Zero-MELO: Test-Time Evidence Calibration with Multimodal LLMs for Zero-Shot Micro-Gesture Recognition

Anonymous Author(s)
Submission Id: 6240

Abstract

While Multi-modal Large Language Models (MLLMs) excel in general video understanding, their capability in fine-grained and motion-centric tasks remains limited. This limitation is particularly critical in micro-gesture recognition (MGR), where **micro-gestures** (MG)—subtle, short-duration, and spatially localized human movements—serve as key discriminative signals for implicit affective analysis, yet are easily neglected following common prompting practices. Although MGR has been intensively studied by many discriminative approaches, the use of MLLMs on MGR is underexplored, with notably poor performance. We hypothesize that the motion-sensitive representation ability of MLLMs is constrained by its inherent single-pass forward inference, which can be substantially enhanced through carefully designed test-time guidance. Motivated by this, building on our prior findings regarding temporal insensitivity in Video LLMs, we diagnose zero-shot MGR errors in the Negative Log-Likelihood (NLL) space. We observe that MLLMs suffer from two bottlenecks: **1) insufficient localized evidence** and **2) severe score biases** driven by language and motion-agnostic appearances. Thus, we propose a novel test-time evidence calibration framework that improves both reasoning details and prediction reliability. Specifically, we introduce a tree search mechanism to progressively acquire localized, fine-grained visual evidence, coupled with a test-time calibration module to mitigate score biases. The multi-cue fusion module then integrates evidence from multiple cues without relying on a single cue for final prediction. Our framework achieves a mean-class accuracy of 26.84% on iMiGUE and 22.10% on MA52, significantly outperforming the Qwen2.5-VL baseline that produces 16.15% and 10.20%, respectively. The code will be available at https://zero-melo.github.io/Zero-MELO.

CCS Concepts

• **Computing methodologies** → **Computer vision**.

Keywords

Gesture Recognition, Micro-Gesture Recognition, MLLM, LLM

1 Introduction

Micro-gestures (MG) are subtle, short-duration, and spatially localized human movements that serve as key indicators of underlying behavioral and affective states [3], which has drawn increasing research attention in recent years. Compared with general human activities [23, 31, 32, 45], MG shows finer granularity in both temporal and spatial domains, making its reliable recognition difficult [15]. The field has progressed rapidly with benchmarks of increasing scale and coverage, including iMiGUE [35], SMG [9], BBSI [6], MA-52 [17], and MMA-52 [27], and supervised models have achieved strong in-domain performances [15, 16, 26, 28].

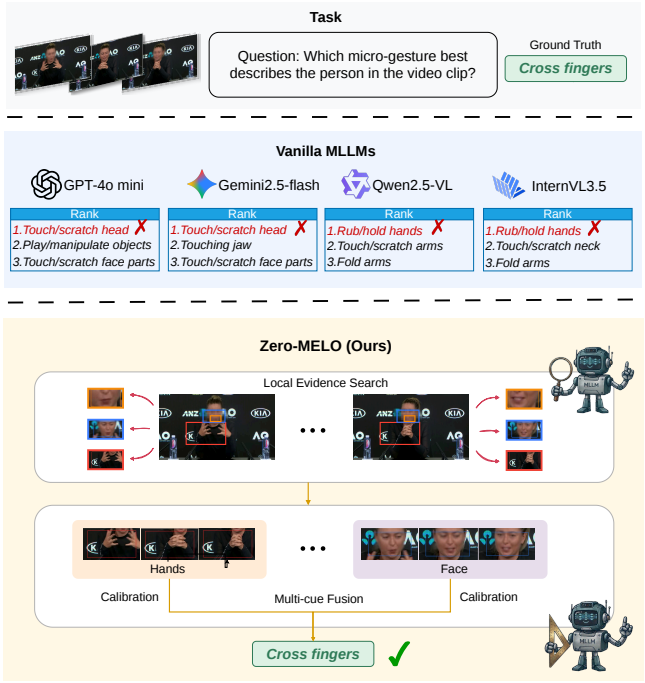

Figure 1: Comparison between existing methods and Zero-MELO. Given the input video clips and question, the target is to predict the correct label in MGR. Compared to the single-pass forward inference of other MLLMs, Zero-MELO can utilize the relevant local evidence and predict the correct label via calibration at inference time.

Nevertheless, micro-gesture recognition (MGR) remains fundamentally challenging due to several intrinsic factors. First, MGs exhibit substantial cross-subject variability, where subtle motion patterns are heavily entangled with individual-specific appearance and motion styles. Second, MG categories typically follow a long-tailed distribution, where rare and fine-grained gestures lack sufficient training samples, leading to biased decision boundaries in supervised settings. Third, real-world MG scenarios continuously introduce **unseen** or **novel** gesture semantics, which cannot be exhaustively covered by predefined label spaces. As a result, conventional fully supervised MGR paradigms, which rely on fixed label sets and sufficient annotated data, struggle to scale to realistic settings. These limitations naturally motivate **zero-shot MGR**, where models are required to recognize fine-grained and potentially novel micro-gestures without additional task-specific training.

Meanwhile, large language models (LLMs) and MLLMs have demonstrated great zero-shot and few-shot capability across a wide range of tasks [4, 7, 20, 59], including both general-purpose [36] and

specific downstream tasks [13, 63, 74], which naturally motivates the zero-shot MGR. Despite the broad transfer capability shown in LLMs and MLLMs, their capability on fine-grained motion-centric video understanding remains limited [11]. In our preliminary study, we also observe that *directly applying MLLMs* on MGR leads to poor performance on both supervised and zero-shot settings (see more in Sec. 4.2).

Thus, in this work, we revisit this discrepancy and ask a fundamental question: **Why do MLLMs struggle to capture subtle, motion-critical cues in micro-gesture recognition under standard prompting settings?** To better understand this problem, we start from standard prompt-based free-form generation, and further analyze MLLM predictions from both free-form generation and Negative Log-Likelihood (NLL)-based label scoring that ranks candidate labels by NLL and selects the minimum-NLL label. Our prior qualitative studies have argued that MLLMs can recognize obvious micro-gestures on iMiGUE but often fail on subtle ones (e.g., when the movement is extremely subtle), suggesting that common general prompting is insufficient for preserving decisive local motion cues in zero-shot MGR [37]. Furthermore, the NLL-based analysis quantitatively reveals a second issue underlying why MLLMs fail in MGR, namely a strong prior-driven bias in general MLLMs. Specifically, we find that **when visual input is removed, the model continues to produce confident predictions, revealing a language prior bias**, indicating its tendency to rely on semantic expectations. This is consistent with existing research findings [29, 49]. Conversely, when using static (rank-1) visual cues that preserve appearance but remove motion, the model exhibits an appearance bias, which is globally weaker but can be locally stronger for certain classes. These two biases heavily affect the behavior of MLLMs, especially under long-tailed label distributions. Based on this analysis, we show that reliable zero-shot MGR requires both better **extraction of localized motion evidence** and **bias-aware score/evidence calibration**.

