing multimodal tokens. To mitigate this issue, they introduce training-free attention reallocation strategies that adjust the attention distribution between generated tokens and modality tokens at inference time.

To investigate the impact of attention allocation on MER, we follow the above attention reallocation paradigm during inference. Specifically, we perform a direct redistribution of the attention mass on the normalized attention weights of the final layer. We partition the tokens into two disjoint sets: modality tokens $\mathcal{M}$ (including video H_v, audio H_a, and multimodal features H_z if there is) and generated tokens $\mathcal{G}$ produced during decoding. Let $A_{i,j}$ denote the normalized attention weight at the last layer from the current query position i to token j.

To enable controlled reallocation, we first quantify the total attention mass assigned to modality tokens and compute a scaled portion governed by a hyperparameter $\alpha \in [0, 1]$:

$$\Delta = \alpha \sum_{j \in \mathcal{M}} A_{i,j}. \tag{4}$$

The attention weights corresponding to modality tokens are subsequently scaled down in a proportional manner:

$$A'_{i,j} = (1 - \alpha) A_{i,j}, \quad \forall j \in \mathcal{M}. \tag{5}$$

The extracted attention mass Δ is then redistributed to generated tokens in accordance with their original relative proportions:

$$A'_{i,j} = A_{i,j} + \frac{A_{i,j}}{\sum_{k \in \mathcal{G}} A_{i,k}} \cdot \Delta, \quad \forall j \in \mathcal{G}. \tag{6}$$

Finally, the updated attention weights are normalized to preserve a valid probability distribution:

$$A^{\text{final}}_{i,j} = \frac{A'_{i,j}}{\sum_k A'_{i,k}}. \tag{7}$$

Building upon the above reallocation procedure, we conduct extensive experiments on MER-UniBench [24] to evaluate its effectiveness across different modality settings. We observe a notable phenomenon that contradicts the common assumption: the proposed reallocation strategy improves performance in two-modality scenarios (e.g., video–text or audio–text), whereas it leads to performance degradation in three- or multi-modality settings (e.g., video–audio–text, as shown in Fig. 1). This phenomenon consistently appears across different benchmarks and modality configurations in our experiments. However, it remains largely underexplored in existing studies, and its underlying mechanism has not been systematically investigated.

Method	Baseline	$\mathcal{G}\rightarrow\mathcal{M}$	$\mathcal{M}\rightarrow\mathcal{G}$	Ours
Two-Modality				
Score	78.17	79.04 ↑	78.11 ↓	**79.48** ↑
Ratio	1.4015	1.3969 ↓	1.4023 ↑	**1.3480** ↓
Three-Modality				
Score	76.52	76.52 ↓	80.53 ↑	**84.49** ↑
Ratio	0.2128	0.2130 ↑	0.2116 ↓	**0.2099** ↓

Table 1: Comparison of performance and contribution ratio across different modality settings. In all scenarios where the attention reallocation (in either direction) can improve performance, the contribution of modality tokens consistently increases. Our method obtains the lowest contribution ratio among all methods.

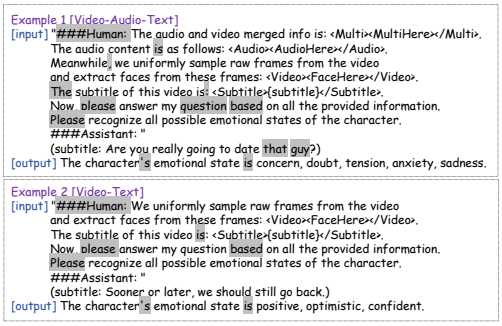

Fig. 2: Visualization of examples in MER task. Numerous tokens with high attention weights contribute little meaningful information in the corresponding sentence context.

3.3 Exploring the Contradictory Reallocation Phenomenon

In particular, the sensitivity to attention reallocation suggests that attention weights themselves may not faithfully reflect the true informational contribution of tokens. This raises a fundamental question: do the tokens receiving higher attention indeed contain more information?

To better account for token-level information strength, we move beyond raw attention weights and consider both attention allocation and information within the Transformer aggregation process. Accordingly, we quantify token-level contribution by combining attention weights with value L_2 magnitude. For each token type $\mathcal{G}$, we define its contribution as follows:

$$C_{\mathcal{G}} = \frac{1}{|\mathcal{G}|} \sum_{j \in \mathcal{G}} \left(\mathrm{Attn}_j \cdot \mathrm{Val}_j \right), \tag{8}$$

Using the proposed contribution metric, we further analyze the behavior of attention reallocation strategies that yield performance improvements. Taking the results on the MER2023 dataset [25] as an example (Tab. 1), we observe a consistent decrease in the ratio between the contributions of generated tokens and modality tokens compared to the baseline.

