Amplify, Aggregate, and Adjust: VideoMAE-based Holistic-Subtle Aggregation for Micro-Action Recognition

Yan Zhang[1,2,3,4], Nan Pu[1✉], Wenjing Li[1], Zhun Zhong[1✉], and Meng Wang[1]

[1] Hefei University of Technology, Hefei, China
[2] Jianghuai Advance Technology Center, Hefei, China
[3] Anhui Provincial Key Laboratory of Humanoid Robots, Hefei, China
[4] Anhui Provincial Industry Innovation Center of Humanoid Robots, Hefei, China

Abstract. Micro-Action Recognition (MAR) aims to identify transient and subtle bodily movements that occur during interpersonal communication. Existing MAR methods attempt to emphasize low-amplitude motion by injecting skeletal priors and have achieved certain progress. However, they (1) rely on additional modalities and costly pose annotations, and (2) employ coarse joint graphs that fail to capture sub-joint micro-movements and fine-grained appearance cues (*e.g.*, finger tremors). In contrast, raw RGB videos preserve such subtle cues, and pretrained VideoMAE provides strong spatiotemporal priors without requiring additional supervision. Motivated by this, we first establish a strong MAR baseline by task-adaptively fine-tuning VideoMAE. To further enhance its capacity to perceive subtle motion cues, we propose **A^3-MAE**, a VideoMAE-based holistic-subtle collaborative framework that integrates **A**mplification, **A**ggregation, and **A**djustment in a unified design. Specifically, 1) we design a Temporal Gradient Local Amplification (TGLA) module to amplify subtle-motion regions directly from RGB inputs, forming a subtle-motion branch that complements the holistic-motion counterpart; 2) we develop a Holistic-Subtle Motion Aggregation (HSMA) module with dual cross-attention, enabling reciprocal conditioning between holistic and subtle branches while maintaining their complementarity; and 3) we introduce a Confidence-Aware Dynamic Adjustment (CADA) module to adaptively calibrate aggregation mismatch between the two branches in an instance-aware manner. Together, these components establish a unified amplify-aggregate-adjust paradigm for comprehensive micro-action understanding. Extensive experiments demonstrate that A^3-MAE achieves new state-of-the-art performance on both MA-52 and iMiGUE benchmarks. Code is available at https://github.com/zy-hfut/A3MAE.

Keywords: Micro-action recognition · VideoMAE

✉ Corresponding authors.

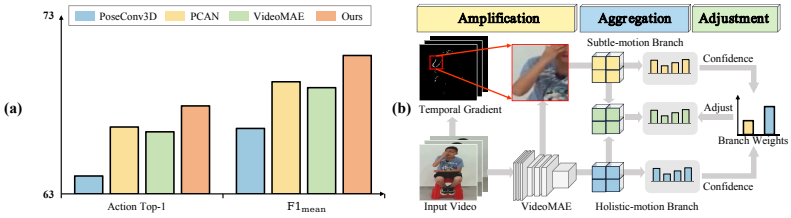

Fig. 1: (a) Comparison of different methods on MA-52 dataset. (b) Our **A^3-MAE** integrates three key stages: (1) **Amplification** enhances subtle motions via temporal gradient; (2) **Aggregation** aligns holistic and subtle semantics while keeping them complementary; (3) **Adjustment** adaptively reweights branches based on instance-wise confidence.

1 Introduction

Despite substantial progress in video action recognition [35, 36, 39], most existing studies primarily focus on high-amplitude, macro-scale activities [14, 16, 30] (e.g., running, cycling), leaving micro human movements largely underexplored. Micro-actions [12, 20, 24] refer to subtle, short-duration, and low-salience bodily motions that naturally occur during interpersonal communication. Accurate micro-action recognition (MAR) is crucial for emotion understanding and psychological analysis, as these fleeting cues often convey genuine intentions and latent affective states [10, 19, 27].

Pre-trained 3D CNNs [3, 32] and video Transformers [1, 25, 40] have demonstrated strong performance in conventional action recognition but transfer poorly to MAR, as low-amplitude and short-lived cues are easily overwhelmed by dominant appearance features and long-range temporal aggregation. To address this limitation, recent approaches (*e.g.*, PCAN [19], MMN [11]) incorporate skeletal priors to emphasize localized motion patterns. However, these methods rely on coarse joint graphs and thus fail to capture fine-grained sub-joint micro-movements [4, 23] (*e.g.*, finger tremors). Furthermore, their reliance on additional modalities or costly pose annotations significantly restricts their practical applicability. In contrast, VideoMAE [31], pre-trained through masked spatiotemporal reconstruction, inherently learns from partial observations and excels at capturing informative local motion and structural patterns. Motivated by this, we develop a VideoMAE-based framework tailored for MAR. As shown in Fig. 1 (a), a carefully designed VideoMAE model achieves competitive performance without auxiliary modalities [8, 19] or sophisticated task-specific losses [19]. Nevertheless, the discriminative cues in MAR remain extremely subtle, and the holistic aggregation within the patch-based encoder tends to smooth out these fine-grained signals, ultimately weakening the encoder's discriminative power.

