

GMoT: Gated Motion-Aware Tokenization for Fine-Grained Micro-Gesture Video Reasoning with Multimodal LLMs

Taorui Wang^{1,2} Wei Xia² Hui Ma² Zijia Song²
 Jiayu Zhang² Zeheng Wang² Yong Xu^{1,*} Zitong Yu^{2,3,*}

¹Harbin Institute of Technology, Shenzhen, Shenzhen, China

²Great Bay University, Dongguan, China

³Dongguan Key Laboratory for Intelligence and Information Technology, Dongguan, China

wangtaorui@stu.hit.edu.cn laterfall@hit.edu.cn zitong.yu@ieee.org

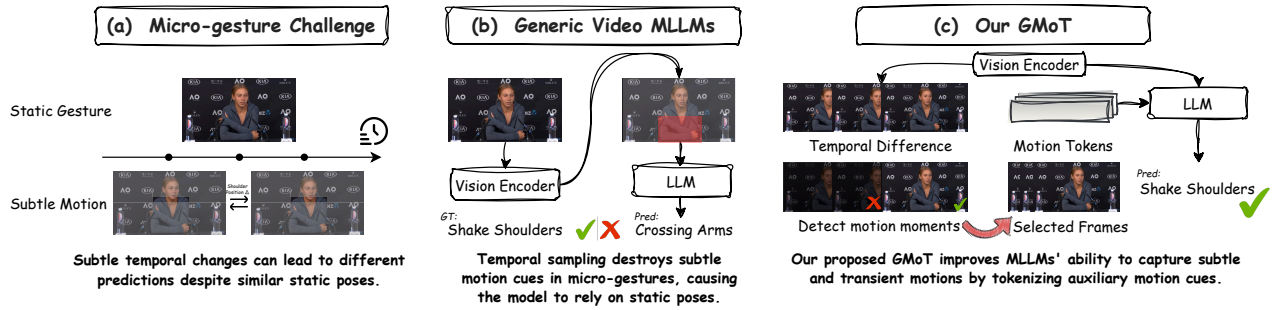


Figure 1: Problem setting and motivation. Generic video MLLMs often over-emphasize static appearance, whereas GMoT explicitly distills weak local motion into compact motion-aware tokens for evidence-grounded micro-gesture reasoning.

Abstract

Micro-gesture recognition demands the detection of fleeting, spatially localized movements that are frequently overwhelmed by dominant static appearances and background noise. While Multimodal Large Language Models (MLLMs) excel at general video understanding, they inherently struggle with subtle kinematics and often rely on static posture priors. To this end, we propose GMoT, a Gated Motion-Aware Tokenization module that explicitly distills sparse kinematic evidence into a compact sequence prior to temporal modeling. GMoT dynamically spotlights action-relevant regions via spatially weighted pooling, extracts adjacent-frame temporal differencing to capture precise motion energy, and adaptively fuses these cues into the visual stream using a conservatively initialized semantic gate. To transition from simple classification to evidence-grounded reasoning, we further introduce a progressive reward-guided policy refinement paradigm, supported by a semi-supervised annotation pipeline that generates anatomically focused captions. Beyond achieving the best Top-1 accuracy among the compared methods on iMiGUE (67.32%) and SMG (73.11%), improving the Qwen3-VL-8B baseline by +6.80 and +3.11 points, our framework introduces Body-Region Grounding (BRG) Recall as an anatomical-grounding proxy conditioned on correct predictions, together with an overlapping-label cross-domain transfer protocol between iMiGUE and SMG. Extensive evaluations demonstrate that our GMoT-augmented model improves in-domain accuracy, retains clear gains under label-preserving corruptions, and improves accuracy-oriented cross-domain transfer under explicit small-split

caveats while maintaining high anatomical grounding in its generated rationales. All code will be publicly released in our GitHub repository.

CCS Concepts

• Computing methodologies → Computer vision.

Keywords

Multimodal Large Language Models, Micro-Gesture Recognition, Temporal Motion Modeling

1 Introduction

Micro-gestures (MGs) are subtle, spontaneous, and highly localized human actions that reveal latent affective and cognitive states during daily communication [1, 10, 33, 45]. Unlike macroscopic actions (e.g., running or jumping), MGs, such as a brief finger twitch, a slight shoulder shrug, or a tense lip press, manifest in extremely narrow spatial regions and occur across just a few frames [11, 45]. Accurately recognizing these movements provides fine-grained behavioral cues essential for downstream applications such as emotion understanding, psychological interaction analysis, and multimodal deception detection [24, 25, 38, 56]. However, visually similar categories often differ only in subtle kinematic details, which makes reliable discrimination under real-world conditions particularly challenging.

Recently, Multimodal Large Language Models (MLLMs) have revolutionized video understanding by coupling powerful visual

*Corresponding authors: Yong Xu and Zitong Yu.

encoders with the reasoning capabilities of Large Language Models (LLMs) [2, 32]. In theory, MLLMs offer an attractive foundation for micro-gesture recognition, which can go beyond merely outputting generic class labels to connect visual evidence with language-based explanations and articulate what moved and how it changed [23, 25]. However, directly applying existing MLLMs to micro-gesture recognition often yields suboptimal and misleading results, because their reliance on dense frame-level spatial patch extraction causes weak and sparse kinematic signals to be either missed by sparse sampling or diluted by temporal pooling [2, 9]. Without sufficient temporal sensitivity, models tend to resort to what we term *hasty guessing*, namely making quick but unreliable judgments based on salient yet coarse appearance patterns before capturing fine-grained motion evidence, for example, classifying a crossed-arm posture while overlooking a subtle but discriminative finger tap. Even worse, because standard paradigms only reward the final classification accuracy, as Figure 1 shows, the models are incentivized to hallucinate plausible-sounding but visually ungrounded reasoning traces. Moreover, most existing evaluations emphasize within-dataset Top-1 accuracy, leaving it unclear whether a model truly captures transferable micro-dynamics or merely memorizes dataset-specific appearance biases.

These critical limitations highlight an urgent need for a paradigm shift from purely label-driven recognition to evidence-based kinematic reasoning. To achieve this, the MLLM requires a strong inductive bias that explicitly forces the model to focus on localized temporal changes rather than static visual contexts. To this end, we propose **GMoT**, a lightweight and highly targeted **G**ated **M**otion-Aware **T**okenization module that distills sparse spatio-temporal cues into a compact motion-sensitive token sequence without relying on computationally heavy 3D convolutions or dense space-time attention that easily overfits to spatial biases. Specifically, GMoT establishes a spatial saliency scorer to spotlight action-relevant regions dynamically, computes explicit adjacent-frame temporal differences to isolate precise kinematic energy, and injects these cues into the MLLM’s visual stream via a conservatively initialized adaptive semantic gate. Furthermore, we argue that evaluating micro-gesture recognition requires moving beyond Top-1 accuracy within a single dataset. To improve the visual grounding of generated explanations, we build a semi-supervised annotation pipeline to generate Chain-of-Thought (CoT) descriptions focused on observable motion evidence. We then optimize the MLLM using a progressive reward-guided policy refinement strategy [40], guided by a reward formulation that penalizes lazy guessing and encourages anatomical region focus. To quantify one transparent aspect of these capabilities, we introduce Body-Region Grounding (BRG) Recall, a lexical anatomical-grounding proxy computed on correctly classified samples. We further establish an overlapping-label cross-domain transfer protocol between iMiGUE and SMG as a diagnostic of robustness under dataset shift.