Specifically, in the two-modality setting, increasing attention to modality tokens consistently leads to higher modality contribution and improved performance, suggesting that modality tokens indeed exhibit higher L_2 magnitude (i.e., more information). Therefore, attention allocation effectively reflects the relative informational strength of tokens. However, this relationship becomes less stable in the three-modality setting. In several datasets, performance improvements are not achieved by further increasing modality attention; instead, reallocating attention toward other tokens yields better results. This observation indicates that, under more-modality scenarios, modality tokens may not exhibit L_2 magnitude large enough to reallocate attention to. That is, with more modalities,

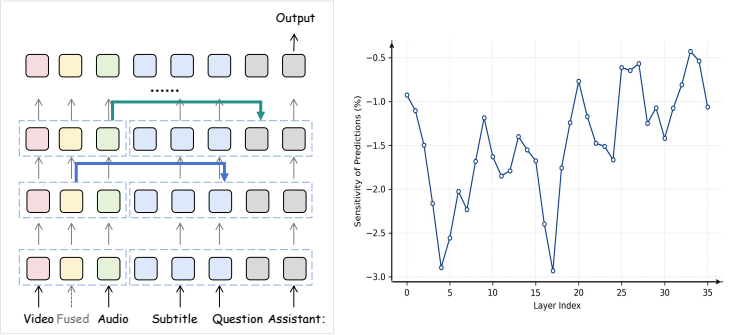

Fig. 3: Layer-wise evolution of attention. As representations move from early to deep layers, cross-modal interactions undergo multiple stages of refinement, shifting from structured attention patterns in early layers to increasingly entangled distributions in deeper layers.

the model cannot extract enough information from modality tokens, but still assigns higher attention to them.

Notably, this phenomenon extends beyond modality and generated tokens, affecting the broader token distribution. From the visualization in Fig. 2, we observe that numerous tokens with high attention weights contribute little meaningful information in the corresponding sentence context. In all, with more modalities involved, it becomes increasingly difficult for the model to consistently assign higher attention weights to tokens that indeed contain richer information.

3.4 Shallow Attention as a Structural Prior

To investigate how this problem occurs, we analyze the layer-wise evolution of attention from shallow to deep layers, following prior work [57]. Specifically, we intervene in the information flow from modality tokens to subsequent tokens during decoding, simulating the disruption of modality-level signals. Within this framework, we investigate how attention perturbations influence information flow during decoding by applying attention knockout to the connections from modality tokens to subsequent tokens. We measure the performance degradation ratio under both the original setting (without intervention) and the knockout setting to quantify the impact of removing modality-token connections, i.e., the lower the Y-axis value, the stronger the influence of removing modality tokens. The detailed experimental setup is provided in the Appendix.

Based on this setup, we further analyze how information flows across transformer layers. As shown in Fig. 3, we can observe that as representations propagate from shallow to deep layers, cross-modal interactions undergo multiple stages of refinement, leading to substantial information exchange and reorganization. In particular, in shallow layers, removing modality tokens has a large influence on the model predictions, while in deep layers, removing these tokens almost has no influence on predictions. These observations indicate that early-

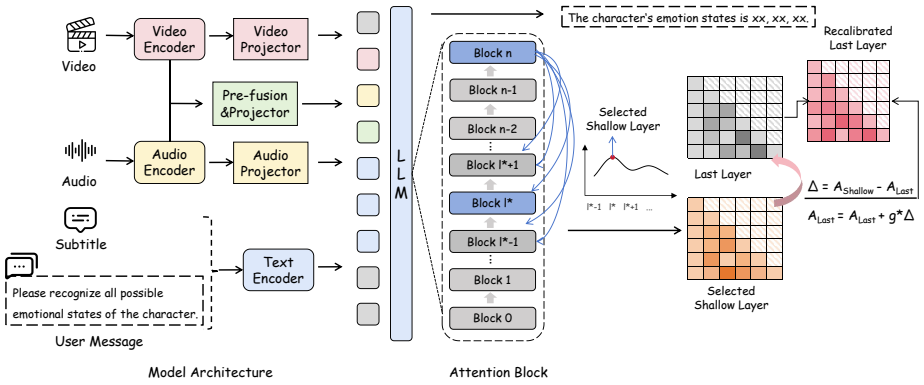

Fig. 4: Overview of our method. Our framework includes: (1) a shallow-layer attention extraction module that captures early-stage structural cues, and (2) an attention refinement module that adaptively integrates this guidance into the last-layer attention map.

layer representations tend to preserve relatively structured attention patterns on modality tokens. In contrast, as information flows into deeper layers, cross-modal interactions become progressively more complex, leading to increasingly entangled attention distributions. In other words, information is mixed in tokens, confusing the contribution of modality tokens, which makes it hard for the model to assign correct attention to these tokens, thereby giving rise to the contradictory reallocation phenomenon.