Motivated by these insights, we propose to address the above bottlenecks by expanding *test-time* reasoning to unlock the latent capability of MLLMs for zero-shot MGR. Specifically, instead of treating the model's first-pass response as final, we introduce a more elaborate inference procedure that searches for localized motion evidence, aggregates clues across spatial-temporal regions rather than relying on a single cue, and recalibrates predictions against dominant semantic priors. Thus, we propose **Zero-MELO**, a training-free MLLM framework for zero-shot MGR via test-time refinement. It incorporates a cue-guided tree search mechanism to acquire the zoomed local evidence from informative body cues, a dual test-time calibration module to mitigate the prior-driven score bias, and a cue-conditioned evidence fusion module to aggregate NLL scores across different cues to predict the final label. The whole framework comparison with conventional MLLMs is shown in Fig. 1. Under the Qwen2.5-VL [5] backbone, our framework improves mean-class accuracy from 16.15% to 26.84% on iMiGUE and from 10.20% to 22.10% on MA-52.

In summary, our contributions can be concluded as follows:

(1) We revisit the failure of MLLMs in micro-gesture recognition and show that the bottleneck is not solely due to insufficient representation ability of MLLMs, but is largely caused by insufficient localized evidence and severe score biases driven by language and motion-agnostic appearances.

(2) We introduce the idea of expanding test-time reasoning in MLLMs for zero-shot MGR, demonstrating that enhancing evidence exploration and aggregation at inference time can effectively reveal the latent capability of MLLMs.

(3) We propose Zero-MELO, a training-free MLLM framework that couples cue-shared zoom search with dual-reference calibration to address missing local evidence and prior-driven score bias in zero-shot MGR with a predefined label list, coupled with a cue-conditioned evidence fusion strategy, enabling robust multi-cue recognition that avoids single-branch decisions under shared context.

(4) Extensive experiments show that Zero-MELO consistently improves over standard MLLM inference, suggesting that the limitations of MLLMs on micro-gesture recognition are highly related to the insufficient localized evidence and severe biases.

2 Related Work

Zero-shot Micro-Gesture Recognition. Existing MGR methods are predominantly supervised and learn dataset-specific spatiotemporal representations from RGB clips [12, 71], skeleton [55, 65, 79] sequences, cross-modality fusion [16, 41, 56] or the joint modeling of spatial structure and temporal dynamics [15, 28, 33, 64, 78]. Despite this progress, these methods typically depend on labeled training data and are not directly applicable in a zero-shot setting.

Existing MLLM studies on MGR-related datasets mainly focus on affective understanding or semantic analysis, rather than action-level MGR [14, 25, 46]. Moreover, previous qualitative evidence suggests that global-view prompting can recognize obvious MG yet often fails on subtle ones in iMiGUE [37]. These observations indicate that directly applying a generic video MLLM to MGR is insufficient, but they do not provide a quantitative diagnosis of why zero-shot MGR fails. Different from these works, we study the predefined-set zero-shot MGR based on MLLM label-scoring formulation, and quantitatively analyze its two core bottlenecks: insufficient localized evidence and reference-conditioned bias in label scores.

Test-Time Inference in MLLMs. Test-time inference in LLMs has become an important research direction [38, 50, 67]. A major line of recent work addresses hallucination and prior-driven errors in LLMs and MLLMs at test time. Prior work mainly mitigates hallucination and prior-driven errors through contrastive decoding [24, 29, 39, 49], attention-space or hidden-state interventions [1, 22, 34, 51, 62, 66, 68, 80], dynamic decoding [53], leveraging external generation models [8]. Compared to these image-based methods mentioned above, video-specific variants have been explored to suppress temporally insensitive responses and improve temporal reasoning [21, 42, 72].

Although highly relevant, these methods primarily intervene in token-level decoding or hidden-state dynamics in open-ended generation settings. In contrast, we consider the MGR with a predefined label set. Accordingly, rather than intervening in generation-time decoding, we reformulate the idea as bias calibration on label NLLs, where the calibration procedure directly adjusts label scores rather

than generation-time token logits. This ensures all calibrated predictions remain valid and allows the correction signal to be naturally integrated with multi-cue evidence fusion.

MLLMs with Fine-grained Visual Search. Visual encoders trained through contrastive learning, such as CLIP [43] and SigLip [70], often discard fine-grained semantics [2, 43, 70]. To address this issue, prior work augments perception with local crops [58, 63], search-based zooming [47, 69], model-internal focusing mechanisms [48, 73, 77], retrieval-assisted perception [60], or training-based active perception [19, 57, 61, 75, 76]. These methods broadly follow a "thinking with images" paradigm [52].

Our setting is fundamentally different due to the video-based and predefined label list nature of MGR. Accordingly, the goal is not to find arbitrary question-relevant regions, but to acquire discriminative local evidence for classification. Moreover, unlike instance-specific search targets, our search is cue-shared across labels, allowing localized evidence to be reused across multiple actions. Besides, existing methods typically use one selected crop as auxiliary evidence. In contrast, zero-shot MGR does not provide an oracle cue prior: the challenge is not only localized search, but also cross-cue decision under branch-specific score distortion and shared global context.

3 Methodology

3.1 Problem Formulation

The zero-shot MGR with an MLLM is defined as follows: given an original video clip g and a prompt t, the model predicts a corresponding MG label $y \in \mathcal{Y}$,

$$y^* = \arg\max_{y \in \mathcal{Y}} p_{\mathcal{M}}(y \mid g, t), \tag{1}$$

where $p_{\mathcal{M}}(y \mid g, t)$ denotes the label-level prediction for class y under a specific inference scheme, and $p_{\mathcal{M}}$ represents the MLLM with parameters $\mathcal{M}$.

To avoid notation ambiguity, we distinguish three types of visual input. We use g for the original/global video clip, z for a localized cue view derived from g by our method, and x for a generic visual input when a formulation is shared across input types. Accordingly, x can instantiate the global clip g, a localized view z, or a joint input such as (g, z).

Prompt-based free-form generation vs. NLL-based label scoring. An intuitive approach to zero-shot MGR is the common general prompting setting referred to as prompt-based free-form generation in this work, where the MLLM directly produces predictions conditioned on a candidate label list. While this paradigm serves as a competitive baseline following common prompting practices, it *only exposes the final generated output*, without access to internal token-level likelihoods or logits. As a result, it cannot provide a calibrated and comparable score distribution over the full label space. This limitation is particularly restrictive for our framework, which requires label-wise score vectors to diagnose bias and perform subsequent correction. Moreover, free-form generation is inherently misaligned with a predefined label inventory, especially in closed-source LLMs where internal scoring signals are not accessible. See Supplementary Material A (SM A) for misaligned failure cases.

To address these issues, we choose a different technical path that reformulates zero-shot MGR as a classification problem. Specifically, we compute a score for each candidate label by evaluating the teacher-forced negative log-likelihood (NLL) of its verbalized form. This enables consistent and comparable scoring across all labels, while leveraging the model's internal probabilistic signals for more transparent decision-making. Details are presented in Sec. 3.4.