Our contributions are summarized as follows:

- **Motion-aware tokenization.** We propose GMoT, which injects motion-salient pooling, adjacent-frame differencing, and gated fusion into an MLLM without disrupting its pretrained visual-language interface.
- **Reasoning-oriented training.** We construct semi-automatic CoT supervision and a progressive reward-guided refinement recipe that encourages label correctness, grounded explanations, and resistance to lazy appearance-based guessing.
- **Broader evaluation.** We introduce BRG Recall as a scoped anatomical-grounding proxy and an overlapping-label cross-domain transfer protocol between iMiGUE and SMG. GMoT reaches 67.32% / 73.11% Top-1 accuracy, respectively, and remains stronger than its Qwen3-VL backbone under the tested label-preserving corruptions.

2 Related work

2.1 Micro-gesture Recognition

Recent progress in micro-gesture recognition (MGR) has been largely shaped by emerging datasets and the evolution of modeling choices across modalities and learning paradigms [24, 25, 48]. RGB-based approaches typically adapt spatio-temporal backbones to capture subtle motion cues, while skeleton-based methods model joint dynamics with sequential models or graph networks [24, 26]. To handle highly similar classes and noisy observations, later work increasingly emphasizes stronger temporal attention and multi-modal fusion between appearance and pose signals [8, 17, 20, 38]. Despite these advances, most studies still cast the task as closed-set classification and therefore provide limited support for evidence-grounded micro-gesture reasoning. Broader surveys of generative affective modeling and human motion video generation further highlight the growing role of generative and language models in human-centered video analysis [36, 46, 49, 51].

2.2 MLLMs for Video Reasoning

MLLMs have recently extended from image understanding to video by combining vision encoders with language decoders and training on instruction-following data [2, 28, 53, 62, 64]. Video MLLMs combine vision encoders with language decoders under strict token budgets, so they often rely on sparse sampling, temporal pooling, or lightweight temporal adapters [2, 7, 28, 47, 54, 57, 61, 63]. These designs are effective for coarse video understanding, yet they often blur the brief localized dynamics that distinguish MG classes. Domain-specific systems have begun to explore LLMs for subtle behavioral understanding: AU-LLM maps fused 3D-CNN features to a compact visual token for micro-expression action-unit detection, while DeceptionX organizes fine-grained audiovisual evidence into interpretable deception reasoning [35, 58]. Recent systems further extend LLM-based multimodal learning to cross-modal physiological sensing and rationale-driven affective reasoning [14, 50], while Reflect-R1 grounds self-correction in retrieved visual evidence for long-video understanding [12]. Complementary continual-learning methods use predictive prompting and task-oriented generation to balance adaptation and retention across class increments [21, 22]. Although GMoT is not designed for class-incremental learning, these methods provide a broader view of controlled prompt- and representation-level adaptation. Nevertheless, most video MLLMs are optimized for answer correctness rather than evidence tracing, making it hard to produce reliable and interpretable decisions for subtle micro-gestures.

3 Methodology

In this section, we first introduce the preliminaries of multimodal large language models. We then present **GMoT**, a gated motion-aware token module that enables a vanilla MLLM to capture subtle temporal cues from micro-gesture videos by distilling frame-level visual tokens into compact motion representations. Finally, we describe a four-stage training strategy that progressively aligns classification, shapes the chain-of-thought pattern, and enhances micro-gesture reasoning via reward-guided policy refinement.

3.1 Preliminary

Multimodal large language models. Since the introduction of LLaVa [32], multimodal large language models (MLLMs) have attracted increasing attention from the community. Among recent advances, Qwen3-VL [3] has become a strong and widely adopted backbone because of its strong multimodal understanding capability. Following the LLaVa paradigm, the dense version of Qwen3-VL has three modules: a SigLIP-2 architecture vision encoder [43], an MLP-based aligner, and the Qwen3 LLM [52].

To process an image or a video, the vision encoder first converts each frame into a sequence of patch-level visual tokens. Given a video clip $V = \{I_t\}_{t=1}^T$, the encoder produces patch tokens $X_t \in \mathbb{R}^{N \times d_v}$ for each frame, and the aligner maps them into the language space as $Z_t \in \mathbb{R}^{N \times d}$.

With the LLaVa-style multimodal prompting, the aligned visual tokens are concatenated with textual tokens and fed into the LLM:

$$\mathbf{H} = \text{LLM}([\mathbf{Z}_1, \dots, \mathbf{Z}_T; \mathbf{E}(\mathbf{p})]), \quad (1)$$

where $\mathbf{p}$ is the input prompt and $\mathbf{E}(\cdot)$ denotes the text tokenizer. The model finally generates an output sequence $\hat{\mathbf{y}}$ in an autoregressive manner, which can be either a class name in a recognition task or a reasoning-augmented response in a reasoning scenario.

However, directly encoding patch tokens from frames may dilute subtle micro-gesture signals. The temporal variation of micro-gestures often exhibits low amplitude and short duration, which are easily overlooked during temporal modeling or diluted during visual token aggregation. This motivates us to distill time-sensitive information into an auxiliary representation that can be seamlessly injected into the MLLM.

3.2 Semi-automatic Reasoning Annotation

We construct reasoning-aware supervision with the semi-automatic pipeline illustrated in the left part of Figure 3. For each labeled clip, two independent LMMs generate candidate motion descriptions under prompts that suppress background and identity cues while emphasizing action-relevant body regions. Qwen3-VL-MoE [3] then selects the better candidate using class correctness, visual grounding, and non-hallucination, conditioned on a class-level reference description. Borderline cases are manually reviewed, and a human validation study based on sampled accepted descriptions is summarized in Appendix A. The resulting annotations supervise both label prediction and evidence-grounded CoT generation.

More concretely, for each clip, using two candidates improves diversity and reduces the risk of under-specified or hallucinated details from a single case. In prompts, we explicitly discourage any description of background, scene context, or identity-related facial

information. Instead, the models are required to focus on action-relevant regions, with primary attention to the hands and arms, and to describe only observable motion patterns that support the MG label. We additionally provide a class-level description from Deepseek [40] as a reference, summarizing typical visual patterns and key regions for each MG category. This serves as a semantic anchor to check class consistency and reduce label drift.

3.3 GMoT: Gated Motion-Aware Tokenization Module

While vanilla MLLMs are strong at general video understanding, they still struggle with fleeting and localized MG cues. As illustrated in Figure 2, **GMoT** addresses this mismatch by injecting explicit kinematic bias into the visual stream and distilling sparse spatio-temporal evidence into compact motion-sensitive tokens.

3.3.1 Motion-Prompted Spatial Saliency. Given the localized nature of micro-gestures, spatially homogeneous pooling will dilute the discriminative cues with irrelevant background noise. Instead of relying on the LLM to implicitly search for these tiny regions across dense patches, GMoT establishes a lightweight spatial scorer to explicitly spotlight motion-relevant areas. Specifically, we project each patch token from the vision encoder through a linear layer to obtain a relevance score. A spatial softmax is then applied to produce normalized attention weights, dynamically aggregating patch features into a frame-level representation via weighted summation. This mechanism functions as a learned spatial filter, aggressively amplifying localized evidence while suppressing visual redundancy.