To address this limitation, we further propose a novel VideoMAE-based holistic-subtle collaborative framework, termed **A^3-MAE**, which integrates **A**mplification, **A**ggregation, and **A**djustment into a unified design (see Fig. 1 (b)). ① **Amplification:** We design a Temporal Gradient Local Amplification **(TGLA)**

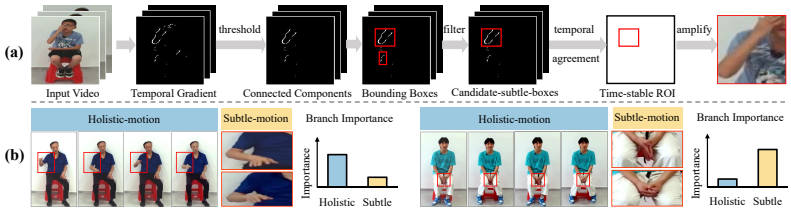

Fig. 2: (a) Temporal gradient (TG) local amplification. We first generate TG maps, from which motion regions are thresholded, labeled as bounding boxes, and filtered. The remaining boxes are integrated based on temporal agreement to yield a time-stable ROI for the subtle-motion branch. **(b) Illustration of different importance of holistic and subtle branches.** Limb-torso interactions depend more on holistic motion, whereas finger-centered movements rely primarily on the subtle-motion branch.

module to enhance subtle-motion regions while suppressing noise. We select the temporal gradient (TG) as the motion cue since it is modality-free, directly derived from RGB frames, and strongly correlated with micro-movements [34, 38]. However, raw TG often highlights irrelevant fluctuations (*e.g.*, mild body tremors), as illustrated in Fig. 2 (a). To suppress these noises, TGLA converts per-frame TG maps into a temporally consistent spatial proposal through thresholding and temporal agreement, followed by region-level feature refinement. Then, the refined regions are cropped and amplified to match the spatial scale of the holistic-motion branch, forming a subtle-motion branch that retains critical local cues with less background redundancy.

② **Aggregation:** To aggregate fine-grained cues from the subtle branch without compromising the holistic semantics of the RGB branch, we develop a Holistic-Subtle Motion Aggregation **(HSMA)** with dual cross-attention. Concretely, each branch queries the other to import complementary context and then applies a gated residual update to its own tokens. This preserves branch identity while injecting cross-branch information. Then, we concatenate the updated representations to form the aggregated vector. This design aligns scales and semantics across branches, mitigates single-branch dominance, and yields a representation that retains holistic characteristics yet effectively captures subtle motion cues.

③**Adjustment:** In our experiments, we observe that the importance of the holistic and subtle-motion branches varies significantly across different instances (as shown in Fig. 2 (b)). However, HSMA treats them as equally reliable, which leads to a mismatch that degrades accuracy, either by over-amplifying a weak branch or underweighting the truly diagnostic one. To address this issue, we introduce a Confidence-Aware Dynamic Adjustment **(CADA)** module that learns an instance-wise weight vector. Specifically, we design a small gating MLP to map per-branch confidence to instance-adaptive weights for each instance, which are used to reweight the contributions of the holistic- and subtle-motion branches during both aggregation and supervision. This dynamic, instance-aware calibration mitigates branch mismatch and produces more reliable micro-action predictions. Our main contributions are summarized as follows:

- We show that MAR can be effectively addressed from raw RGB videos by properly adapting VideoMAE, establishing a strong baseline that avoids reliance on pose annotations and auxiliary modalities.
- We propose A^3-MAE, a VideoMAE-based holistic-subtle collaborative framework that explicitly captures the complementarity between global motion context and subtle dynamics via Amplification, Aggregation, and Adjustment.
- Extensive experiments validate the effectiveness of our approach. Our A^3-MAE outperforms existing MAR methods and the strong VideoMAE baseline, achieving state-of-the-art results on both MA-52 and iMiGUE benchmarks.

2 Related Work

Video Action Recognition. Early video action recognition works include two-stream CNNs combining RGB and optical flow [29] and 3D CNNs that learn spatiotemporal filters end-to-end (e.g., I3D [3]). Subsequent designs focus on improving efficiency and temporal coverage (e.g., SlowFast [9]) and leveraging human poses to emphasize motion structure (e.g., PoseConv3D [8]). Recently, transformer-based approaches have become popular for capturing long-range dependencies in videos (e.g., VSwimT [25]), while self-supervised pretraining with VideoMAE [31] offers powerful general-purpose video representations. However, these pipelines are largely tuned for high-amplitude, holistic activities and can blur or miss the low-amplitude, transient cues central to micro-actions. *In this work, we first establish a strong MAR baseline on VideoMAE and then enhance it by explicitly injecting subtle-motion cues during training.*

Micro-Action Recognition. Micro-Action Recognition (MAR) focuses on capturing subtle, short-duration motions, requiring finer spatiotemporal perception than conventional action recognition. Recent efforts have significantly advanced this field. MA-52 [12] introduces a comprehensive whole-body micro-action benchmark with a MANet [12] baseline, establishing a solid foundation for MAR research. PCAN [19] mitigates sample ambiguity via a hierarchical action tree and prototype-guided calibration. MMN [11] integrates motion cues through motion-guided modulation, and enforces motion-consistency learning to emphasize subtle dynamics. APLT [41] extends MAR into a semi-supervised setting (SSMAR) and accordingly designs memory-based pseudo-label learning to train a robust model with limited annotations. *Different from them, we design a novel framework that extracts subtle-motion cues by temporal gradient and effectively bridges holistic- and subtle-motion cues for robust MAR.*

Motion Cue Extraction in Video Action Recognition. Various approaches have been explored to enhance motion cue modeling. The two-stream paradigm [29] pairs appearance with dense flow but is compute-heavy and sensitive to camera motion. Dynamic images [2] compress a clip into one frame, losing frame-level locality. Pose/skeleton pipelines [8] capture joint kinematics and are robust to appearance, but require extra modalities and costly annotations. In contrast, LTG [38] uses lightweight per-frame TG maps as a training-time auxiliary modality, boosting perception to transient motions. *Building on TG,*

we design a subtle-motion branch that directly extracts TG from RGB frames to amplify subtle-motion regions and complement the holistic-motion branch.