3.2 Overview of Zero-MELO

The overall framework of Zero-MELO is illustrated in Fig. 2. The data flow of Zero-MELO starts from the global input video MG clip and prompt (g, t), where MLLM computes label-wise scores $s_g(y)$ or equivalently NLL $\ell_g(y)$ over the set of labels $\mathcal{Y}$. To address missing fine-grained evidence, we propose a ***1) cue-guided tree search*** that generates local views $\{z_i\}_{i=1}^n$ from semantic cues c_i, forming paired inputs $x_i = (g, z_i)$, each producing scores $s_{x_i}(y)$. To mitigate prior bias, ***2) test-time calibration*** introduces the calibration inputs of x_b (blank) and x_r (low-rank), yielding calibrated NLL $\tilde{\ell}_{x_b}(y)$ and $\tilde{\ell}_{x_r}(y)$, which are further combined into aggregated calibration energies $\bar{\ell}_x(y)$. Finally, ***3) multi-cue fusion*** aggregates global and local evidence into a unified score $\ell_{\text{fuse}}(y)$, and the final prediction is obtained as y_{final}. Below, we introduce the Zero-MELO following the data flow in detail.

3.3 Cue-Guided Tree Search

As revealed by our analytical experiments in Sec. 4.2, we can see that a key failure mode of MLLMs in MGR lies in their improper cue allocation: under common prompting practices, the model often attends to irrelevant regions while neglecting informative, motion-critical cues, leading to insufficient discriminative evidence. To address this, inspired by the general zoom-search idea of some existing works [47, 69, 73], we design a novel technique called *Cue-Guided Tree Search* customized for video-based MGR.

Given an input video, we first select a motion-salient temporal window and uniformly sample T frames from it to form the global clip g. Rather than searching on a single frame, for each cue c_i (e.g., "pay attention to the lips"), we perform cue-guided search on several anchor frames $\mathcal{A} = \{a_1, \ldots, a_m\}$ sampled from the clip:

$$\mathcal{B}_{i,a} = \text{search}(g_a, c_i), \qquad a \in \mathcal{A}, \tag{2}$$

where $\mathcal{B}_{i,a}$ denotes the high-confidence proposals returned on anchor a. This anchor-based design improves temporal stability for transient motions and avoids relying on a single snapshot.

The retained anchor-wise proposals are then aggregated into one temporally stable region,

$$b_i = \text{Agg}\Big(\bigcup_{a \in \mathcal{A}} \hat{\mathcal{B}}_{i,a} \Big), \tag{3}$$

where $\hat{\mathcal{B}}_{i,a} \subseteq \mathcal{B}_{i,a}$ is a small retained set, and $\text{Agg}(\cdot)$ performs robust cross-anchor aggregation by suppressing inconsistent proposals and unioning the remaining ones. If the aggregated region is unreliable or becomes excessively large, we conservatively fall back to the root region. The final cue-conditioned local view is obtained by applying the same crop to all sampled frames:

$$z_i = \text{CropResize}(g, b_i). \tag{4}$$

Compared with generic zooming on a single image, this design adapts cue search to video MGR by combining motion-aware temporal selection, anchor-based localization, and robust cross-anchor

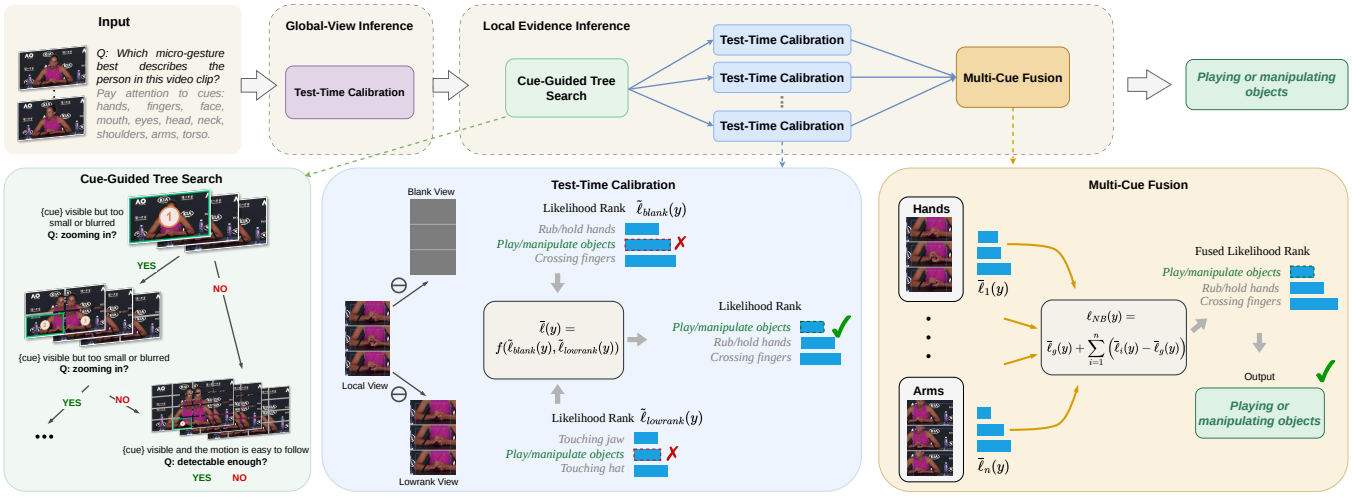

Figure 2: The overall algorithm of Zero-MELO. It mainly consists of three key modules: the cue-guided search (Sec. 3.3), the test-time calibration (Sec. 3.4), and the multi-cue fusion (Sec. 3.5). The global video input will first be fed to the MLLM for the global inference. Then, the model will acquire local evidence via the cue-guided tree search. Next, the test-time calibration will be conducted, and the calibrated scores will be yielded. Based on the global and local calibrated scores, the multi-cue fusion contributes to the final prediction.

aggregation. In practice, the resulting cue-conditioned views are reused for all-label NLL scoring and later fused by the multi-cue inference module. Furthermore, several efficiency-oriented implementation optimizations are employed, including caching and batched inference, without changing the scoring objective. The implementation details are discussed in the SM B.

3.4 NLL-based Test-Time Calibration

Our analysis in Sec. 4.2 shows that the failure of vanilla MLLMs in zero-shot MGR is not only caused by missing local evidence, but also by severe bias in label prediction. For instance, the model tends to collapse to a few dominant labels. These observations suggest that raw label ranking is substantially affected by prior-driven bias, rather than being determined by the visual evidence. As a result, directly using the raw label scores can lead to unreliable predictions.

To address this issue, we introduce an NLL-based test-time calibration module. The core idea is to compare the NLL of an input view x with two controlled calibration variants: a blank calibration input x_b obtained by masking the video input and a lowrank calibration input x_r (motion-based calibration) obtained by repeating a single frame across time. The blank calibration input suppresses visual content and mainly reflects language prior, while the lowrank calibration input preserves static appearance but removes motion information.