3.3.2 Kinematics-Inspired Temporal Differencing. To capture the precise moment a micro-gesture occurs, the model must be sensitive to temporal gradients rather than static appearances. While standard video transformers rely on complex temporal mixing, we argue that explicit temporal differencing offers a more direct and robust inductive bias for subtle motion. Rather than treating this difference as a monolithic feature, we decouple it into a two-stream kinematic representation. We compute the mean magnitude to explicitly quantify motion energy, while retaining the signed difference to preserve the directional kinematics. By concatenating these decoupled streams and applying a learnable projection, we compress the raw temporal gradients into a fixed-length, dense motion token. This design forces the network to focus exclusively on visual changes, effectively isolating micro-dynamics from static human appearances.

3.3.3 Semantic-Preserving Adaptive Gating. A critical challenge in adapting pretrained MLLMs is that abruptly injecting new modality tokens can disrupt the aligned visual-semantic space. We therefore use a near-closed gated residual as a stabilization component rather than treating the gate itself as a standalone architectural novelty. Let z_t denote the aligned original visual feature at time t , g denote the learned gate parameter, and m_t denote the distilled motion token. We fuse them through

$$\tilde{z}_t = z_t + \sigma(g) \odot m_t, \quad (2)$$

where σ is the sigmoid function and $\odot$ denotes element-wise multiplication. The gate is initialized in its near-closed regime so that $\tilde{z}_t \approx z_t$ at the start of adaptation. This conservative starting point

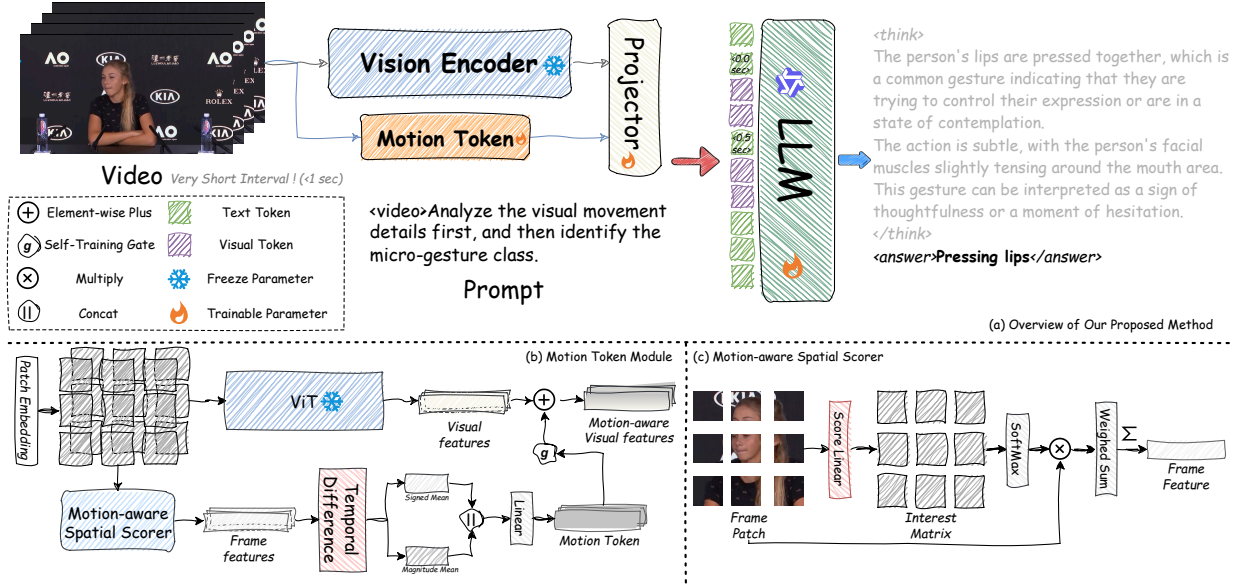


Figure 2: Overview of GMoT. GMoT extracts motion-sensitive evidence from short video clips and injects the resulting motion token into the MLLM through a gated fusion pathway for motion-aware reasoning. (a) presents the overall pipeline, where a compact motion token is derived from a short clip and incorporated into the MLLM together with visual and textual inputs. (b) illustrates the motion token module, which integrates motion-guided spatial saliency, adjacent-frame temporal differencing, and a learnable self-training gate. (c) shows the motion-aware spatial scorer, which predicts a soft importance map over frame patches and aggregates them into frame-level features.

preserves the pretrained representation while allowing the model to learn how much motion evidence to introduce. The contribution of GMoT lies in the task-specific composition of localized spatial scoring, explicit temporal differencing, and conservative residual injection before language reasoning.

3.4 Training Strategy

Building on the annotation pipeline and GMoT, we adopt a staged recipe that separates motion token learning, domain adaptation, reasoning format acquisition, and policy refinement, as shown in Figure 3. This decomposition avoids destructive interference, where sparse label-only supervision proves insufficient for newly introduced motion tokens. Specifically, we adopt a four-stage schedule.

3.4.1 Stage 0: GMoT Warm-up. We initialize from a pretrained MLLM and warm up the proposed GMoT module. During this stage, we freeze the LLM and the vision encoder and update only the GMoT module on general video datasets. Given a video clip and a prompt, the model is trained to predict the ground-truth label tokens under the standard language modeling loss,

$$\mathcal{L}_{\text{lm}} = - \sum_t \log \pi_{\theta}(s_t | x, s_{<t}), \quad (3)$$

which encourages GMoT to produce motion-aware tokens while keeping the pretrained model stable.

3.4.2 Stage 1: Video-Label Pairing. We then adapt the model to the micro-gesture domain with supervised fine-tuning on MG video and label pairs. In this stage, we allow the MLLM components to be trainable in the LoRA [19] setting and continue to optimize

the language modeling objective to align visual representations with MG-specific label semantics. This stage primarily improves domain adaptation and reduces the gap between generic video understanding and micro-gesture recognition.

3.4.3 Stage 2: CoT-format SFT. Next, we teach the model a consistent reasoning pattern using CoT-style annotations. We freeze the vision encoder to avoid drifting visual features and fine-tune the LLM together with GMoT to generate the target CoT text under the language modeling loss. The supervision is constructed to focus on observable motion evidence rather than background or identity cues, so the model learns an evidence-grounded reasoning path that will be further reinforced in Stage 3.

3.4.4 Stage 3: Reward-guided Policy Refinement. After Stage 2, the model already follows a structured reasoning format, but SFT still optimizes token-level likelihood rather than directly favoring globally better responses. We therefore perform a reward-guided policy refinement stage with grouped rollouts and a reference-regularized update objective [16, 40] to improve reasoning reliability and reduce failure cases on subtle micro-gestures.