Motivated by this insight, we seek to leverage the inherent structural properties of shallow-layer attention to guide and refine the attention distribution in the final layer, aiming to mitigate accumulated distortions and enhance the stability of the decoding process.

3.5 Conclusion

In summary, our analysis reveals a modality-dependent contradiction in attention reallocation and highlights the limitations of relying solely on attention weights to measure token informativeness. We further show that deeper layers exhibit increasingly entangled cross-modal interactions, which can distort attention allocation in complex multimodal settings, leading to the contradictory reallocation phenomenon. These findings offer a structural perspective on attention behavior and motivate a shallow-guided refinement strategy for more stable multimodal decoding.

4 Method

Based on the above analysis, we aim to rectify the attention allocation influenced by increasingly entangled cross-modal interactions in deeper layers. As revealed

by our layer-wise study, deeper representations exhibit complex and entangled attention patterns, whereas early-layer attention remains relatively structured and closer to the initial token-level signals. Inspired by this observation, we decide to incorporate shallow-layer attention as a structural prior to regularize and refine the final-layer attention distribution, thereby improving the stability of multimodal decoding. Our method consists of two complementary modules: (1) a shallow-layer attention extraction module that captures early-stage structural cues, and (2) an attention refinement module that adaptively integrates this guidance into the last-layer attention map, as illustrated in Fig. 4.

4.1 Shallow Layer Extraction

To identify the intermediate layer that captures the most informative early-stage structural cues, we introduce a selection mechanism grounded in *Attentional Trajectory Divergence*. Through empirical analysis, we observe that as the input propagates through the Transformer, attention maps gradually shift from capturing fine-grained multimodal interactions to high-level task-aligned representations. This semantic transition often leads to an *information bottleneck*, where subtle but informative signals in the multimodal tokens are progressively attenuated.

To formalize this, we define the candidate set of layers within the Multimodal Formative Phase according to Fig. 3, i.e., the shallow range where visual, acoustic, and textual features first interact, as $\mathcal{L} = \{l \mid l \in [l_s, l_e)\}$. Our goal is to identify a premature layer l^* that preserves the maximal volume of these vanishing multimodal signals.

Let $A^{(L)}$ denote the attention distribution of the final layer, and $A^{(l)}$ the attention distribution of an intermediate layer l. The discrepancy between these layers is quantified using the Jensen–Shannon divergence ($\mathcal{D}_{JS}$):

$$\mathcal{D}_{JS}(A^{(L)} \,\|\, A^{(l)}) = \frac{1}{2}\mathrm{KL}(A^{(L)} \,\|\, M) + \frac{1}{2}\mathrm{KL}(A^{(l)} \,\|\, M), \tag{9}$$

where $M = \frac{1}{2}(A^{(L)} + A^{(l)})$ represents the latent consensus distribution, and $\mathrm{KL}(\cdot\|\cdot)$ denotes the Kullback–Leibler divergence [18].

The optimal premature layer l^* is then selected as

$$l^* = \arg\max_{l \in \mathcal{L}} \mathcal{D}_{JS}(A^{(L)} \,\|\, A^{(l)}), \tag{10}$$

ensuring that l^* captures the largest attentional variance suppressed by deeper layers, and thus retains critical multimodal signals for downstream reasoning.

4.2 Attention Refinement

Upon identifying the premature layer l^*, we introduce an Attention Refinement mechanism to restore multimodal cues that are attenuated in the final-layer attention. We define the *Attentional Residual Map* ΔA as the discrepancy between the attention distributions of the selected intermediate layer and the final layer:

$$\Delta A = A^{(l^*)} - A^{(L)}. \tag{11}$$

In this formulation, ΔA acts as an *Attentional Gradient Signal*, highlighting the contributions of modality-specific tokens that are underrepresented in deeper layers. To reintegrate these signals without perturbing the pretrained representations, the final-layer attention is recalibrated by first injecting the residual signal:

$$\tilde{A}^{(L)} = A^{(L)} + g \cdot \Delta A. \tag{12}$$

To maintain a valid attention distribution, we then apply a normalization operator $\Psi(\cdot)$:

$$\tilde{A}^{(L)} \leftarrow \Psi(\tilde{A}^{(L)}), \tag{13}$$

where $\Psi(\cdot)$ denotes a normalization operator that ensures each row of the attention matrix sums to one. By performing this residual injection followed by normalization, we effectively restore the attenuated multimodal signals while preserving the probabilistic structure of the attention map, ensuring that the final token predictions reflect both high-level semantic reasoning and the relatively clean and well-structured attention patterns retained in early layers. The refined attention matrix is integrated into the final-layer self-attention computation of the decoder (as described in Sec. 3.1), and consequently affects the next-token prediction process.