Let $\mathbf{v}(y)$ denote the verbalized label string. For a general input view x, we define the per-label NLL as

$$\ell_x(y) = -\frac{1}{|\mathbf{v}(y)|} \log p_{\mathcal{M}}(\mathbf{v}(y) \mid x, t), \tag{5}$$

with token-level factorization under teacher forcing. Hence, the prediction is

$$y^* = \arg\min_{y \in \mathcal{Y}} \ell_x(y). \tag{6}$$

We further define the induced label distribution:

$$\hat{p}_x(y) = \frac{\exp(-\ell_x(y))}{\sum_{y'} \exp(-\ell_x(y'))}. \tag{7}$$

We then perform calibration in NLL space by

$$\tilde{\ell}_{x_b}(y) \triangleq \ell_x(y) - \ell_{x_b}(y), \qquad \tilde{\ell}_{x_r}(y) \triangleq \ell_x(y) - \ell_{x_r}(y). \tag{8}$$

The calibrated scores are aggregated by:

$$\bar{\ell}_x(y) \triangleq -\tau_d \log\left(\pi_x \exp\left(-\frac{\tilde{\ell}_{x_b}(y)}{\tau_d}\right) + (1-\pi_x)\exp\left(-\frac{\tilde{\ell}_{x_r}(y)}{\tau_d}\right)\right), \tag{9}$$

where $\tau_d > 0$ is a temperature parameter, and $\pi_x \in (0,1)$ controls the relative contribution of the blank and lowrank references. This aggregated energy is retained for subsequent cue-conditioned fusion. See the full derivation in SM C.

3.5 Multi-Cue Fusion

Given the cue-guided zoomed-in views $z_{1:n}$ extracted from the global clip g in the tree search stage, we present the multi-cue fusion module that fuses multiple local evidences into a single label distribution over the candidate set $\mathcal{Y}$.

For each cue view z_i, a paired input $x_i \triangleq (g, z_i)$ is constructed and scored by the MLLM using the same teacher-forced objective in Sec. 3.4, yielding per-label loss $\ell_{x_i}(y)$, and the induced distribution $\hat{p}_{x_i}(y)$. The module is designed to derive an aggregation rule fusing $\{\hat{p}_{x_i}(y)\}_{i=1}^n$ to predict the y_{final} instead of selecting one cue branch.

The product of experts [18] is adopted to model the joint conditional probability,

$$p(y \mid x_1, \dots x_n) \approx \prod_{i=1}^{n} p(y \mid x_i). \tag{10}$$

However, simply adopting products of $p(y \mid x_i)$ will lead to the problem that the information of the global view will accumulate repeatedly because $\{x_i \mid i = 1, \ldots, n\}$ share the same global view. To address this issue, we adopt the Naive Bayes approximation [44], (see detailed derivation in SM D) with terms independent of y (e.g., $\prod_i p(z_i \mid g)$) and propose an ideal posterior factorization:

$$p(y \mid g, z_{1:n}) \propto p(y \mid g) \prod_{i=1}^{n} \frac{p(y \mid g, z_i)}{p(y \mid g)} = p(y \mid g)^{1-n} \prod_{i=1}^{n} p(y \mid g, z_i). \tag{11}$$

Implementation in Calibrated NLL Space. Eq. (11) gives an ideal posterior factorization under the conditional independence approximation. In practice, however, cue-conditioned branches are neither independent nor equally reliable; their rankings are not directly comparable, and the set of cues is label-dependent. Therefore, Eq. (11) is retained as the probabilistic motivation, while instantiating the final inference rule in a calibrated energy space.

Using the notation in Sec. 3.4, for the global branch g, the calibrated losses defined previously are directly used $\tilde{\ell}_{g_b}(y)$ and $\tilde{\ell}_{g_r}(y)$. For the i-th paired branch $x_i = (g, z_i)$, we abbreviate $\bar{\ell}_{(x_i)_b}(y)$ by $\bar{\ell}_{i,b}(y)$.

Hence, $\bar{\ell}_x(y)$ is a temperature-smoothed lower envelope of the two bias calibration energies, and serves as the branch-wise energy used in the subsequent fusion.

Let

$$\bar{p}_x(y) \triangleq \frac{\exp(-\bar{\ell}_x(y))}{\sum_{y' \in \mathcal{Y}} \exp(-\bar{\ell}_x(y'))} \tag{12}$$

be the surrogate posterior induced by $\bar{\ell}_x(y)$. Substituting Eq. (12) into the factorization in Eq. (11) yields, up to a Naive-Bayesian (NB) label-independent constant,

$$\ell_{\mathrm{NB}}(y) \doteq \sum_{i=1}^{n} \bar{\ell}_i(y) - (n-1)\bar{\ell}_g(y) = \bar{\ell}_g(y) + \sum_{i=1}^{n} \big(\bar{\ell}_i(y) - \bar{\ell}_g(y)\big), \tag{13}$$

which shows that each cue branch contributes a residual correction relative to the shared global branch.

However, Eq. (13) is not yet suitable as the final rule, because not all cues are semantically relevant to every label. We propose a refined multi-cue fusion strategy, called **Label-Conditioned Normalized Fusion** (please see the detailed derivation in SM D).

Let $\mathcal{I}(y) = \{i : c_i \in C(y)\}$ denote the compatible cue set, where $C(y)$ is defined once from label semantics and body-part association (See full mapping in SM E.1), and $w_i = \lambda_{\phi(c_i)}$ the family weight. We fuse global and local evidence as

$$\ell_{\mathrm{fuse}}(y) = \frac{\bar{\ell}_g(y) + \sum_{i \in \mathcal{I}(y)} w_i \, \bar{\ell}_i(y)}{1 + \sum_{i \in \mathcal{I}(y)} w_i}. \tag{14}$$

The final prediction is

$$y_{\mathrm{final}} = \arg\min_{y \in \mathcal{Y}} \ell_{\mathrm{fuse}}(y). \tag{15}$$

4 Experiments

4.1 Experimental Setup

Datasets. In the experiments, we evaluate our method on the iMiGUE [35] test set and MA52 [17] validation set as they align with our clip-level MGR setting, while other available datasets follow different or derivative protocols. The primary MLLM backbone is Qwen2.5-VL-7B [5] (by default unless otherwise specified), while results with additional backbones are provided in the SM E.1. For iMiGUE, experiments are conducted on a single NVIDIA A100 GPU with 40GB memory, whereas MA52 experiments are conducted on AMD MI250X GPUs due to the larger size. See more details of experimental configurations, including the values of parameters and compatible cues used in multi-cue fusion, in SM E.1.

Evaluation. Since MG datasets are strongly long-tailed, to ensure fair evaluation, we report mean-class Top-1 (mCA@1), mean-class Top-5 (mCA@5), and macro-F1 (MF1), which provide a more balanced assessment than overall accuracy.

Prompt Design. We compare two prompt variants for NLL-based inference: with and without the candidate label list. The results show that this choice leads to noticeable performance differences. To keep a fair comparison, we adopt the NLL prompt whose semantic content is aligned with that used in free-form generation, meaning that we try to keep those prompts the same with the least modification. Detailed results are provided in SM E.2.

4.2 Why Vanilla MLLMs Struggle with MGR: A Two-View Analysis

We first analyze why vanilla MLLM inference has poor performance on MGR based on two bottlenecks.