3.4.5 Reward Formulation. Our Stage 3 reward combines label correctness, lazy-prediction penalty, format compliance, and observation-focus terms:

$$R_i = R_i^{\text{label}} + R_i^{\text{lazy}} + R_i^{\text{format}} + R_i^{\text{focus}} \quad (4)$$

Accuracy Reward (R_i^{label}). To prevent the MLLM from hallucinating non-existent micro-gesture concepts, we strictly constrain the output space. Let $\mathcal{C}$ denote the set of standard categories, c be the ground-truth class, and $\hat{c}_i$ be the predicted class parsed from

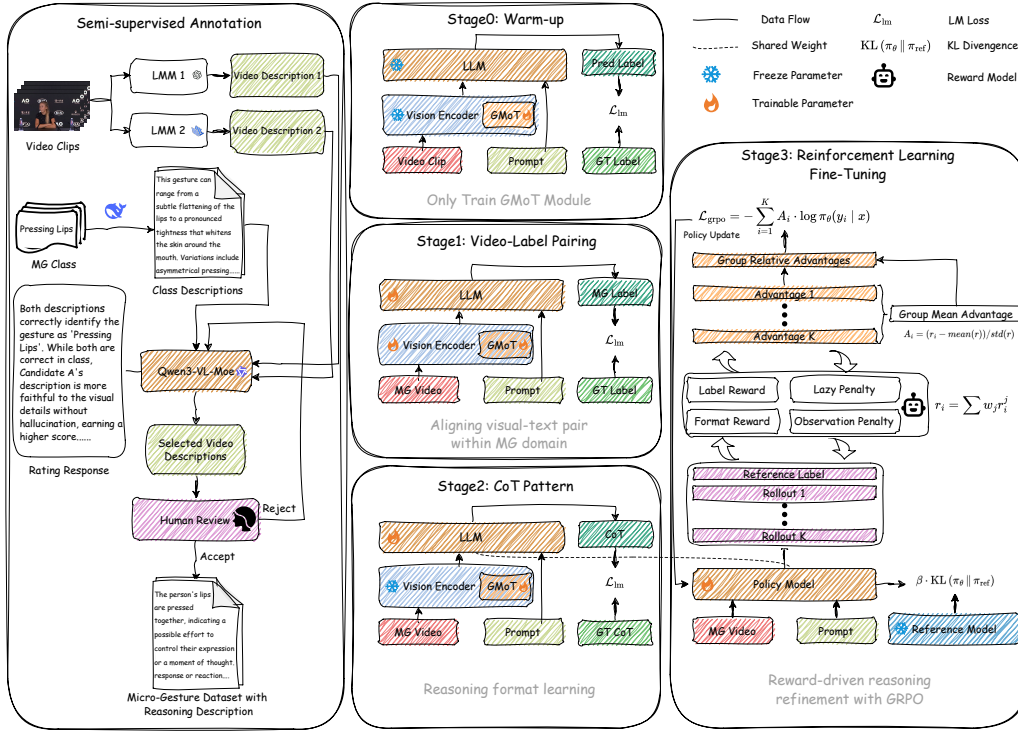


Figure 3: Overview of the four-stage training schedule. The training process progressively moves through GMoT warm-up, micro-gesture domain adaptation, CoT-style reasoning learning, and finally reward-guided reinforcement fine-tuning for reasoning refinement.

the `<answer>` field. The label reward applies a large positive signal for exact matches, a standard penalty for incorrect guesses, and a severe penalty for out-of-domain hallucinations:

$$R_i^{\text{label}} = \begin{cases} +3.0, & \hat{c}_i = c \\ -1.0, & \hat{c}_i \neq c \wedge \hat{c}_i \in C \\ -2.0, & \hat{c}_i \notin C \end{cases} \quad (5)$$

Lazy-Prediction Penalty (R^{lazy}). During early training stages, the model tends to exploit the long-tailed distribution by repeatedly guessing high-frequency head classes such as *Illustrative hand gestures*. Let $\mathcal{H} \subset C$ denote the predefined high-frequency set of such classes. We penalize incorrect predictions that fall into this set:

$$R_i^{\text{lazy}} = \begin{cases} -0.5, & \hat{c}_i \neq c \wedge \hat{c}_i \in \mathcal{H} \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

Formatting Reward (R^{format}). We encourage the completion to follow the CoT template:

$$R_i^{\text{format}} = \delta_{\text{think}} + \delta_{\text{answer}} \quad (7)$$

where $\delta_{\text{think}} = 0.5$ if the `<think>` block is present and -0.5 otherwise, while $\delta_{\text{answer}} = 0$ if the `<answer>` block is well formed and -1.5 if it is missing.

Observation Penalty (R^{focus}). We penalize background-oriented hallucination in the reasoning trace. If the `<think>` field contains

irrelevant environmental keywords $\mathcal{K}_{\text{bg}}$, then

$$R_i^{\text{focus}} = \begin{cases} -1.0, & \exists k \in \mathcal{K}_{\text{bg}} \text{ appearing in } \text{<think>} \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

Group rollouts and loss. For each input x , we sample K completions, compute group-normalized advantages

$$A_i = \frac{r_i - \mu}{\sigma + \epsilon}, \quad \mu = \frac{1}{K} \sum_{i=1}^K r_i, \quad \sigma = \sqrt{\frac{1}{K} \sum_{i=1}^K (r_i - \mu)^2}, \quad (9)$$

and optimize

$$\mathcal{L}_{\text{stage3}} = -\frac{1}{K} \sum_{i=1}^K A_i \log \pi_{\theta}(y_i | x) + \beta D_{\text{KL}}(\pi_{\theta}(\cdot | x) \| \pi_{\text{ref}}(\cdot | x)), \quad (10)$$

where the reference policy π_{ref} is adapted from Stage 2. Appendix B provides the full keyword sets and auxiliary details.

Each reward term targets a specific failure mode that we repeatedly observed in MG reasoning. R^{label} keeps the model inside the valid label space, R^{lazy} focuses on long-tail classes, R^{format} stabilizes the RL template, and R^{focus} suppresses background-oriented hallucinations. Together with group-normalized advantages, Stage 3 optimizes relative response quality within each prompt rather than raw token regression, while the KL term preserves the grounded reasoning behavior already established in Stage 2.

Table 1: Top-1 accuracy on iMiGUE [33] and SMG [11].

Method	Model Type	iMiGUE	SMG
		Top-1 Acc. (%) $\uparrow$	Top-1 Acc. (%) $\uparrow$
C3D [42]	CNN	30.13	36.07
I3D [6]	CNN	35.08	28.36
TSN [44]	CNN	51.31	50.49
TSM [31]	CNN	56.23	58.69
MA-Net [18]	CNN	58.08	48.69
P&C [41]	Encoder-Decoder	31.67	39.30
U-S-VAE [33]	Encoder-Decoder	32.43	39.59
Timesformer [5]	Transformer	50.66	47.89
Video SwinT [34]	Transformer	60.11	48.03
UniformerV2 [29]	Transformer	59.91	46.07
VideoMamba [27]	SSM	58.13	53.28
MSF-Mamba-T ⁺ [24]	SSM	58.33	52.29
MSF-Mamba-M ⁺ [24]	SSM	62.98	60.13
GPT-4o [37]	MLLM	56.89	60.71
Gemini 2.5 Pro [15]	MLLM	58.41	62.57
InternVL2.5-8B [13]	MLLM	59.25	56.21
Video-LLaMA2 [55]	MLLM	39.74	51.80
TinyLLaVA-Video [59]	MLLM	57.82	52.87
Affect-GPT [30]	MLLM	58.45	54.12
Qwen2.5-VL-7B [4]	MLLM	59.71	53.27
Qwen3-VL-2B [3]	MLLM	51.25	20.98
Qwen3-VL-8B [3]	MLLM	60.52	70.00
Qwen2.5-VL-7B+ GMoT	MLLM	63.23	60.85
Qwen3-VL-2B+ GMoT	MLLM	58.51	68.85
Qwen3-VL-8B+ GMoT	MLLM	67.32	73.11

4 Experiments

4.1 Implementation Details

We build our method on Qwen3-VL [3] and all stages are implemented with `ms-swift` [60] and `DeepSpeed ZeRO-3` [39] on NVIDIA H100 GPUs. Since most video clips in the datasets are shorter than 3 seconds, we sample each clip at 4 FPS with at most 12 frames, cap the pixels per frame at 786,432, and limit visual tokens to 896. Appendix B provides the complete training details.