5 Experiments

5.1 Tasks and Datasets

For performance comparison, we evaluate the model on MER-UniBench [24], assessing its capabilities in open-vocabulary and generalized emotion understanding. MER-UniBench covers three representative tasks: fine-grained emotion recognition, basic emotion recognition, and sentiment analysis. For fine-grained emotion recognition, we adopt OV-MERD+ [24], which extends emotion prediction beyond basic categories. For basic emotion recognition, we evaluate on four widely used datasets, including MER2023 [25], MER2024 [27], IEMOCAP [1], and MELD [35]. For sentiment analysis, which focuses on polarity prediction, we report results on CMU-MOSI [54], CMU-MOSEI [55], CH-SIMS [53], and CH-SIMS v2 [32]. Additional details are provided in Appendix.

5.2 Evaluation Metrics

For emotion reasoning and understanding, following AffectGPT [24], we adopt Qwen3 [50] to extract emotion-related keywords from the final conclusions of the generated explanations. These keywords are then grouped and aligned with the ground-truth annotations for evaluation. For fine-grained emotion recognition, evaluation is conducted using Precision, Recall, and F-score, with F-score serving as the primary metric [26]. For basic emotion recognition, we adopt Hit

Table 2: Performance on MER-UniBench. This table presents the results for the primary metrics, with other metrics listed in the Appendix. "MOSI", "MOSEI", "SIMS", and "SIMS v2" refer to CMU-MOSI, CMU-MOSEI, CH-SIMS, and CH-SIMS v2, respectively. The last column shows the average score across all datasets.

Methods	Basic				Sentiment				Fine-grained	Avg
	MER2023	MER2024	MELD	IEMOCAP	MOSI	MOSEI	SIMS	SIMS v2	OV-MERD+	
Input Modality: Audio, Text										
OneLLM [9]	25.52	17.21	28.32	33.44	64.01	54.09	63.39	61.98	22.25	41.14
SECap [48]	40.95	52.46	25.56	36.92	55.76	54.18	59.51	57.41	36.97	46.64
PandaGPT [40]	33.57	39.04	31.91	36.55	66.06	61.33	62.93	58.88	31.33	46.84
Qwen-Audio [4]	41.85	31.61	49.00	35.47	70.09	46.90	70.73	65.26	32.36	49.26
SALMONN [41]	55.53	45.38	45.62	46.84	81.00	67.03	68.69	65.93	45.00	57.89
AffectGPT [24]	72.94	73.41	56.63	55.68	**83.46**	**80.74**	82.99	83.75	59.98	72.18
Ours	**75.78**	**80.49**	**62.74**	**63.06**	81.79	80.34	**87.48**	**84.43**	**62.33**	**75.38**
Input Modality: Video, Text										
Otter [20]	16.41	14.65	22.57	29.08	52.89	50.44	57.56	53.12	16.63	34.82
Video-LLaVA [28]	36.93	30.25	30.73	38.95	56.37	61.64	53.28	57.45	34.00	44.40
PandaGPT [40]	39.13	47.16	38.33	47.21	58.50	64.25	62.07	65.25	35.07	50.77
Video-ChatGPT [33]	44.86	46.80	37.33	56.83	54.42	63.12	64.82	65.80	39.80	52.64
VideoChat2 [22]	33.67	54.50	36.64	48.70	66.84	54.32	69.49	70.66	39.21	52.67
LLaMA-VID [23]	50.72	57.60	42.75	46.02	61.78	63.89	69.35	67.48	45.01	56.07
VideoChat [21]	48.73	57.30	41.11	48.38	65.13	63.61	69.52	72.14	44.52	56.71
Chat-UniVi [17]	57.62	65.67	45.61	52.37	54.53	63.18	68.15	66.36	48.00	57.94
mPLUG-Owl [52]	56.86	59.89	49.11	55.54	72.40	72.91	72.13	75.00	48.18	62.45
AffectGPT [24]	74.58	75.29	57.63	62.19	**82.39**	**81.57**	87.20	86.29	61.65	74.31
Ours	**79.48**	**82.56**	**62.06**	**64.25**	82.27	80.27	**90.72**	**87.71**	**63.52**	**76.98**
Input Modality: Video, Audio, Text										
PandaGPT [40]	40.21	51.89	37.88	44.04	61.92	67.61	68.38	67.23	37.12	52.92
R1-Omni [59]	64.17	67.43	43.20	51.58	58.02	56.48	71.82	68.58	55.24	59.61
Emotion-LLaMA [3]	59.38	73.62	46.76	55.47	66.13	67.66	78.32	77.23	52.97	64.17
AffectGPT [24]	78.54	78.80	55.65	60.54	81.30	**80.90**	88.49	86.18	62.52	74.77
Ours	**84.49**	**83.79**	**62.25**	**65.69**	**83.31**	79.07	**89.77**	**89.04**	**65.32**	**78.08**