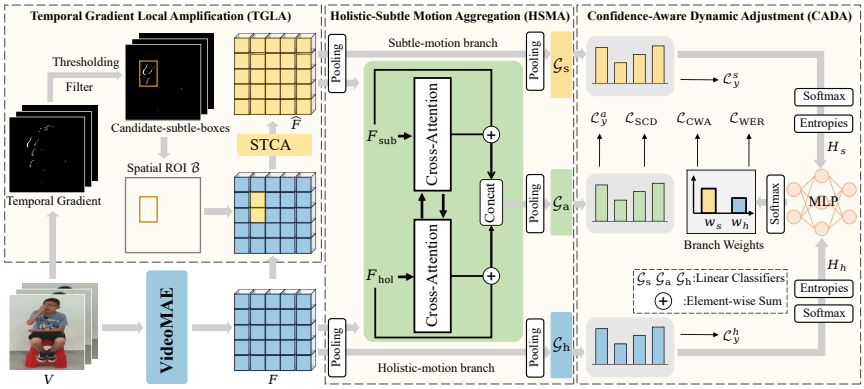

Fig. 3: The overview of A^3-MAE. Given an input video V, we propose **TGLA** to generate temporal gradient (TG) from RGB. Then, we obtain spatial regions and utilize spatio-temporal consistency amplifier (STCA) to amplify subtle features, forming the subtle-motion branch. In parallel, the VideoMAE backbone yields a holistic-motion branch. Holistic and subtle branches are aggregated by **HSMA**, aligning semantics while preserving complementarity. **CADA** produces instance-wise branch weights $[w_s, w_h]$ to adapt aggregation and supervision via $\mathcal{L}_{\text{SCD}}$, $\mathcal{L}_{\text{WER}}$, and $\mathcal{L}_{\text{CWA}}$. In addition, tri-head classification losses $\mathcal{L}_y^h$, $\mathcal{L}_y^s$, $\mathcal{L}_y^a$ supervise the classifiers of three branches $\mathcal{G}_h, \mathcal{G}_s, \mathcal{G}_a$, respectively. The aggregated head $\mathcal{G}_a$ is used for final evaluation.

3 Method

Task Definition. Given a video V, the goal of the Micro-Action Recognition (MAR) is to predict two levels of micro-action labels. The *action-level* category $Y_{\text{act}} \in \{1, 2, \ldots, N_A\}$ and the *body-level* category $Y_{\text{body}} \in \{1, 2, \ldots, N_B\}$, where N_A and N_B denote the total numbers of action-level and body-level classes, respectively.

VideoMAE Baseline. In this work, we establish our strong MAR baseline by fine-tuning VideoMAE [31] with the MAR dataset. Specifically, we select ViT-B/16 as the backbone, which is pre-trained on the official Kinetics-400 [15] via high-ratio masked video reconstruction. During training, we employ light data augmentations and uniformly sample 16 frames per clip to preserve motion cues. In addition, we adopt a cosine learning rate schedule, which stabilizes ViT convergence under small batch sizes and enhances perception to subtle micro-motions. *Despite its simplicity, our baseline delivers strong and competitive results compared with prior methods.*

Framework Overview. Our A^3-MAE framework enhances the strong VideoMAE baseline by integrating three key modules: Temporal Gradient Local Ampli-

fication (**TGLA**), Holistic-Subtle Motion Aggregation (**HSMA**), and Confidence-Aware Dynamic Adjustment (**CADA**). As illustrated in Fig. 3, given an input video, we first compute temporal gradient (TG) to amplify motion-salient cues, which guides TGLA to form a subtle-motion branch. The holistic-motion branch and subtle-motion branch are then aggregated through HSMA with dual cross-attention. Finally, CADA adaptively adjusts the contributions of the holistic- and subtle-motion branches based on instance-wise confidence, achieving accurate and robust predictions.

3.1 Temporal Gradient Local Amplification

We design TGLA to explicitly leverage TG for localizing subtle-motion regions and amplifying their cues, forming a subtle-motion feature branch that remains spatio-temporally aligned with the holistic-motion branch.