4.2 Quantitative Results

Intra-dataset testing. Table 1 compares our **GMoT**-enhanced MLLM with representative micro-gesture recognition approaches on iMiGUE and SMG datasets. All trainable open-source MLLM baselines use the same iMiGUE and SMG training splits. CNN-based backbones such as C3D [42] and I3D [6] perform poorly, suggesting that purely local spatio-temporal convolutions are insufficient to capture subtle and instantaneous cues. Stronger CNN variants like TSN [44], TSM [31], and MA-Net [18] improve steadily, yet still lag behind stronger temporal modeling on fine-grained MG settings. Transformer-based models benefit from global attention, but they remain limited on SMG, where weak motion evidence is easily overwhelmed by irrelevant appearance variation. Unsupervised learning methods, e.g., P&C [41] and U-S-VAE [33], benefit from label-free training but typically achieve suboptimal accuracy.

State-space models offer a stronger trade-off between accuracy and efficiency, with the VideoMamba [27] and MSF-Mamba [24] variants consistently outperforming most CNN and Transformer baselines. On top of this landscape, MLLMs provide a competitive alternative by combining strong visual encoders with language-centric reasoning priors. With significant pretraining data, Qwen2.5-VL-7B [4] reaches 59.71% on iMiGUE and 53.27% on SMG, while

Qwen3-VL-8B [3] further improves to 60.52% and 70.00%. Gemini 2.5 Pro [15] and InternVL2.5-8B [13] provide additional closed- and open-model reference points. GMoT improves Qwen2.5-VL-7B to 63.23% and 60.85%, and Qwen3-VL-8B to 67.32% and 73.11%, showing positive portability across backbones and model scales. Appendix C provides the extended per-family discussion.

For the transfer and reasoning analyses below, we focus on reasoning-capable MLLM backbones evaluated under a unified prompting interface because matched transfer settings and rationale outputs are not currently available for the traditional non-MLLM baselines in Table 1.

Cross-domain transfer. Table 2 reports the transfer results in two directions. Qwen3-VL-8B+**GMoT** achieves the best Accuracy and Macro-F1 in both directions. The metric ranking is nevertheless mixed: vanilla Qwen3-VL-8B retains a slightly higher Weighted-F1 on iMiGUE $\rightarrow$ SMG (0.8130 vs. 0.8053) and SMG $\rightarrow$ iMiGUE (0.6762 vs. 0.6720). These decreases are 0.0077 and 0.0042, respectively, while Accuracy and Macro-F1 improve in both directions. We therefore treat this experiment as an overlapping-label transfer diagnostic rather than broad proof of cross-domain generalization.

The iMiGUE $\rightarrow$ SMG target contains only 22 samples, so one additional correct prediction changes Accuracy by 4.55 points. Qwen3-VL-8B + GMoT reaches 86.36% Accuracy (19/22), compared with 81.82% (18/22) for the vanilla backbone. The target is also highly skewed, with 15 Folding arms, 4 Playing or adjusting hair, 2 Touching or covering suprasternal notch, and 1 Touching or scratching neck samples. We therefore report Accuracy, Macro-F1, and Weighted-F1 together and avoid statistical-strength claims from this split. On SMG $\rightarrow$ iMiGUE, the small Weighted-F1 fluctuation is concentrated in a class with only three test samples, further motivating per-class and class-support-aware interpretation.

Qwen2.5-VL-7B+**GMoT** remains a strong baseline. It raises iMiGUE $\rightarrow$ SMG Accuracy to 77.27% and Weighted-F1 to 0.7786, but it still trails Qwen3-VL-8B+**GMoT** on both metrics in that direction; on SMG $\rightarrow$ iMiGUE it reaches 52.24% Accuracy, and both Macro-F1 and Weighted-F1 remain lower. The comparison with Qwen2.5-VL-7B (w/o GMoT) further highlights the contribution of the motion token branch: removing GMoT reduces iMiGUE $\rightarrow$ SMG Accuracy from 77.27% to 50.00% and Weighted-F1 from 0.7786 to 0.5578, while SMG $\rightarrow$ iMiGUE Weighted-F1 drops from 0.5378 to 0.5034.

The task-specific MM-Gesture RGB baseline [17] also drops sharply under the same direct-transfer splits, which shows that the diagnostic is challenging for a dedicated recognition model as well as general MLLMs. Within this restricted protocol, GMoT improves accuracy and Macro-F1 over its Qwen3-VL backbone, but the mixed Weighted-F1 results and small target support remain explicit limitations. Appendix C provides the split audit and extended protocol.

Additional repeat-run stability and controlled corruption results are reported in Appendix C.2.

Analysis of Reasoning Ability. An anatomically grounded micro-gesture rationale should mention the body region relevant to the predicted class. We therefore define **Body-Region Grounding (BRG) Recall** by mapping each class c to a region-specific keyword set K_c and checking whether the generated reasoning trace

Table 2: Cross-domain transfer results on iMiGUE [33] and SMG [11].

Model	iMiGUE $\rightarrow$ SMG ($N = 22$)			SMG $\rightarrow$ iMiGUE ($N = 312$)		
	Acc. (%)	Macro-F1	Weighted-F1	Acc. (%)	Macro-F1	Weighted-F1
Affect-GPT [30]	4.55	0.0217	0.0040	5.13	0.0195	0.0050
Video-LLaMA2 [55]	4.55	0.0235	0.0053	0.00	0.0000	0.0000
MM-Gesture RGB [17]	13.64	0.0952	0.1558	21.79	0.1426	0.2217
Qwen2.5-VL-7B [4]	50.00	0.3743	0.5578	50.00	0.2894	0.5034
Qwen3-VL-8B [3]	81.82	0.3325	0.8130	61.86	0.3335	0.6762
Qwen2.5-VL-7B+ GMoT	77.27	0.3457	0.7786	52.24	0.2720	0.5378
Qwen3-VL-8B+ GMoT	86.36	0.4257	0.8053	62.18	0.3371	0.6720

Table 3: Reasoning evaluation on iMiGUE [33] after policy refinement. BRG is the body-region keyword-overlap rate conditioned on a correct prediction; Avg. Len. is a diagnostic rather than an optimization target.

Method	Acc. (%) $\uparrow$	BRG $\uparrow$	Avg. Len.
GPT-4o [37]	56.89	56.21	30.91
Qwen3-VL-8B [3]	62.96	96.82	45.08
Qwen3-VL-8B+GMoT	67.32	97.70	64.84

T_i overlaps with K_c :

$$BRG = \frac{1}{N_{correct}} \sum_{i=1}^{N_{correct}} \mathbb{I}(K_c \cap T_i \neq \emptyset) \quad (11)$$

where $\mathbb{I}$ is the indicator function. BRG is computed only over correctly classified samples, so it is a lexical anatomical-grounding proxy rather than a general measure of causal reasoning faithfulness. We report *Avg. Length*, the average number of words inside the <think> block, alongside BRG to expose the potential verbosity confound.