First, under the common general prompting setting, i.e., prompt-based free-form generation, vanilla MLLMs exhibit weak grounding on discriminative local evidence. To better analyze this limitation, we employ TAM [30] for token-level visual attribution. The resulting maps suggest that, even when the model partially captures the correct semantics, it often fails to consistently attend to the truly decisive regions, such as the lips area. Such insufficient localization therefore leads to misclassification. A representative failure case visualized by TAM is shown in Fig. 3. In contrast, when effective local evidence is explicitly leveraged, the correct prediction can be recovered, which is provided in SM F.1.

Second, under free-form generation, Qwen2.5-VL shows a clear prediction-collapse pattern. As shown in Fig. 4a, a large portion of test clips is mapped to a few dominant labels, leading to poor coverage of the label space and a distribution inconsistent with the ground-truths (GTs). This suggests that **the visual input is often insufficient to override the language priors of the model, namely the language bias**. Besides, this prediction-collapse pattern is observed across various MLLMs, as revealed by Fig. 4b. Moreover, the bias-calibrated NLL analysis is conducted to quantify prior-driven score distortion. Due to the space limitation, relevant results and figures are listed in SM F.2. The NLL analysis indicates that the two calibrations introduced address overlapping but distinct subsets of errors, suggesting that language bias and appearance bias are complementary rather than redundant. These observations motivate the two key components of Zero-MELO: localized evidence acquisition and dual-reference calibration.

4.3 Experimental Results

Comparison with SOTA Models. We first evaluate zero-shot MGR performance on both iMiGUE and MA52. As shown in Table 1, Zero-MELO is compared with SOTA vanilla MLLMs and

Table 1: Zero-shot performance comparison on iMiGUE testset and MA52 valid set.

Method	iMiGUE			MA52		
	mCA@1	mCA@5	MF1	mCA@1	mCA@5	MF1
Prompt-based Zero-shot						
GPT-4o mini [40]	0.1113	0.3392	0.0699	0.0576	0.2362	0.0363
Gemini2.5-Flash [10]	0.1917	0.3689	0.1367	0.1666	0.3072	0.1217
MiMo-VL-7B [54]	0.2445	0.5643	0.1506	0.1869	0.3814	0.1042
InternVL3.5-8B [59]	0.2034	0.4358	0.0927	0.1796	0.3677	0.0914
Qwen2.5-VL-7B [5]	0.1615	0.4101	0.0728	0.1020	0.3133	0.0680
NLL-Based						
MiMo-VL-7B	0.0325	0.2480	0.0008	0.0402	0.1521	0.0144
InternVL3.5-8B	0.0329	0.2408	0.0007	0.0721	0.2387	0.0414
Qwen2.5-VL-7B	0.0523	0.2524	0.0072	0.1178	0.3299	0.0714
Test-Time Method[*]						
ZoomEye [**] [47]	0.0760	0.2340	0.0326	0.1185	0.3508	0.0723
VCD [24]	0.2077	0.4717	0.1045	0.1516	0.3364	0.0888
Zero-MELO	**0.2684**	**0.5775**	0.1325	**0.2210**	**0.4359**	**0.1221**

[*] All test-time methods in this table use Qwen2.5-VL as the common backbone for fair comparison.
[**] ZoomEye was originally proposed for image inputs. For this comparison, we adapt it to process video inputs and conduct it in the NLL-based setting.

Figure 3: The TAM visualization of a failure case for the gesture *Pressing lips*. The wrong prediction is *Bulging face, deep breath*. We show the token-wise activation maps for three successive generated tokens during free-form generation in (a), (b), and (c), respectively. Although the model suggests the GT in Top-5, the highlighted regions are not consistently grounded on the lip area or the most discriminative motion phase, indicating that the naive single-pass MLLM fails to capture compact motion-centric local evidence.

representative test-time baselines, including ZoomEye and VCD. The results demonstrate that Zero-MELO achieves state-of-the-art performance, effectively mitigating the limitations of global-view MLLM inference in zero-shot MGR.

NLL-Based vs. Prompt-based Generation. As discussed in Sec. 3 regarding the technical path choice of MLLMs, we compare the results of NLL-based zero-shot and prompt-based free-form zero-shot in Table 1. For most compared models, NLL-based inference is weaker than prompt-based free-form generation under the current prompt design, due to the amplified prediction-collapse problem. For instance, MiMo-VL ranks *Buckle button, pulling shirt collar, adjusting tie* as the Top-1 prediction for nearly all samples despite minor score variation across samples. In contrast, applying blank calibration substantially improves the NLL-based results and significantly narrows the gap to free-form generation, which supports the effectiveness of the proposed calibration. For instance, InternVL3.5 improves from 3.29% to 17.68%, yielding an absolute gain of 14.39 percentage points. More detailed analyses are discussed in SM G.

Moreover, we also observe a dataset-related reversal in Table 1. While prompt-based baselines are consistently stronger on iMiGUE, NLL-based vanilla MLLMs are consistently stronger on MA52. This counterintuitive gap suggests that the behavior of raw NLL ranking is not determined solely by visual difficulty, but is also strongly affected by dataset-dependent prior structure in the label space. After calibration, this reversal becomes less consistent, indicating that calibration and evidence refinement partially reduce the bias. Detailed analyses are also provided in SM G.

Zero-shot MLLMs vs. Supervised Discriminative Models. We further report supervised discriminative models for reference. As shown in Table 2, although these methods are not directly comparable to our zero-shot setting, Zero-MELO outperforms BlockerGCN(J) and remains within a modest gap of stronger supervised baselines in mCA@1. However, supervised methods still show clear advantages on mCA@5 and MF1. The full comparison, including other modality variants, is provided in SM H. In the full comparison, Zero-MELO surpasses all BlockerGCN variants in mCA@1, approaches MMN(2s), and remains below PoseC3D.

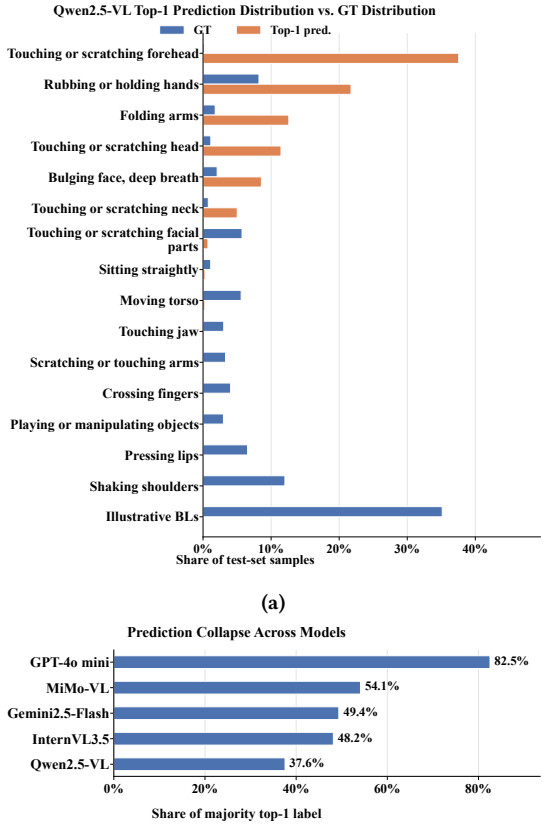

Figure 4: Failure analysis on iMiGUE of the prompt-based free-form generation. (a) The comparison of distributions of Qwen2.5-VL predictions and the ground-truths (GTs) of iMiGUE. Due to the space limitations, only the appearing frequency of labels is larger than 1% in either GT or Top-1 prediction is listed. (b) The comparison of the severity of the prediction-collapse across different vanilla MLLMs.