Subtle-Motion Extraction via Temporal Gradient. Given a sequence of frames $V = \{f_t\}_{t=1}^{S}$, where S denotes the number of frames, and each frame $f_t \in \mathbb{R}^{H\times W\times C}$ represents the t-th RGB image with height H, width W, and channel number $C = 3$, we denote by $f_t(i, j, c)$ the pixel value at spatial location (i, j) and channel c. We then compute a per-frame TG map TG_t for each frame f_t $(t = 2, \ldots, S)$, defined as the per-pixel, channel-averaged ℓ_1 difference between consecutive frames:

$$TG_t(i,j) \;=\; \frac{1}{C}\sum_{c=1}^{C} \big|\, f_t(i,j,c) - f_{t-1}(i,j,c) \,\big|\,. \tag{1}$$

For completeness, we set $TG_1 = TG_2$. For each frame, we then obtain a binary mask M_t by thresholding TG_t at the τ-th percentile (default: $\tau = 90$). We extract connected components $\{B_{t,m}\}$ from M_t, where m is the number of components in frame t, and convert each region to its tight axis-aligned bounding box. We assign a score to each box based on its TG value:

$$s(B_{t,m}) \;=\; \sum_{(i,j)\in B_{t,m}} TG_t(i,j), \tag{2}$$

and then select the box with the highest score as the candidate-subtle-box for each frame t, *i.e.*, $b_t = B_{t,\,\arg\max_m s(B_{t,m})}$,

To obtain a time-stable region of interest (ROI) for video, we aggregate $\{b_t\}_{t=1}^{S}$ by measuring temporal-agreement. Let $B^{\cap}$ and $B^{\cup}$ denote the intersection and union of all b_t, and define the overlap ratio as:

$$\rho \;=\; \frac{\mathrm{area}(B^{\cap})}{\mathrm{area}(B^{\cup}) + \varepsilon}\,, \tag{3}$$

where a small ε prevents division by zero. If $\mathrm{area}(B^{\cap}) > 0$ and $\rho > \eta$ (default: $\eta = 0.2$), we set the final ROI $\mathcal{B}=B^{\cap}$. Otherwise, ROI $\mathcal{B}$ is the candidate-subtle-box with the highest TG score among all frames.

Spatio-Temporal Consistency Amplifier (STCA). Feeding V into the VideoMAE backbone yields holistic-motion features $F\in\mathbb{R}^{T\times D\times H'\times W'}$, where T is the post-backbone temporal length (typically $T = S/r$ for tubelet size r), D

is the feature dimension, and H' and W' are the spatial resolutions. We apply the same spatial ROI $\mathcal{B}$ to every temporal slice via ROIAlign [13], cropping the region and bilinearly resizing it back to (H', W') to obtain temporally aligned subtle-motion features $\hat{F} \in \mathbb{R}^{T\times D\times H'\times W'}$. TGLA thereby enforces consistent spatial support over time, suppresses background drift, and produces a subtle-motion feature branch $\hat{F}$ that complements the holistic-motion RGB features F for subsequent aggregation.

Discussion. (i) We choose the TG as our motion cue because it is modality-free, computed directly from RGB, and highly correlated with subtle movements. *Crucially, micro-actions typically occur against largely static backgrounds, so frame differencing suppresses stationary clutter and isolates motion-only evidence.* (ii) Spatio-temporal consistency amplification is important because it enforces consistent spatial support over time and suppresses background drift. *Without it, per-frame TG tends to produce unstable and fragmented masks that can even amplify transient noise, degrading the effective aggregation in subsequent modules.*

3.2 Holistic-Subtle Motion Aggregation

In our preliminary results, we find that training the holistic- and subtle-motion branches independently or fusing them with simple concatenation provides only marginal gains, indicating that their complementary cues remain under-exploited. To address this, we develop a HSMA module with dual cross-attention, enabling reciprocal conditioning between holistic and subtle branches while maintaining their complementarity.

Gated Residual Cross-Updates. Given the holistic-motion and subtle-motion features F and $\hat{F}$, we first convert them into per-frame tokens by spatial pooling:

$$\begin{aligned} F_{\text{hol}} &= \text{GAP}(F) \in \mathbb{R}^{T\times D}, \\ F_{\text{sub}} &= \text{GAP}(\hat{F}) \in \mathbb{R}^{T\times D}, \end{aligned} \tag{4}$$

where GAP averages over the spatial dimensions (H', W'). We propose a dual cross-attention module that aggregates complementary cues from holistic- and subtle-motion, enabling mutual enhancement between the two branches. Specifically, subtle-motion features attend to the holistic-motion features and vice versa. Using the subtle branch as an example, the cross-attention is formulated as:

$$\tilde{F}_{\text{sub}} \;=\; \textbf{Cross-Attention}\big(F_{\text{sub}} \leftarrow F_{\text{hol}}\big), \tag{5}$$

where **Cross-Attention**(A $\leftarrow$ B) denotes that the queries come from A while the keys/values come from B in a cross-attention module [26]. Moreover, we propose a gated residual connection to balance the original feature (F_{sub}) and cross-updated feature ($\tilde{F}_{\text{sub}}$):

$$F'_{\text{sub}} \;=\; F_{\text{sub}} + \gamma_{\text{sub}}\big(\tilde{F}_{\text{sub}} + \text{FFN}(\text{LN}(\tilde{F}_{\text{sub}}))\big), \tag{6}$$

where γ_{sub} is a learnable gating coefficient, LN$(\cdot)$ denotes Layer Normalization and FFN$(\cdot)$ denotes a Feed-Forward Network [33]. The holistic branch is updated analogously, yielding F'_{hol}.