Table 3 shows that both Qwen3-VL variants attain high BRG among their correct predictions. GMoT improves accuracy from 62.96% to 67.32% after the same policy-refinement stage, while BRG changes from 96.82% to 97.70% and average rationale length increases from 45.08 to 64.84 words. Because a longer rationale has more opportunities to contain a body-region keyword, the length increase is reported as a confound and is not itself treated as evidence of better reasoning. A fully matched-length BRG comparison is not available from the current logs.

We also move the blind human check into the main evaluation discussion. On 181 accepted SMG descriptions, both annotators agreed that 137 were visually grounded. For severe hallucination, 44 samples were flagged by both annotators and one additional sample was flagged by one annotator, yielding 89 positive votes out of 362. Under a conservative item-level rule that treats the single disagreement as positive, the severe-hallucination count is 45/181 (24.86%; 95% Wilson CI: 19.13%–31.64%). This human check supports the use of the annotations as intermediate supervision but does not make them gold-standard rationales. Appendix A gives the sampling and agreement details.

4.3 Qualitative Analysis

To better understand why **GMoT** improves micro-gesture reasoning, we qualitatively compare the generated rationales and predictions across representative MLLMs. Table 4 shows a challenging

clip labeled as *Shake double shoulders*, where the discriminative motion is subtle and localized around the shoulder region, while the overall posture contains a strong yet misleading cue of folded arms. As shown in Table 4, general-purpose MLLMs often default to salient appearance cues and produce rationales that are temporally insensitive. Instead of isolating the crucial shoulder dynamics, they describe the clip as largely unchanged over time and consequently predict visually plausible but motion-agnostic classes such as *Holding arms*, *Touching arm*, or *Crossing arms*.

In contrast, our method produces a motion-grounded rationale that explicitly attributes the decision to shoulder dynamics, such as the raised and slightly hunched shoulders, and correctly predicts *Shake double shoulders*. This is consistent with our design goal, which distills motion-relevant evidence into a compact token sequence before temporal modeling. By emphasizing regions that change over time and suppressing static background cues, GMoT reduces distraction from dominant but non-discriminative appearance signals, yielding more motion-aligned explanations in this qualitative case. Appendix D provides additional examples and limitations.

Figure 4 provides a qualitative comparison between models with and without GMoT on a representative ‘*Moving torso*’ sample. Compared with the model without GMoT, the model equipped with GMoT produces more concentrated visual responses around the upper-body motion region while reducing scattered activations on irrelevant background areas. The response difference map further indicates that GMoT mainly amplifies motion-relevant cues, suggesting its effectiveness in guiding the model toward subtle regions.

4.4 Ablation Study

In the ablation study, we ask whether the four-stage strategy is necessary and quantify the contribution of each GMoT component. The secondary Video-LLaMA2 transfer check is discussed in Appendix C.

4.4.1 Stage-wise ablation. Table 6 progressively enables each stage while keeping the backbone, sampling protocol, and optimization budget comparable. Video-label pairing alone slightly reduces accuracy to 59.85%, CoT-format SFT recovers it to 61.19%, and policy refinement then reaches 67.32%. The larger gain from refinement after reasoning-aware initialization supports the staged training design. Appendix C provides the extended interpretation.

Table 4: Qualitative reasoning comparison on a challenging *Shake double shoulders* clip. Incorrect reasoning is highlighted in red and the correct answer is highlighted in green.

Video Sample	
	Ground Truth: Shake double shoulders
ChatGPT	<think> Both arms are crossed in front of the torso, with one hand gripping the opposite upper arm/shoulder in a self-hug posture; no clear face/neck/leg-focused action. </think> <answer>Holding arms</answer>
Gemini	<think>I observe the person's right hand moving across her body to touch and rest upon her left upper arm.</think> <answer>Touching arm</answer>
Qwen	<think>The video shows a woman sitting at a press conference table. Her posture is static throughout the clip. The most prominent feature is her arm position: her arms are folded across her chest. Her right hand is grasping her left upper arm, and her left arm is tucked underneath, holding her right arm. </think> <answer>Crossing arms</answer>
GMoT (Ours)	<think>The person's shoulders are raised and slightly hunched, indicating a possible feeling of discomfort or unease. The head is slightly tilted forward, and the upper body appears tense, with the arms crossed in front. This posture can be associated with a need to self-soothe or a moment of contemplation during the conversation. </think> <answer>Shake double shoulders</answer>

Table 5: Transfer study on Video-LLaMA2.

Method	iMiGUE	SMG
VideoLLaMA2-Qwen2.5-7B	39.74	51.80
+ Motion Token	39.52	50.33
+ Motion Token + CoT mixed training	42.75	52.13

Table 6: Stage-wise ablation on iMiGUE.

Setting	Acc. (%)
S0: Qwen3-VL SFT w/o GMoT	60.52 (0.00)
S1: S0 + Policy Refinement	62.96 (+2.44)
S2: GMoT warm-up + Video-Label Pairing	59.85 (-0.67)
S3: S2 + CoT-format SFT	61.19 (+0.67)
S4: S3 + Policy Refinement	67.32 (+6.80)

Table 7: GMoT component ablation on iMiGUE.

Setting	Acc. (%)
Full GMoT (Stage 2)	61.19
w/o Spatial Scorer	56.94 (-4.25)
w/o Temporal Diff	55.65 (-5.54)
w/o Semantic Gate	56.28 (-4.91)



Figure 4: Qualitative visualization on a *Moving torso* sample. Compared with w/o GMoT, the model with GMoT shows more concentrated visual responses on the head-torso region. GMoT Focus line denotes the response difference map between models.

5 Limitations

The iMiGUE→SMG diagnostic has only 22 target samples and a highly skewed class distribution, so its result should not be generalized to broad cross-domain robustness. BRG is a lexical body-region proxy conditioned on correct predictions and can be affected by rationale length; it does not establish causal faithfulness. The sampled annotation audit is category-skewed and leaves a non-negligible

residual grounding and hallucination error rate. The corruption study covers three controlled operators rather than unconstrained real-world degradation, and adjacent-frame differencing is not a complete model of long-range temporal dependency. Future work should evaluate larger aligned cross-domain sets, matched-length grounding metrics, and broader natural corruptions.

6 Conclusion

We propose GMoT, a motion-aware tokenization module that distills subtle kinematics into a compact sequence before language reasoning. Coupled with semi-automatic rationale supervision and reward-guided refinement, GMoT improves in-domain recognition on iMiGUE and SMG, retains gains under the tested corruptions, and improves Accuracy and Macro-F1 in a restricted overlapping-label transfer diagnostic. Its generated rationales maintain high lexical anatomical grounding, while the explicit limitations of BRG, annotation quality, and small-split transfer remain part of the reported evidence.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62576076), the CCF-Tencent Rhino-Bird Open Research Fund, and the Guangdong Research Team for Communication and Sensing Integrated with Intelligent Computing (Project No. 2024KCXTD047). Computational resources were provided by the SongShan Lake HPC Center (SSL-HPC) at Great Bay University.