Rate as the evaluation criterion to measure whether the predicted emotional category successfully captures the ground-truth label. We report Accuracy (ACC) together with Weighted Average F-score (WAF), prioritizing WAF as the main metric [54, 55]. Comprehensive descriptions of the evaluation procedures and metric formulations are presented in the Appendix.

5.3 Implementation Details

Our implementation follows the configuration of AffectGPT [24]. We adopt CLIP ViT-L [36] as the visual encoder and HuBERT-L [11] as the acoustic encoder. We employ Qwen3 [50] as the backbone LLM and retrain the model under the same multimodal framework. For efficiency, we fine-tune only the additional LoRA modules within the LLM, the multimodal projector, and the pre-fusion branch, while keeping the parameters of the unimodal encoders frozen. The model is optimized for up to 60 epochs, and each epoch comprises 5,000 training steps with a batch size of 3. We employ the AdamW optimizer and use a learning rate of 1e-5. More implementation details are provided in the Appendix.

5.4 Comparison with State-of-the-Art Methods

As shown in Tab. 2, our method consistently outperforms existing state-of-the-art approaches across all modality configurations on MER-UniBench. Under au-

Table 3: Ablation Study of the proposed method under three-modality settings.

Methods	Fine-grained	Basic	Sentiment	MER-UniBench
Baseline	64.5	67.07	84.06	74.34
+ Shallow Layer Guidance	65.23	71.19	84.60	76.48
+ Residual Refinement	**65.32**	**74.06**	**85.30**	**78.08**
(a) Shallow Layer Replacement	65.26	73.64	84.81	77.68
(b) Uniform Prior Injection	65.27	72.50	85.04	77.27
(c) Layer-wise Residual Aggregation	65.28	70.05	84.65	76.01

dio–text, video–text, and video-audio–text settings, we achieve the best overall average performance, demonstrating strong robustness and cross-modal generalization ability. Compared with prior methods, especially AffectGPT [24], our model achieves a 4.4% improvement in overall average performance, showing consistent gains in basic emotion recognition while maintaining competitive results in sentiment analysis and fine-grained emotion recognition.

5.5 Ablation Studies

Effectiveness of each design To evaluate the contribution of each component in our framework, we conduct ablation studies under the three-modality setting. As shown in Tab. 3, introducing shallow layer guidance consistently improves performance over the baseline across all evaluation metrics, demonstrating the effectiveness of incorporating early-layer structural cues as a prior for attention refinement. Furthermore, adding the residual refinement module yields additional and substantial improvements, achieving the best results on both sub-tasks and the overall MER-UniBench benchmark. These results indicate that the combination of shallow guidance and residual-based attention refinement plays a critical role in stabilizing attention allocation and enhancing multimodal reasoning under complex modality interactions.

Comparison with alternative refinement strategies To further evaluate the effectiveness of our design, we compare the proposed method with several alternative strategies that modify the final-layer attention(Tab. 3). *Shallow Layer Replacement* substitutes the final-layer attention with the shallow-layer map. While it improves over the baseline, performance is still slightly lower than our method, indicating that direct replacement may disturb deeper-layer semantics. *Uniform Prior Injection* introduces a uniform distribution into the final-layer attention as a regularization signal. Although it slightly improves performance by mitigating attention concentration, the lack of modality-aware structure limits its effectiveness. *Layer-wise Residual Aggregation* combines attention from multiple shallow layers with weighted aggregation. The results show that it performs worse than using a single selected shallow layer, indicating that simply aggregating multiple layers is less effective. In contrast, our residual refinement selectively injects shallow-layer signals while preserving the structure of the final-layer at-

Strategies	Fine-grained	Basic Sentiment		MER-UniBench
Baseline	64.5	67.07	84.06	74.34
Length-Constrained Decoding	63.9	66.03	79.26	71.67
Layer-Guided Logit Refinement	64.82	67.07	83.98	74.33
Ours	**65.32**	**74.06**	**85.30**	**78.08**
Baseline w/o fused token	60.77	67.10	83.03	73.47
Ours w/o fused token	**61.94**	**69.36**	**83.48**	**74.81**

Table 4: Comparison of different strategies.