Next, the enhanced features (F'_{sub} and F'_{hol}) are average-pooled over time and then integrated by concatenation:

$$f_{\text{a}} = \big[\frac{1}{T}\sum_{t=1}^{T} F'_{\text{hol},t}; \frac{1}{T}\sum_{t=1}^{T} F'_{\text{sub},t}\big] \in \mathbb{R}^{2D}, \tag{7}$$

which is then passed through a linear classifier for final predictions. This bidirectional conditioning enables holistic-motion features to leverage subtle-motion cues and vice versa, generating a more robust and complementary representation for MAR. Additionally, we retain two auxiliary classifiers operating on the temporally averaged features $f_{\text{h}} = \frac{1}{T}\sum_{t=1}^{T} F_{\text{hol},t}$ and $f_{\text{s}} = \frac{1}{T}\sum_{t=1}^{T} F_{\text{sub},t}$, respectively.

Tri-Head Classification Loss. The classification loss is applied to the generated three types of features, *i.e.,* holistic feature f_{h}, subtle feature f_{s}, and aggregated feature f_{a}. The loss function can be formulated as:

$$\begin{aligned} \mathcal{L}_y^b &= \text{CE}\big(\mathcal{G}_b(f_b),\, y\big), \quad b \in \{a, s, h\}, \\ \mathcal{L}_y &= \mathcal{L}_y^a + \frac{1}{2}\big(\mathcal{L}_y^s + \mathcal{L}_y^h\big), \end{aligned} \tag{8}$$

where CE is cross-entropy loss and each head $\mathcal{G}_b$ produces logits in $\mathbb{R}^N$ with $N \in \{N_A,\, N_B\}$ depending on the task.

3.3 Confidence-Aware Dynamic Adjustment

Motivation: Although HSMA aligns the holistic and subtle branches at the token level, we observe that naive aggregation with fixed or implicit weights often leads to a *cross-branch* ***mismatch***. That is, the importance of the holistic- and subtle-motion branches varies significantly across instances. However, the aggregation of HSMA treats them as equally reliable. For instance, in head-centric micro-actions such as *"head up"*, highly localized cues around the head suffice for discrimination, and the subtle-motion branch is more reliable than the holistic branch. In contrast, actions involving coordination between limbs and torso, such as *"scratching or touching hindbrain"*, require broader postural context. In such cases, the holistic-motion branch provides the critical evidence while purely local cues are ambiguous (see Fig. 4 (a)). This *mismatch* degrades accuracy by either over-amplifying a weak branch or underweighting the truly diagnostic one.

To rectify this *mismatch*, we propose the CADA module. It dynamically weights the holistic and subtle branches based on their instance-wise confidence, adaptively balancing their contributions in feature aggregation and supervision.

Entropy-based Confidence Estimation. We obtain the probabilities of holistic branch and subtle branch, which are denoted as $p_h = \text{softmax}(\mathcal{G}_h(f_{\text{h}}))$ and $p_s = \text{softmax}(\mathcal{G}_s(f_{\text{s}}))$, respectively. The corresponding entropy is represented as:

$$H_b \;=\; -\sum_{c=1}^{N} p_b(c)\, \log p_b(c), \qquad b \in \{s, h\}, \tag{9}$$

where N is the number of classes for the current task ($N \in \{N_A,\, N_B\}$). The confidence vector is defined as:

$$s = \big[-H_s,\, -H_h\big] \in \mathbb{R}^2, \tag{10}$$

where a larger value indicates higher confidence. Then, a lightweight MLP layer $\mathrm{MLP}(\cdot)$ is employed to produce instance-wise branch weights W:

$$W = \big[w_s,\, w_h\big] = \mathrm{softmax}(\mathrm{MLP}(s)). \tag{11}$$

The generated weights W are used to fuse the logits from the two branches into a unified output:

$$o_{\mathrm{teacher}} = w_s\, \mathcal{G}_s(f_{\mathrm{s}}) + w_h\, \mathcal{G}_h(f_{\mathrm{h}}). \tag{12}$$

Selective Confidence Distillation. We supervise the aggregated head via a selective confidence distillation, which treats o_{teacher} as a confidence-weighted teacher:

$$\mathcal{L}_{\mathrm{SCD}} \;=\; \mathrm{KL}\big(\mathrm{softmax}(o_{\mathrm{teacher}}) \,\big\|\, \mathrm{softmax}(\mathcal{G}_a(f_{\mathrm{a}}))\big). \tag{13}$$

Here $\mathrm{KL}(\cdot\|\cdot)$ is the Kullback-Leibler divergence [17]. This term encourages the aggregated head to trust the branch that is more confident for the current sample, while still learning from both branches. Note that, during the distillation, gradients are propagated only to the aggregation head, and not to the holistic or subtle heads.

Weight Entropy Regularization. To avoid collapsing to a single branch, we also perform weight entropy regularization:

$$\begin{aligned} \mathcal{L}_{\mathrm{WER}} &= \max\big(0,\, \delta - \Phi(W)\big), \\ \Phi(W) &= -\sum_{i\in\{h,s\}} w_i \log\big(w_i{+}\varepsilon\big), \end{aligned} \tag{14}$$

where δ is the lower bound set for entropy (default: $\delta = 0.4$) and $\varepsilon > 0$ is a small constant to prevent $\log(0)$.