References

- [1] Roger E. Axtell and Mike Fornwald. 1998. *Gestures: The Do's and Taboos of Body Language Around the World*. John Wiley & Sons, New York.
- [2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv:2308.12966 [cs.CV]
- [3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaoai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibao Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. 2025. Qwen3-VL Technical Report. arXiv:2511.21631 [cs.CV] <https://arxiv.org/abs/2511.21631>
- [4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaoai Li, Jianqiang Wan, 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>
- [5] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is Space-Time Attention All You Need for Video Understanding? arXiv:2102.05095 [cs.CV] <https://arxiv.org/abs/2102.05095>
- [6] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.
- [7] Boyu Chen, Zhengrong Yue, Siran Chen, Zikang Wang, Yang Liu, Peng Li, and Yali Wang. 2025. Lvagent: Long video understanding by multi-round dynamical collaboration of mllm agents. arXiv preprint arXiv:2503.10200 (2025).
- [8] Guoliang Chen, Fei Wang, Kun Li, Zhiliang Wu, Hehe Fan, Yi Yang, Meng Wang, and Dan Guo. 2024. Prototype learning for micro-gesture classification. In *Proceedings of the IJCAI 2024 Workshop and Challenge on Micro-gesture Analysis for Hidden Emotion Understanding (CEUR Workshop Proceedings, Vol. 3848)*. https://ceur-ws.org/Vol-3848/paper_3.pdf
- [9] Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Juntong Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, and Limin Wang. 2023. VideoLLM: Modeling Video Sequence with Large Language Models. arXiv:2305.13292 [cs.CV] <https://arxiv.org/abs/2305.13292>
- [10] Haoyu Chen, Xin Liu, Xiaobai Li, Henglin Shi, and Guoying Zhao. 2019. Analyze Spontaneous Gestures for Emotional Stress State Recognition: A Micro-gesture Dataset and Analysis with Deep Learning. In *2019 14th IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 1–8.
- [11] 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.
- [12] Shuimu Chen, Yuteng Chen, Yuanshen Guan, Zebang Cheng, Zeyu Zhang, Shengqian Qin, Bin Xia, Jiaran Li, Wenming Yang, and Fei Ma. 2026. Reflect-R1: Evidence-Driven Reflection for Self-Correction in Long Video Understanding. In *European Conference on Computer Vision*. arXiv:2606.27922 [cs.CV] doi:10.48550/arXiv.2606.27922
- [13] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. arXiv:2412.05271 [cs.CV] doi:10.48550/arXiv.2412.05271
- [14] Zebang Cheng, Shuimu Chen, Boxue Yang, Yuanshen Guan, Jingyi Chen, Zheng Lian, Xiaojiang Peng, Fei Ma, Laizhong Cui, and Qi Tian. 2026. OmniOPSD: Rationale-Privileged On-Policy Self-Distillation for Affective Computing. arXiv preprint arXiv:2606.15920 (2026). arXiv:2606.15920 [cs.CV] doi:10.48550/arXiv.2606.15920
- [15] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, et al. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. arXiv:2507.06261 [cs.CL] doi:10.48550/arXiv.2507.06261
- [16] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. 2025. Video-R1: Reinforcing Video Reasoning in MLLMs. arXiv:2503.21776 [cs.CV] <https://arxiv.org/abs/2503.21776>
- [17] Jihao Gu, Fei Wang, Kun Li, Yanyan Wei, Zhiliang Wu, and Dan Guo. 2025. MM-Gesture: Towards Precise Micro-Gesture Recognition through Multimodal Fusion. In *Proceedings of the IJCAI-2025 Workshop and Challenge on Human Behavior Analysis for Emotion Understanding (CEUR Workshop Proceedings, Vol. 4168)*. https://ceur-ws.org/Vol-4168/paper_2.pdf
- [18] 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.
- [19] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685 [cs.CL] <https://arxiv.org/abs/2106.09685>
- [20] Hexiang Huang, Yuhang Wang, Kerui Linghu, and Zhaoqiang Xia. 2024. Multi-modal micro-gesture classification via multiscale heterogeneous ensemble network. In *Proceedings of the IJCAI 2024 Workshop and Challenge on Micro-gesture Analysis for Hidden Emotion Understanding (CEUR Workshop Proceedings, Vol. 3848)*. https://ceur-ws.org/Vol-3848/paper_2.pdf
- [21] Libo Huang, Xiangqi Li, Jiarui Zhao, Zhulin An, Chuanguang Yang, Boyu Diao, Fei Wang, Yan Zeng, Zhifeng Hao, and Yongjun Xu. 2026. PrePrompt: Predictive Prompting for Class-Incremental Learning. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.
- [22] Libo Huang, Yan Zeng, Chuanguang Yang, Zhulin An, Boyu Diao, and Yongjun Xu. 2024. eTag: Class-Incremental Learning via Embedding Distillation and Task-Oriented Generation. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 11 (2024), 12591–12599. doi:10.1609/aaai.v38i11.29153
- [23] Deng Li, Xin Liu, Bohao Xing, Baiqiang Xia, Yuan Zong, Bihan Wen, and Heikki Kälviäinen. 2024. EALD-MLLM: Emotion Analysis in Long-sequential and De-identity videos with Multi-modal Large Language Model. arXiv:2405.00574 [cs.CV] doi:10.48550/arXiv.2405.00574
- [24] Deng Li, Jun Shao, Bohao Xing, Rong Gao, Bihan Wen, Heikki Kälviäinen, and Xin Liu. 2026. MSF-Mamba: Motion-Aware State Fusion Mamba for Efficient Micro-Gesture Recognition. *IEEE Transactions on Multimedia* (2026), 1–12. doi:10.1109/TMM.2026.3668511
- [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*. Association for Computing Machinery, 5707–5716. doi:10.1145/3746027.3755411
- [26] Kun Li, Dan Guo, Guoliang Chen, Xinge Peng, and Meng Wang. 2023. Joint Skeletal and Semantic Embedding Loss for Micro-gesture Classification. arXiv:2307.10624 [cs.CV] <https://arxiv.org/abs/2307.10624>
- [27] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. 2024. Videomamba: State space model for efficient video understanding. In *European conference on computer vision*. Springer, 237–255.
- [28] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22195–22206.
- [29] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. 2023. Uniformerv2: Unlocking the potential of image vits for video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1632–1643.
- [30] Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, Jiangyan Yi, and Jianhua Tao. 2025. AffectGPT: A New Dataset, Model, and Benchmark for Emotion Understanding with Multimodal Large Language Models. In *Proceedings of the 42nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 267)*. PMLR, 36993–37014. <https://proceedings.mlr.press/v267/lian25a.html>
- [31] Ji Lin, Chuang Gan, and Song Han. 2019. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*. 7083–7093.
- [32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. arXiv:2304.08485 [cs.CV] <https://arxiv.org/abs/2304.08485>
- [33] 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.
- [34] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2021. Video Swin Transformer. arXiv:2106.13230 [cs.CV] <https://arxiv.org/abs/2106.13230>
- [35] Zhishu Liu, Kaishen Yuan, Bo Zhao, Hui Ma, and Zitong Yu. 2026. AULLM++: Structured-Token-Conditioned Large Language Models for Micro-Expression Action Unit Detection. arXiv preprint arXiv:2603.08387 (2026).
- [36] Fei Ma, Yucheng Yuan, Yifan Xie, Hongwei Ren, Ivan Liu, Ying He, Fuji Ren, Fei Richard Yu, and Shiguang Ni. 2025. Generative Technology for Human Emotion Recognition: A Scoping Review. *Information Fusion* 115 (2025), 102753. doi:10.1016/j.inffus.2024.102753
- [37] OpenAI. 2024. GPT-4o System Card. arXiv:2410.21276 [cs.CL] doi:10.48550/arXiv.2410.21276
- [38] 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. In *Proceedings of the IJCAI-2025 Workshop and Challenge on Human Behavior Analysis for Emotion Understanding (CEUR Workshop Proceedings, Vol. 4168)*. https://ceur-ws.org/Vol-4168/paper_11.pdf