Table 2: Comparison with supervised discriminative models on iMiGUE dataset. For space, we list only the representative variants here; the full comparison is provided in SM H.

Method	mCA@1	mCA@5	MF1
BlockerGCN(J) [79]	0.2324	0.6080	0.2375
PoseC3D [12]	0.3402	0.7159	0.3583
MMN(J) [15]	0.2991	0.6626	0.3062
Zero-MELO (Ours)	**0.2684**	0.5775	0.1325

SFT vs Zero-shot for MLLM. Since zero-shot performance does not represent the upper bound of MLLM capability on MGR, we further examine how far a vanilla MLLM can be improved through supervised fine-tuning. Specifically, we fine-tune Qwen2.5-VL under two task formulations: an NLL-style label-scoring formulation and a VQA-style formulation. As shown in Table 3, both formulations improve over the vanilla zero-shot baseline, and the VQA-style formulation performs better than the label-scoring formulation. However, both remain below Zero-MELO, suggesting that our test-time framework is more effective than straightforward SFT under the long-tailed label distribution.

Table 3: Comparison between our method and the SFT method on MA52 dataset.

Method	mCA@1	mCA@5	MF1
VQA	0.177	0.4201	0.0842
Label Scoring	0.0899	0.2545	0.0623
Ours	**0.2210**	**0.4359**	**0.1221**

4.4 Ablation Study

Different Backbone. Different backbones are explored with Zero-MELO. The full results are provided in SM I, where Zero-MELO consistently improves over the corresponding vanilla MLLMs.

Different Test-Time Calibration Strategies. The blank and lowrank calibrations are ablated, respectively. The results are listed in SM J since the page limitation is strict. The experiment indicates that the joint fusion of these two calibrations can contribute to better performance, aligning with Sec. 4.2 that the two calibrations proposed aim to mitigate different biases and are complementary.

Component Validity. Each component of Zero-MELO is ablated in Table 4. Note that the vanilla baseline in this table is not numerically identical to that in Table 1, due to the structural difference of default visual input and the minor prompt alternatives. Check SM K for details.

Note that purely adding the tree search module leaves mCA@1 almost unchanged. This is reasonable since the language prior bias is dominant and thus changes of visual input do not affect Top-1 prediction, while the additional local evidence mainly moves the ground-truth label upward in the ranking, improving mCA@5 (more useful information). Considering such a strong language prior, we use blank calibration to first mitigate the bias in order to compare the benefits of the search module fairly. As such, the more sufficient local evidence acquired via the search algorithm contributes to the higher mCA@1. Notably and encouragingly, simply applying blank calibration (remove language bias) allows significant improvement of mCA@1 from 0.0357 to 0.2040, supporting the existence of the dominant language prior. The effectiveness of the search module under blank calibration also suggests that the insufficient local evidence and severe biases are not identical but complementary reasons for the failures of MLLMs in MGR. Besides, aggregate calibration further improves performance, which supports the analysis in Sec. 4.2. Finally, multi-cue fusion yields consistent gains, and our strategy achieves the best performance. Due to the space limitation, details of the experiment are illustrated in SM K.

4.5 Case Analysis

Case Analysis. Fig. 5 provides case studies for the complementary roles of the three components in Zero-MELO, namely the tree

Figure 5: Case study of Zero-MELO. Each panel shows the global video clip, the zoomed-in view of the most supportive cue, and the trajectory of the prediction Top-1 across global inference, calibration, and final multi-cue fusion. The (a), (b), (c) correspond to the successful samples, while (d) is the failure case.

Table 4: Ablation study of different components of Zero-MELO on iMiGUE dataset. Cali means calibration, and the TS represents tree search. F is the abbreviation of Fusion. Max means label-wise selecting the cue with the lowest calibrated NLL. Average means averaging over compatible cues.

Method	mCA@1	mCA@5	MF1
Vanilla	0.0357	0.2084	0.0094
+ TS	0.0358	0.2629	0.0100
+ Cali (Blank)	0.2040	0.4943	0.0855
+ Cali (Aggregate)	0.2363	0.5436	0.1097
+ TS + Cali (Blank)	0.2131	0.5078	0.0924
+ TS + Cali + F (Max)	0.2287	0.5365	0.1227
+ TS + Cali + F (Average)	0.2508	0.5621	0.1278
+ TS + Cali + F (**Ours**)	**0.2684**	**0.5775**	**0.1325**

search mechanism, the test-time calibration module, and the multi-cue fusion module. In the subfigure **(a)**, the calibration shifts the prediction from the correct fine-grained label to the broader class *Illustrative BLs*, while the localized finger cue preserves the critical crossing pattern, allowing multi-cue fusion to recover the correct prediction. In the subfigure **(b)**, the global branch yields a spurious hand-centric prediction (*Crossing fingers*), whereas test-time calibration suppresses the bias induced by prior and corrects the top-1 prediction to *Pressing lips*. In the subfigure **(c)**, both global inference and calibration remain biased toward *Bulging face, deep breath*; however, tree search localizes the mouth region and exposes the discriminative lip-compression cue, enabling the final prediction to switch to the ground-truth class. Taken together, these examples show that calibration mitigates prior-driven bias; tree search identifies discriminative local evidence, and multi-cue fusion consolidates cue-specific evidence to improve final recognition.

However, Zero-MELO still depends on whether the localized cue crop preserves sufficient relevant evidence. As shown in the subfigure **(d)**, both the global branch and calibration correctly predict *Playing or manipulating objects*, yet the localized hand crop fails to retain sufficient object-related evidence and mainly captures an ambiguous hand pose. As a result, cue-conditioned fusion overemphasizes the local cue and shifts the final prediction to the nearby hand-centric class *Minaret gesture*.

4.6 Limitation

Although Zero-MELO compares favorably with existing MLLM-based methods, it still has two main limitations. First, its performance is sensitive to cue localization quality: if the most informative cue is missed, weakly captured, or degraded after zoom-in, multi-cue fusion often cannot recover the correct prediction, especially in highly fine-grained cases where pose alone is insufficient and object or context cues are required. Improving localization robustness is therefore an important direction. Second, Zero-MELO introduces nontrivial test-time overhead due to multi-cue search and repeated label scoring, despite practical caching and batching. We leave standardized runtime comparison to future work, as comparable latency statistics are rarely reported in prior work.

5 Conclusion

In this paper, we first revisit the failure patterns of MLLMs in MGR. We observe that MLLMs suffer from two bottlenecks: insufficient localized evidence and severe score biases driven by language and motion-agnostic appearances. Accordingly, we present Zero-MELO. Extensive experiments validate the effectiveness of Zero-MELO over different backbones and datasets. Our results suggest that the limitations of current MLLMs on MGR are twofold: a substantial portion comes from insufficient elicitation under standard inference, while the remaining hard cases indicate that current models still have limited capability on extremely fine-grained motion-centric understanding.