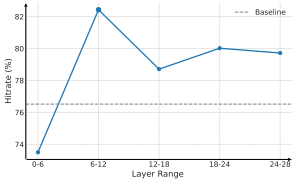

Fig. 5: Effect of layer selection.

tention. It achieves the best performance across all metrics, demonstrating its superiority over naive replacement or aggregation strategies.

5.6 Analysis and Discussion

Improvements Arise from Attention Refinement Rather Than Longer Outputs As our method produces more accurate and fine-grained emotional expressions, motivating us to investigate whether the improvement is merely due to generating longer outputs, which could superficially increase matches with ground-truth labels. We conducted an experiment by constraining the baseline to produce longer responses, and the results in Tab. 4 show no performance gain, demonstrating that the improvements stem from the refined attention mechanism rather than output length.

Comparison with Logit-Level Decoding We compare our method with DeCo [45], which adjusts final-layer logits by integrating preceding layers knowledge. As shown in Tab. 4, this logit-level adjustment does not perform well in our setting. In contrast, our method operates directly at the attention level, refining the attention map before the final prediction is formed. By rectifying misallocated attention earlier in the reasoning process, our approach better preserves modality-relevant information and leads to more accurate predictions, resulting in consistent performance improvements.

Layer Selection Analysis We analyze the impact of layer selection, which further aligns with our previous observations. The results in Fig. 5 indicate that applying the method to the earliest layers is suboptimal, as the model has not yet learned sufficiently informative representations. In contrast, selecting intermediate layers yields better performance, likely because these layers capture the first significant cross-modal interactions and information exchange. This observation is consistent with our analysis of attention dynamics, highlighting the importance of operating at appropriate depth levels.

Effect of Removing the Pre-fusion Module To exclude the influence of AffectGPT's specific architectural design, we remove its pre-fusion module and conduct experiments using only the three original modalities (audio, video, and text). As shown in Tab. 4, our method still yields consistent improvements over

Table 5: Comparison with existing attention adjustment methods.

Methods	Fine-grained	Basic Sentiment		MER-Unibench
Baseline	64.50	67.07	84.06	74.34
DoLA [5]	61.94	57.95	80.42	68.38
OPERA [12]	64.93	69.21	84.68	75.61
Entropy-based [38]	63.89	67.95	83.44	74.38
Average-based [34]	65.27	67.24	84.32	74.61
Training-based [58]	62.04	68.44	83.24	74.31
Ours	**65.32**	**74.06**	**85.30**	**78.08**

the baseline. This result suggests that the performance gains mainly come from the proposed method rather than the model-specific design.

Comparison with attention adjustment methods We achieve the best performance in Tab. 5. We differ from them by leveraging shallow-layer attention as a structural prior motivated by the harmful attentional interaction in multimodal deep layers, which is found for the first time.

6 Conclusion

In this work, we revisit attention reallocation for multimodal emotion recognition under the generative paradigm of MLLMs. We find that reallocating attention from generated tokens to modality tokens improves performance in two-modality settings but degrades it in three- or more-modality scenarios, where the opposite direction is more effective. We attribute this to increasingly entangled attention in deeper layers as modalities increase. To address this issue, we propose a training-free attention rectification method that leverages shallow-layer attention to regularize final-layer attention during inference. Extensive experiments demonstrate that our method achieves state-of-the-art performance across diverse multimodal settings.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under grants 62206102; the National Key Research and Development Program of China under grant 2024YFC3307900; the National Natural Science Foundation of China under grants 62436003, 62376103 and 62302184; Major Science and Technology Project of Hubei Province under grant 2025BAB011 and 2024BAA008; Hubei Science and Technology Talent Service Project under grant 2024DJC078; Ant Group through CCF-Ant Research Fund; the Postdoctoral Fellowship Program (Grade C) of China Postdoctoral Science Foundation under grant GZC20252271; the China Postdoctoral Science Foundation under grant 2025M773444; and the Outstanding Seed Cultivation Project of Central China Normal University under grant CCNU25XJ031. The computation is completed in the HPC Platform of Huazhong University of Science and Technology.