- [39] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. ZeRO: Memory Optimizations Toward Training Trillion Parameter Models. arXiv:1910.02054 [cs.LG] <https://arxiv.org/abs/1910.02054>
- [40] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeek-Math: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300 [cs.CL] <https://arxiv.org/abs/2402.03300>
- [41] Kun Su, Xiulong Liu, and Eli Shlizerman. 2020. Predict & cluster: Unsupervised skeleton based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9631–9640.
- [42] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.
- [43] Michael Tschanen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. arXiv:2502.14786 [cs.CV] <https://arxiv.org/abs/2502.14786>
- [44] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2018. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence* 41, 11 (2018), 2740–2755.
- [45] Tao Wang, Xue Lin, Yixing Xu, Qian Ye, Dan Guo, Sergio Escalera, George Khoriba, and Zhiyong Yu. 2026. Micro-gesture recognition: A comprehensive survey of datasets, methods, and challenges. *Machine Intelligence Research* 23, 2 (2026), 308–331. doi:10.1007/s11633-025-1629-x
- [46] Yelin Wang, Zijia Song, Chuanguang Yang, Miaoyu Wang, Zhulin An, Libo Huang, and Yongjun Xu. 2026. DFM: Difference Feature Modeling with Text-Guided Gated Contrastive Loss for Remote Sensing Image Change Captioning. *arXiv preprint arXiv:2606.27410* (2026).
- [47] Yelin Wang, Zijia Song, Shuo Ye, Chuanguang Yang, Miaoyu Wang, Yong Xu, Zhulin An, Yongjun Xu, and Zitong Yu. 2026. RSICCLLM: A Multimodal Large Language Model for Remote Sensing Image Change Captioning. *arXiv preprint arXiv:2606.28266* (2026).
- [48] Zeheng Wang, Zitong Yu, Yijie Zhu, Bo Zhao, Haochen Liang, Taorui Wang, Wei Xia, Jiayu Zhang, Zhishu Liu, Hui Ma, Fei Ma, and Qi Tian. 2026. AffectAgent: Collaborative Multi-Agent Reasoning for Retrieval-Augmented Multimodal Emotion Recognition. arXiv:2604.12735 [cs.CV] <https://arxiv.org/abs/2604.12735>
- [49] Zeheng Wang, Bo Zhao, Yijie Zhu, Zhishu Liu, Hui Ma, Ruixin Zhang, Shouhong Ding, Qianyu Xie, and Zitong Yu. 2026. Navigating the Emotion Tree: Hierarchical Hyperbolic RAG for Multimodal Emotion Recognition. arXiv:2605.18884 [cs.LG] <https://arxiv.org/abs/2605.18884>
- [50] Yiping Xie, Bo Zhao, Mingtong Dai, Jian-Ping Zhou, Yue Sun, Tao Tan, Weicheng Xie, Linlin Shen, and Zitong Yu. 2026. PhysLLM: Harnessing Large Language Models for Cross-Modal Remote Physiological Sensing. In *The Fourteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=aR43t8OEeW>
- [51] Haiwei Xue, Xiangyang Luo, Zhanghao Hu, Xin Zhang, Xunzhi Xiang, Yuqin Dai, Jianzhuang Liu, Zhensong Zhang, Minglei Li, Jian Yang, Fei Ma, Zhiyong Wu, Changpeng Yang, Zonghong Dai, and Fei Richard Yu. 2025. Human Motion Video Generation: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47, 11 (2025), 10709–10730. doi:10.1109/TPAMI.2025.3594034
- [52] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. arXiv:2505.09388 [cs.CL] <https://arxiv.org/abs/2505.09388>
- [53] Qilang Ye, Wei Zeng, Meng Liu, Jie Zhang, Yupeng Hu, Zitong Yu, and Yu Zhou. 2026. When Eyes and Ears Disagree: Can MLLMs Discern Audio-Visual Confusion?. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 11955–11963.
- [54] Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, et al. 2025. TimeSuite: Improving MLLMs for Long Video Understanding via Grounded Tuning. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=nAvejJURqZ>
- [55] Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Singapore, 543–553. doi:10.18653/v1/2023.emnlp-demo.49
- [56] Jiayu Zhang, Xun Lin, Jiajian Huang, Shuo Ye, Xiaobao Guo, Dongliang Zhu, Ruimin Hu, Dan Guo, Yanyan Liang, Zitong Yu, and Xiaochun Cao. 2026. Multimodal Deception Detection: A Survey. *Machine Intelligence Research* 23, 2 (2026), 284–307. doi:10.1007/s11633-025-1625-x
- [57] Jie Zhang, Qilang Ye, Hao Zhou, Haochen Liang, and Fei Luo. 2026. MAVIS: Multi-Agent Video Retrieval via Structured Video Understanding. In *Findings of the Association for Computational Linguistics: ACL 2026*. 21751–21764.
- [58] Jiayu Zhang, Shuo Ye, Jiajian Huang, Yawen Cui, Taorui Wang, Wei Xia, Zeheng Wang, Haowen Tang, Hui Ma, and Zitong Yu. 2026. DeceptionX: Explainable Deception Detection with Multimodal Large Language Models. *arXiv preprint arXiv:2606.11385* (2026). arXiv:2606.11385 [cs.CV] doi:10.48550/arXiv.2606.11385
- [59] Xingjian Zhang, Xi Weng, Yihao Yue, Zhaoxin Fan, Wenjun Wu, and Lei Huang. 2025. TinyLLaVA-Video: Towards Smaller LLMs for Video Understanding with Group Resampler. arXiv:2501.15513 [cs.CV] <https://arxiv.org/abs/2501.15513>
- [60] Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Hong Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. 2025. SWIFT: A Scalable lightWeight Infrastructure for Fine-Tuning. arXiv:2408.05517 [cs.CL] <https://arxiv.org/abs/2408.05517>
- [61] Yijie Zhu, Jie He, Rui Shao, Kaishen Yuan, Tao Tan, Xiaochen Yuan, and Zitong Yu. 2026. AVL: Prior-Guided Vision-Language-Action Models via World Knowledge Variation. *arXiv preprint arXiv:2603.08361* (2026).
- [62] Yijie Zhu, Yibo Lyu, Zitong Yu, Rui Shao, Kaiyang Zhou, and Liqiang Nie. 2025. EmoSym: A Symbiotic Framework for Unified Emotional Understanding and Generation via Latent Reasoning. In *Proceedings of the 33rd ACM International Conference on Multimedia*.
- [63] Yijie Zhu, Rui Shao, Ziyang Liu, Jie He, Jizhihui Liu, Jiuru Wang, and Zitong Yu. 2026. H-GAR: A Hierarchical Interaction Framework via Goal-Driven Observation-Action Refinement for Robotic Manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- [64] Yijie Zhu, Lingsen Zhang, Zitong Yu, Rui Shao, Tao Tan, and Liqiang Nie. 2025. UniEmo: Unifying Emotional Understanding and Generation with Learnable Expert Queries. *arXiv preprint arXiv:2507.23372* (2025).