Confidence-Weight Alignment. To further calibrate the learned weights W, we align them with a confidence target derived from per-branch losses, where a lower loss indicates higher confidence. This alignment prevents the MLP-based weighting (Eq. (11)) from learning biased or inconsistent weights and ensures that the aggregation process in Eq. (12) remains consistent with branch reliability. Concretely, we convert the cross-entropy losses of the holistic and subtle branches into a normalized, loss-derived confidence target:

$$I = \mathrm{softmax}\Big(-\big[\,\mathcal{L}_{\mathrm{y}}^{s},\, \mathcal{L}_{\mathrm{y}}^{h}\,\big]\Big) \in \mathbb{R}^2. \tag{15}$$

We then regularize the mismatch between I and W via confidence-weight alignment:

$$\mathcal{L}_{\mathrm{CWA}} = \mathrm{KL}\big(I \,\|\, W\big). \tag{16}$$

Overall Objective. Taking Eqs. (8), (13), (14) and (16), the overall training objective is formulated as:

$$\mathcal{L}_{\mathrm{A^3\text{-}MAE}} = \mathcal{L}_y + \mathcal{L}_{\mathrm{SCD}} + \mathcal{L}_{\mathrm{WER}} + \lambda\, \mathcal{L}_{\mathrm{CWA}}, \tag{17}$$

where λ is a hyperparameter that balances the importance of confidence-weight alignment.

4 Experiments

4.1 Experimental Setup

Micro-Action and Micro-Gesture Datasets. We conduct experiments on two benchmark datasets that focus on subtle human movements. **MA-52** [12] is a human-centric interview-style video dataset specifically designed for whole-body micro-action recognition. It contains 22,422 annotated clips collected from 205 participants and provides two levels of annotations: 7 body-level categories and 52 fine-grained action-level categories. We follow the official split with 11,250 training samples, 5,586 validation samples, and 5,586 test samples. The dataset covers a broad range of subtle and unconscious whole-body movements, making it a challenging benchmark for fine-grained micro-action analysis. In addition, we evaluate our method on **iMiGUE** [24], a micro-gesture dataset that focuses primarily on upper-body movements. The dataset was collected from post-competition athlete interview scenarios and contains 32 micro-gesture categories. Compared with MA-52, which emphasizes whole-body micro-actions, iMiGUE mainly captures fine-grained upper-limb gestures. The dataset consists of 12,893 training samples, 777 validation samples, and 4,562 test samples. The inclusion of both MA-52 and iMiGUE allows us to evaluate our framework across different granularities and spatial scopes of subtle human movements, ranging from whole-body micro-actions to upper-body micro-gestures.

Evaluation Protocol. Following prior work [11, 12, 19], we report multiple evaluation metrics on MA-52, including Action Top-1 and Top-5 accuracy, Body Top-1 accuracy, and both macro- and micro-F1 scores computed at the body level and the action level. $\mathrm{F1}_{\mathrm{macro}}$ is the unweighted mean of per-class F1, assigning equal importance to each action category regardless of its sample size. In contrast, $\mathrm{F1}_{\mathrm{micro}}$ is computed from globally pooled counts, so each instance contributes equally and frequent categories have greater influence than rare ones. We also report the mean of the four F1 scores, denoted as: $\mathrm{F1}_{\mathrm{mean}} = \frac{\mathrm{F1}^{\mathrm{body}}_{\mathrm{macro}}+\mathrm{F1}^{\mathrm{body}}_{\mathrm{micro}}+\mathrm{F1}^{\mathrm{action}}_{\mathrm{macro}}+\mathrm{F1}^{\mathrm{action}}_{\mathrm{micro}}}{4}$, where "body" / "action" denotes the body-level / action-level prediction task. For iMiGUE, only single-level gesture labels are provided; therefore we report action-level metrics only.

Implementation Details. We implement our models in MMAction2 [6], which is a video toolbox. The backbone is VideoMAE [31] with ViT-B/16. Inputs are sampled as 16 frames per clip, resized to 224×224. During training, we use the AdamW optimizer with a base learning rate 3e-4, weight decay 1e-2, and a batch size of 24. The model is trained for 42 epochs with a cosine decay schedule. Training is conducted with 4 NVIDIA 4090 GPUs. All results are reported on the official test splits using the evaluation protocol mentioned above. Unless otherwise noted, we use the following hyperparameters for all settings: TG-mask percentile threshold $\tau = 90$; temporal-agreement threshold $\eta = 0.2$; confidence-weight alignment $\mathcal{L}_{\mathrm{CWA}}$ weight $\lambda = 0.1$; lower bound set for entropy $\delta = 0.4$. To avoid over-tuning hyperparameters to individual datasets, we set τ, η, λ, and δ once based on MA-52 and keep them unchanged when evaluating on iMiGUE.

Table 1: Performance comparison on the MA-52 and iMiGUE datasets. V denotes RGB video frames; P denotes human pose (skeleton/keypoints). The best are highlighted in **bold**, while the second-best results are underlined. † reported by the original paper but not reproducible; therefore excluded from all comparisons. Results marked with * are reproduced with official code. Note that PCAN [19] is not evaluated on iMiGUE since its hierarchical body–action learning framework is specifically designed for MA-52 and is not directly applicable to single-level micro-gesture annotations.