References

[1] Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. 2025. Mitigating object hallucinations

in large vision-language models with assembly of global and local attention. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 29915–29926.

[2] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. 2025. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661* (2025).

[3] Hillel Aviezer, Yaacov Trope, and Alexander Todorov. 2012. Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science* 338, 6111 (2012), 1225–1229.

[4] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631* (2025).

[5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923 [cs.CV] https://arxiv.org/abs/2502.13923

[6] Michal Balazia, Philipp Müller, Ákos Levente Tánczos, August von Liechtenstein, and François Brémond. 2022. Bodily behaviors in social interaction: Novel annotations and state-of-the-art evaluation. In *Proceedings of the 30th ACM International Conference on Multimedia*. 70–79.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[8] Zongsheng Cao, Yangfan He, Anran Liu, Jun Xie, Zhepeng Wang, and Feng Chen. 2025. CoFi-Dec: Hallucination-Resistant Decoding via Coarse-to-Fine Generative Feedback in Large Vision-Language Models. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 10709–10718.

[9] Haoyu Chen, Henglin Shi, Xin Liu, Xiaobai Li, and Guoying Zhao. 2023. Smg: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis. *International Journal of Computer Vision* 131, 6 (2023), 1346–1366.

[10] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025).

[11] Yipeng Du, Tiehan Fan, Kepan Nan, Rui Xie, Penghao Zhou, Xiang Li, Jian Yang, Zhenheng Yang, and Ying Tai. 2025. Motionsight: Boosting fine-grained motion understanding in multimodal llms. *arXiv preprint arXiv:2506.01674* (2025).

[12] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. 2022. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2969–2978.

[13] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2025. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[14] Rong Gao, Xin Liu, Bohao Xing, Zitong Yu, Bjorn W Schuller, and Heikki Kälviäinen. 2026. Identity-free artificial emotional intelligence via micro-gesture understanding. *IEEE Transactions on Affective Computing* (2026).

[15] Jihao Gu, Kun Li, Fei Wang, Yanyan Wei, Zhiliang Wu, Hehe Fan, and Meng Wang. 2025. Motion matters: Motion-guided modulation network for skeleton-based micro-action recognition. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 5461–5470.

[16] Jihao Gu, Fei Wang, Kun Li, Yanyan Wei, Zhiliang Wu, and Dan Guo. 2025. MM-gesture: towards precise micro-gesture recognition through multimodal fusion. *arXiv preprint arXiv:2507.08344* (2025).

[17] Dan Guo, Kun Li, Bin Hu, Yan Zhang, and Meng Wang. 2024. Benchmarking micro-action recognition: Dataset, methods, and applications. *IEEE Transactions on Circuits and Systems for Video Technology* 34, 7 (2024), 6238–6252.

[18] G.E. Hinton. 1999. Products of experts. In *1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470)*, Vol. 1. 1–6 vol.1. doi:10.1049/cp:19991075

[19] Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. 2025. Deepeyesv2: Toward agentic multimodal model. *arXiv preprint arXiv:2511.05271* (2025).

[20] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. 2025. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006* (2025).

[21] Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. 2024. Self-introspective decoding: Alleviating hallucinations for large vision-language models. *arXiv preprint arXiv:2408.02032* (2024).

[22] Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. 2025. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 25004–25014.

[23] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).

[24] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13872–13882.

[25] Deng Li, Bohao Xing, Xin Liu, Baiqiang Xia, Bihan Wen, and Heikki Kälviäinen. 2025. Deemo: De-identity multimodal emotion recognition and reasoning. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 5707–5716.

[26] Kun Li, Dan Guo, Guoliang Chen, Chunxiao Fan, Jingyuan Xu, Zhiliang Wu, Hehe Fan, and Meng Wang. 2025. Prototypical calibrating ambiguous samples for micro-action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 4815–4823.

[27] Kun Li, Pengyu Liu, Dan Guo, Fei Wang, Zhiliang Wu, Hehe Fan, and Meng Wang. 2025. Mmad: Multi-label micro-action detection in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13225–13236.

[28] Qiankun Li, Xiaolong Huang, Huabao Chen, Feng He, Qiupu Chen, and Zengfu Wang. 2024. Advancing micro-action recognition with multi-auxiliary heads and hybrid loss optimization. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 11313–11319.

[29] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*. 12286–12312.

[30] Yi Li, Hualiang Wang, Xinpeng Ding, Haonan Wang, and Xiaomeng Li. 2025. Token activation map to visually explain multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 48–58.

[31] Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying Liu. 2017. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *arXiv preprint arXiv:1703.07475* (2017).

[32] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. 2019. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence* 42, 10 (2019), 2684–2701.

[33] Pengyu Liu, Kun Li, Fei Wang, Yanyan Wei, Junhui She, and Dan Guo. 2025. Online Micro-gesture Recognition Using Data Augmentation and Spatial-Temporal Attention. *arXiv preprint arXiv:2507.09512* (2025).

[34] Sheng Liu, Haotian Ye, and James Zou. 2025. Reducing hallucinations in large vision-language models via latent space steering. In *The Thirteenth International Conference on Learning Representations*.

[35] Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. 2021. imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10631–10642.

[36] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2024. Mmbench: Is your multi-modal model an all-around player?. In *European conference on computer vision*. Springer, 216–233.

[37] Hao Lu, Xuesong Niu, Jiyao Wang, Yin Wang, Qingyong Hu, Jiaqi Tang, Yuting Zhang, Kaishen Yuan, Bin Huang, Zitong Yu, et al. 2024. Gpt as psychologist? preliminary evaluations for gpt-4v on visual affective computing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 322–331.

[38] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems* 36 (2023), 46534–46594.

[39] Sean O'Brien and Mike Lewis. 2023. Contrastive decoding improves reasoning in large language models. *arXiv preprint arXiv:2309.09117* (2023).

[40] OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2026-04-01.

[41] Santosh Patapati, Trisanth Srinivasan, and Amith Adiraju. 2025. CLIP-MG: Guiding Semantic Attention with Skeletal Pose Features and RGB Data for Micro-Gesture Recognition on the iMiGUE Dataset. *arXiv preprint arXiv:2506.16385* (2025).

[42] Daiqing Qi, Dongliang Guo, Hanzhang Yuan, Handong Zhao, Mengxuan Hu, Lehan Yang, and Sheng Li. 2025. Improve Temporal Reasoning in Multimodal Large Language Models via Video Contrastive Decoding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In

International conference on machine learning. PmLR, 8748–8763.

[44] Stuart Russell, Peter Norvig, and Artificial Intelligence. 1995. A modern approach. *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs* 25, 27 (1995), 79–80.

[45] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. 2016. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1010–1019.

[46] Tuyun Shang, Yanbin Hao, Ming Pei, Kun Li, Huixia Ben, and Shuo Wang. 2025. Cross-modal Feature Enhancement and Contrastive Alignment for Micro-gesture Recognition. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 203–217.

[47] Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. 2025. Zoomeye: Enhancing multimodal llms with human-like zooming capabilities through tree-based image exploration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 6613–6629.