References

1. Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S.: Iemocap: Interactive emotional dyadic motion capture database. Language resources and evaluation **42**(4), 335–359 (2008)
2. Cao, M., Wan, Z.: Retracted: Psychological counseling and character analysis algorithm based on image emotion. IEEE Access (2020)
3. Cheng, Z., Cheng, Z.Q., He, J.Y., Wang, K., Lin, Y., Lian, Z., Peng, X., Hauptmann, A.: Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. Advances in Neural Information Processing Systems **37**, 110805–110853 (2024)
4. Chu, Y., Xu, J., Zhou, X., Yang, Q., Zhang, S., Yan, Z., Zhou, C., Zhou, J.: Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. arXiv preprint arXiv:2311.07919 (2023)
5. Chuang, Y.S., Xie, Y., Luo, H., Kim, Y., Glass, J.R., He, P.: Dola: Decoding by contrasting layers improves factuality in large language models. In: International Conference on Learning Representations. vol. 2024, pp. 54158–54183 (2024)
6. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in neural information processing systems **36**, 49250–49267 (2023)
7. Fan, W., Xu, X., Xing, X., Chen, W., Huang, D.: Lssed: a large-scale dataset and benchmark for speech emotion recognition. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 641–645. IEEE (2021)
8. Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva: Exploring the limits of masked visual representation learning at scale. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19358–19369 (2023)
9. Han, J., Gong, K., Zhang, Y., Wang, J., Zhang, K., Lin, D., Qiao, Y., Gao, P., Yue, X.: Onellm: One framework to align all modalities with language. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26584–26595 (2024)
10. Han, Z., Zhu, B., Xu, Y., Song, P., Yang, X.: Benchmarking and bridging emotion conflicts for multimodal emotion reasoning. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 5528–5537 (2025)
11. Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM transactions on audio, speech, and language processing **29**, 3451–3460 (2021)
12. Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., Lin, D., Zhang, W., Yu, N.: Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13418–13427 (2024)
13. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
14. Hutchison, A., Gerstein, L.: Emotion recognition, emotion expression, and cultural display rules: Implications for counseling. Journal of Asia Pacific Counseling **7**(1) (2017)

15. Imani, M., Montazer, G.A.: A survey of emotion recognition methods with emphasis on e-learning environments. Journal of network and computer applications **147**, 102423 (2019)
16. Jiang, X., Zong, Y., Zheng, W., Tang, C., Xia, W., Lu, C., Liu, J.: Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In: Proceedings of the 28th ACM international conference on multimedia. pp. 2881–2889 (2020)
17. Jin, P., Takanobu, R., Zhang, W., Cao, X., Yuan, L.: Chat-univi: Unified visual representation empowers large language models with image and video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13700–13710 (2024)
18. Kullback, S., Leibler, R.A.: On information and sufficiency. The annals of mathematical statistics **22**(1), 79–86 (1951)
19. Le Ngwe, J., Lim, K.M., Lee, C.P., Ong, T.S., Alqahtani, A.: Patt-lite: Lightweight patch and attention mobilenet for challenging facial expression recognition. IEEE Access **12**, 79327–79341 (2024)
20. Li, B., Zhang, Y., Chen, L., Wang, J., Pu, F., Cahyono, J.A., Yang, J., Li, C., Liu, Z.: Otter: A multi-modal model with in-context instruction tuning. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
21. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023)
22. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al.: Mvbench: A comprehensive multi-modal video understanding benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22195–22206 (2024)
23. Li, Y., Wang, C., Jia, J.: Llama-vid: An image is worth 2 tokens in large language models. In: European Conference on Computer Vision. pp. 323–340. Springer (2024)
24. Lian, Z., Chen, H., Chen, L., Sun, H., Sun, L., Ren, Y., Cheng, Z., Liu, B., Liu, R., Peng, X., et al.: Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. In: Forty-second International Conference on Machine Learning
25. Lian, Z., Sun, H., Sun, L., Chen, K., Xu, M., Wang, K., Xu, K., He, Y., Li, Y., Zhao, J., et al.: Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. In: Proceedings of the 31st ACM international conference on multimedia. pp. 9610–9614 (2023)
26. Lian, Z., Sun, H., Sun, L., Chen, L., Chen, H., Gu, H., Wen, Z., Chen, S., Siyuan, Z., Yao, H., et al.: Open-vocabulary multimodal emotion recognition: Dataset, metric, and benchmark (2024)
27. Lian, Z., Sun, H., Sun, L., Wen, Z., Zhang, S., Chen, S., Gu, H., Zhao, J., Ma, Z., Chen, X., et al.: Mer 2024: Semi-supervised learning, noise robustness, and open-vocabulary multimodal emotion recognition. In: Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing. pp. 41–48 (2024)
28. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Proceedings of the 2024 conference on empirical methods in natural language processing. pp. 5971–5984 (2024)

29. Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al.: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024)
30. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
31. Liu, S., Zheng, K., Chen, W.: Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In: European Conference on Computer Vision. pp. 125–140. Springer (2024)
32. Liu, Y., Yuan, Z., Mao, H., Liang, Z., Yang, W., Qiu, Y., Cheng, T., Li, X., Xu, H., Gao, K.: Make acoustic and visual cues matter: Ch-sims v2. 0 dataset and av-mixup consistent module. In: Proceedings of the 2022 international conference on multimodal interaction. pp. 247–258 (2022)
33. Maaz, M., Rasheed, H., Khan, S., Khan, F.: Video-chatgpt: Towards detailed video understanding via large vision and language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12585–12602 (2024)
34. McDanel, B., Li, S., Khaitan, H.: Claa: Cross-layer attention aggregation for accelerating llm prefill. arXiv preprint arXiv:2602.16054 (2026)
35. Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., Mihalcea, R.: Meld: A multimodal multi-party dataset for emotion recognition in conversations. In: Proceedings of the 57th annual meeting of the association for computational linguistics. pp. 527–536 (2019)
36. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
37. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
38. Song, J., Luo, J., Tang, X.s., Hao, K., Zhao, M.: Segmentation-based attention entropy: Detecting and mitigating object hallucinations in large vision-language models. arXiv preprint arXiv:2603.16558 (2026)
39. Song, P., Guo, D., Yang, X., Tang, S., Wang, M.: Emotional video captioning with vision-based emotion interpretation network. IEEE Transactions on Image Processing **33**, 1122–1135 (2024)
40. Su, Y., Lan, T., Li, H., Xu, J., Wang, Y., Cai, D.: Pandagpt: One model to instruction-follow them all. In: Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants! pp. 11–23 (2023)
41. Tang, C., Yu, W., Sun, G., Chen, X., Tan, T., Li, W., Lu, L., MA, Z., Zhang, C.: Salmonn: Towards generic hearing abilities for large language models. In: The Twelfth International Conference on Learning Representations
42. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
43. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
44. Tu, C., Ye, P., Zhou, D., Bai, L., Yu, G., Chen, T., Ouyang, W.: Attention reallocation: Towards zero-cost and controllable hallucination mitigation of mllms. International Journal of Computer Vision **134**(1), 22 (2026)

45. Wang, C., Chen, X., Zhang, N., Tian, B., Xu, H., Deng, S., Chen, H.: Mllm can see? dynamic correction decoding for hallucination mitigation. In: The Thirteenth International Conference on Learning Representations
46. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
47. Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., et al.: Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. Advances in Neural Information Processing Systems **36**, 61501–61513 (2023)
48. Xu, Y., Chen, H., Yu, J., Huang, Q., Wu, Z., Zhang, S.X., Li, G., Luo, Y., Gu, R.: Secap: Speech emotion captioning with large language model. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 19323–19331 (2024)
49. Yan, Q., Chen, G., Zou, Y.: Start small, think big: Curriculum-based relative policy optimization for visual grounding. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 11550–11558 (2026)
50. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
51. Yang, Q., Bai, D., Peng, Y.X., Wei, X.: Omni-emotion: Extending video mllm with detailed face and audio modeling for multimodal emotion analysis. arXiv preprint arXiv:2501.09502 (2025)
52. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al.: mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178 (2023)
53. Yu, W., Xu, H., Meng, F., Zhu, Y., Ma, Y., Wu, J., Zou, J., Yang, K.: Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In: Proceedings of the 58th annual meeting of the association for computational linguistics. pp. 3718–3727 (2020)
54. Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.P.: Tensor fusion network for multimodal sentiment analysis. In: Proceedings of the 2017 conference on empirical methods in natural language processing. pp. 1103–1114 (2017)
55. Zadeh, A.B., Liang, P.P., Poria, S., Cambria, E., Morency, L.P.: Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2236–2246 (2018)
56. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022)
57. Zhang, Z., Yadav, S., Han, F., Shutova, E.: Cross-modal information flow in multimodal large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19781–19791 (2025)
58. Zhao, H., Si, S., Chen, L., Zhang, Y., Sun, M., Chang, B., Zhang, M.: Looking beyond text: Reducing language bias in large vision-language models via multimodal dual-attention and soft-image guidance. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 19677–19701 (2025)
59. Zhao, J., Wei, X., Bo, L.: R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning. arXiv preprint arXiv:2503.05379 (2025)