Method	Venue	Input	MA-52								iMiGUE
			Body	Action		Body		Action		All	Action
			Top-1	Top-1	Top-5	$F1_{macro}$	$F1_{micro}$	$F1_{macro}$	$F1_{micro}$	$F1_{mean}$	Top-1
C3D [32]	ICCV'15	V	74.04	52.22	86.97	66.60	74.04	40.86	52.22	58.43	30.13
I3D [3]	CVPR'17	V	78.16	57.07	88.67	71.56	78.16	39.84	57.07	61.66	35.08
TSM [22]	ICCV'19	V	77.64	56.75	87.47	70.98	77.64	40.19	56.75	61.39	56.23
SlowFast [9]	ICCV'19	V	77.18	59.08	88.54	70.61	77.18	44.96	59.06	63.09	53.70
TIN [28]	AAAI'20	V	73.26	52.81	85.37	66.99	73.26	39.82	52.81	58.22	55.35
TimesFormer [1]	ICML'21	V	69.17	40.67	82.67	61.90	69.17	34.38	40.67	51.53	50.66
CTRGCN [5]	ICCV'21	P	72.06	52.61	81.22	63.46	72.06	37.79	52.61	56.48	56.12
VSwimT [25]	CVPR'22	V	77.95	57.23	87.99	71.25	77.95	38.53	57.23	61.24	60.11
Uniformer [21]	TPAMI'23	V	79.03	58.89	87.29	71.80	79.03	48.01	58.89	64.43	59.91
HD-GCN [18]	ICCV'23	P	75.76	60.19	86.90	67.32	75.76	44.50	60.19	61.94	53.44
Koopman [37]	CVPR'23	P	75.04	59.70	86.79	66.48	75.04	44.57	59.70	61.45	52.08
FR-Head [42]	CVPR'23	P	76.35	61.17	86.99	68.88	76.35	47.43	61.17	63.46	55.26
SkateFormer [7]	ECCV'24	P	75.67	59.67	87.27	68.33	75.67	45.58	59.76	62.34	47.74
MANet [12]	TCSVT'24	V	78.95	61.33	88.83	72.87	78.95	49.22	61.33	65.59	58.08
MMN [11]	ACM MM'25	P	78.52	62.71	89.83	71.86	78.52	48.27	62.71	65.34	60.21
PCAN [19]	AAAI'25	V	79.36	60.03	87.65	72.37	79.36	43.29	60.03	63.76	-
PCAN† [19]	AAAI'25	V + P	82.30	66.74	91.75	77.02	82.30	53.83	66.74	69.97	-
PCAN* [19]	AAAI'25	V + P	80.74	64.59	90.74	74.37	80.74	51.04	64.59	67.69	-
PCAN (VideoMAE)	-	V	80.79	63.50	90.67	74.77	80.79	52.20	63.50	67.82	-
VideoMAE Baseline (Ours)	-	V	81.67	66.47	91.78	75.53	81.83	52.46	65.92	68.94	66.44
A^3-MAE (Ours)	-	V	**82.62**	**67.92**	**91.89**	**76.36**	**82.57**	**55.97**	**67.64**	**70.64**	**67.08**

4.2 Comparison with the State-of-the-Art

As illustrated in Tab. 1, we evaluate our VideoMAE baseline and A^3-MAE against a comprehensive set of state-of-the-art methods, including: *CNN-based*—C3D [32], I3D [3], TSM [22], SlowFast [9], TIN [28]; *Transformer-based*—TimesFormer [1], VSwimT [25], Uniformer [21]; *Pose/graph-based*— CTRGCN [5], HD-GCN [18], Koopman [37], FR-Head [42], SkateFormer [7]; and *MAR-specific*—MANet [12], MMN [11], PCAN [19]. *Note that, PCAN† denotes the numbers reported in the original paper, while PCAN* denotes our reproduction using the official code under the same evaluation protocol. For fair comparison, we mainly compare with PCAN*.*

VideoMAE Baseline vs. the State-of-the-Art Methods. As shown in Tab. 1, the VideoMAE baseline outperforms previous state-of-the-art methods across all evaluation metrics both on MA-52 and iMiGUE. Importantly, it uses RGB only yet remains competitive—indeed, it outperforms the approaches that rely on pose modality, including FR-Head [42] and SkateFormer [7]. We also evaluate PCAN using the same VideoMAE backbone (named as PCAN (VideoMAE)) and the same input, *i.e.*, both methods take RGB frames as the only input. Even using the same backbone, our VideoMAE baseline yields better results

Table 2: Ablation of the three components.

Method	Action Top-1	$F1^{action}_{macro}$	$F1^{action}_{micro}$
Baseline	66.47	52.46	65.92
+ TGLA	67.04	55.09	66.74
+ TGLA + HSMA	67.63	55.39	67.44
+ TGLA + HSMA + CADA	**67.92**	**55.97**	**67.64**

Table 3: Ablation of the losses in CADA.

Index	$\mathcal{L}_{\mathbf{SCD}}$	$\mathcal{L}_{\mathbf{CWA}}$	$\mathcal{L}_{\mathbf{WER}}$	Action Top-1	$F1^{action}_{macro}$	$F1^{action}_{micro}$
I				67.63	55.39	67.44
II	✓			67.49	55.59	67.08
III	✓	✓		67.60	55.47	67.31
IV	✓		✓	67.69	55.67	67.44
V	✓	✓	✓	**67.92**	**55.97**	**67.64**

than PCAN (VideoMAE) across all metrics on MA-52. This indicates that the advantage of PCAN largely arises from the auxiliary modality, whereas a well-designed RGB-only VideoMAE baseline already provides a strong foundation.