[48] Yuxiang Shen, Hailong Huang, Zhenkun Gao, Xueheng Li, Chengjun Xie, Xuanhua He, and Jie Zhang. 2026. AdaFocus: Knowing When and Where to Look for Adaptive Visual Reasoning. *arXiv preprint arXiv:2603.00171* (2026).

[49] Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*. 783–791.

[50] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314* (2024).

[51] Jingran Su, Jingfan Chen, Hongxin Li, Yuntao Chen, Li Qing, and Zhaoxiang Zhang. 2025. Activation steering decoding: Mitigating hallucination in large vision-language models through bidirectional hidden state intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 12964–12974.

[52] Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. 2025. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918* (2025).

[53] Wei Suo, Lijun Zhang, Mengyang Sun, Lin Yuanbo Wu, Peng Wang, and Yanning Zhang. 2025. Octopus: Alleviating hallucination via dynamic contrastive decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 29904–29914.

[54] Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, Kainan Bao, Hao Tian, Hailin Zhang, Gang Wang, Dawei Zhu, Cici, Chenhong He, Bowen Ye, Bowen Shen, Zihan Zhang, Zihan Jiang, Zhixian Zheng, Zhichao Song, Zhenbo Luo, Yue Yu, Yudong Wang, Yuanyuan Tian, Yu Tu, Yihan Yan, Yi Huang, Xu Wang, Xinzhe Xu, Xingchen Song, Xing Zhang, Xing Yong, Xin Zhang, Xiangwei Deng, Wenyu Yang, Wenhan Ma, Weiwei Lv, Weiji Zhuang, Wei Liu, Sirui Deng, Shuo Liu, Shimao Chen, Shihua Yu, Shaohui Liu, Shande Wang, Rui Ma, Qiantong Wang, Peng Wang, Nuo Chen, Menghang Zhu, Kangyang Zhou, Kang Zhou, Kai Fang, Jun Shi, Jinhao Dong, Jiebao Xiao, Jiaming Xu, Huaqiu Liu, Hongshen Xu, Heng Qu, Haochen Zhao, Hanglong Lv, Guoan Wang, Duo Zhang, Dong Zhang, Di Zhang, Chong Ma, Chang Liu, Can Cai, and Bingquan Xia. 2025. MiMo-VL Technical Report. arXiv:2506.03569 [cs.CL] https://arxiv.org/abs/2506.03569

[55] Renzhi Wang, JianHua Hu, and Zhanjiang Yuan. 2024. Skeleton Based Micro-Gestures Classification. In *2024 7th International Conference on Mechatronics and Computer Technology Engineering (MCTE)*. IEEE, 1426–1431.

[56] Ruosi Wang, Yuhan Wang, and Zhaoqiang Xia. 2025. A Gesture-centered Dual-modal Network for Micro-Gesture Emotion Recognition. In *2025 4th International Conference on Image Processing and Media Computing (ICIPMC)*. IEEE, 109–113.

[57] Shijian Wang, Jiarui Jin, Xingjian Wang, Linxin Song, Runhao Fu, Hecheng Wang, Zongyuan Ge, Yuan Lu, and Xuelian Cheng. 2025. Video-Thinker: Sparking" Thinking with Videos" via Reinforcement Learning. *arXiv preprint arXiv:2510.23473* (2025).

[58] Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. 2025. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 7907–7915.

[59] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. 2025. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265* (2025).

[60] Wenbin Wang, Yongcheng Jing, Liang Ding, Yingjie Wang, Li Shen, Yong Luo, Bo Du, and Dacheng Tao. 2025. Retrieval-Augmented Perception: High-resolution Image Perception Meets Visual RAG. In *International Conference on Machine Learning*. PMLR, 63290–63307.

[61] Lai Wei, Liangbo He, Jun Lan, Lingzhong Dong, Yutong Cai, Siyuan Li, Huijia Zhu, Weiqiang Wang, Linghe Kong, Yue Wang, et al. 2026. Zooming without Zooming: Region-to-Image Distillation for Fine-Grained Multimodal Perception. *arXiv preprint arXiv:2602.11858* (2026).

[62] Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. 2025. Don't miss the forest for the trees: Attentional vision calibration for large vision language models. In *Findings of the Association for Computational Linguistics: ACL 2025*. 1927–1951.

[63] Penghao Wu and Saining Xie. 2024. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13084–13094.

[64] Zhaoqiang Xia, Hexiang Huang, Haoyu Chen, Xiaoyi Feng, and Guoying Zhao. 2026. Hybrid-Supervised Hypergraph-Enhanced Transformer for Micro-Gesture Based Emotion Recognition. *IEEE Transactions on Affective Computing* 17, 1 (2026), 379–393. doi:10.1109/TAFFC.2025.3618639

[65] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[66] Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. 2025. Understanding and Mitigating Hallucination in Large Vision-Language Models via Modular Attribution and Intervention. In *International Conference on Learning Representations*, Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (Eds.), Vol. 2025. 51546–51568.

[67] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems* 36 (2023), 11809–11822.

[68] Hao Yin, Guangzong Si, and Zilei Wang. 2025. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 14625–14634.

[69] Xuan Yu, Dayan Guan, and Yanfeng Gu. 2025. Zoom-Refine: Boosting High-Resolution Multimodal Understanding via Localized Zoom and Self-Refinement. *arXiv preprint arXiv:2506.01663* (2025).

[70] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11975–11986.

[71] Congyue Zhang, Wenjie Fu, Canrong Tian, Xu Cheng, Yuan Tian, and Hao Yu. 2024. 3D Convolutional Network based micro-gesture recognition. In *Proceedings of the ACM Turing Award Celebration Conference-China 2024*. 193–198.

[72] Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Na Zhao, Zhiyu Tan, Hao Li, Xingjun Ma, and Jingjing Chen. 2024. Eventhallusion: Diagnosing event hallucinations in video llms. *arXiv preprint arXiv:2409.16597* (2024).

[73] Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. 2025. Mllms know where to look: Training-free perception of small visual details with multimodal llms. *arXiv preprint arXiv:2502.17422* (2025).

[74] YiFan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. 2025. MME-RealWorld: Could Your Multimodal LLM Challenge High-Resolution Real-World Scenarios that are Difficult for Humans?. In *The Thirteenth International Conference on Learning Representations*.

[75] Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, et al. 2025. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630* (2025).

[76] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. 2025. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362* (2025).

[77] Liangyu Zhong, Fabio Rosenthal, Joachim Sicking, Fabian Hüger, Thorsten Bagdonat, Hanno Gottschalk, and Leo Schwinn. [n. d.]. FOCUS: Internal MLLM Representations for Efficient Fine-Grained Visual Question Answering. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[78] Jian Zhou, Jingchao Yao, Nan Su, Jingchen Lu, Qingyang Yu, Yichi Zhang, and Wenqiang Hu. 2025. KMANet: A spatio-temporal enhancement network for micro-action recognition. *Knowledge-Based Systems* (2025), 114139.

[79] Yuxuan Zhou, Xudong Yan, Zhi-Qi Cheng, Yan Yan, Qi Dai, and Xian-Sheng Hua. 2024. Blockgcn: Redefine topology awareness for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2049–2058.

[80] Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. 2025. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 1624–1633.