A^3-MAE vs. the State-of-the-Art Methods. Building on the strong VideoMAE baseline, A^3-MAE achieves consistent improvements across all metrics on both MA-52 and iMiGUE, attaining the best overall performance among all compared methods (Tab. 1). It performs favorably against MAR-specific designs (e.g., MANet, MMN, PCAN), and even surpasses the PCAN variant built on the same VideoMAE backbone (PCAN (VideoMAE)), indicating that the gains arise not only from the strong VideoMAE backbone but also from our holistic-subtle collaboration pipeline.

4.3 Ablation Study

Ablation of Components. In Tab. 2, we evaluate the effectiveness of each component on MA-52, starting from the VideoMAE baseline and progressively enabling TGLA, HSMA, and CADA. *First*, adding TGLA to the baseline yields consistent improvements across all metrics, indicating that TGLA better highlights subtle motions by focusing on temporally salient regions. *Second*, introducing HSMA on top of TGLA further improves every metric, suggesting that reciprocal conditioning between holistic- and subtle-motion branches produces more discriminative representations. *Third*, enabling CADA on top of (TGLA+HSMA) brings additional gains, evidencing that confidence-aware dynamic adjustment mitigates cross-branch mismatch. *Overall*, the full configuration (TGLA + HSMA + CADA) delivers the best results on all metrics, demonstrating the complementarity and effectiveness of the proposed components.

Ablation of CADA Losses. In Tab. 3, we analyze the losses inside CADA, starting from our method without CADA (**I**) and progressively enabling selective confidence distillation $\mathcal{L}_{\text{SCD}}$, confidence-weight alignment $\mathcal{L}_{\text{CWA}}$, and weight entropy regularization $\mathcal{L}_{\text{WER}}$. *First*, using $\mathcal{L}_{\text{SCD}}$ alone (**II**) slightly underperforms the no-CADA baseline (**I**). This behavior is expected, as $\mathcal{L}_{\text{SCD}}$ selectively amplifies high-confidence branches. Without additional calibration, early-stage confidence estimates may be biased, causing the aggregation to prematurely favor the dominant branch. Therefore, $\mathcal{L}_{\text{SCD}}$ is designed to operate in conjunction with complementary regularization terms. *Second*, adding $\mathcal{L}_{\text{CWA}}$ on $\mathcal{L}_{\text{SCD}}$ (**III**) consistently improves over (**II**) and brings the variant on par with no-CADA (**I**), indicating a calibration effect on the learned weights. *Third*, pairing $\mathcal{L}_{\text{SCD}}$ with $\mathcal{L}_{\text{WER}}$ (**IV**) stabilizes aggregation and mitigates early dominance, outperforming

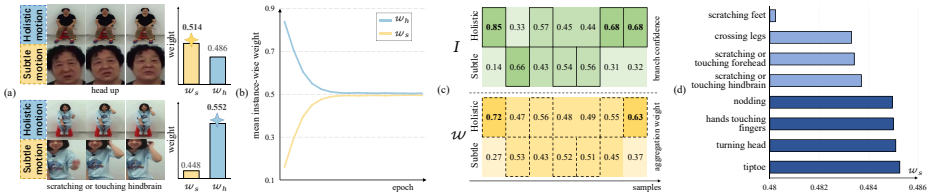

Fig. 4: (a) Instance-wise weights of subtle-motion vs. holistic-motion for examples, showing instance-specific allocation. (b) Evolution of the mean weights over training epochs, indicating dynamic cross-branch rebalancing and stage-wise learning. (c) Visualization of branch confidence and instance-wise weight on MA-52. The maximum value in each branch is highlighted with a border. (d) Class-wise mean subtle-motion weight on MA-52. Localized head/hand micro-actions assign higher subtle-motion weights than posture-dependent actions.

$\mathcal{L}_{\text{SCD}}$ alone (**II**) and exceeding no-CADA (**I**). *Overall*, enabling all three losses (**V**) is higher than all other variants across metrics, outperforming the no-CADA baseline and indicating that $\mathcal{L}_{\text{SCD}}$, $\mathcal{L}_{\text{CWA}}$, and $\mathcal{L}_{\text{WER}}$ are complementary when used together.

4.4 More Investigations

Analysis of Instance-wise Weights. *(a) Instance-specific allocation.* Fig. 4 (a) shows that CADA learned weights vary across samples. For highly localized micro-actions around the head (e.g., *"head up"*), CADA up-weights the subtle-motion branch, as discriminative cues are confined in space and time. In contrast, actions involving coordination between limbs and torso (e.g., *"scratching or touching hindbrain"*) receive more weight on the holistic-motion branch, where broader body context disambiguates look-alike local motions. This pattern indicates that CADA adapts branch contributions per instance according to their confidence, instead of applying a fixed prior weighting. *(b) Mean weight over training.* Fig. 4 (b) shows that early in training CADA assigns higher weight to the holistic-motion branch which cues are easier to learn. As training proceeds, the mean weight allocation stabilizes close to parity while retaining instance-level adjustments. This trajectory reflects CADA dynamic adjustment without a manually designed curriculum.

Visualization Analysis. *(a) Confidence-weight alignment.* As shown in Fig. 4 (c), branch confidences and instance-wise weights exhibit a strong concordance across selected samples: the higher-confidence branch receives a larger weight. This alignment shows that the model uses confidence as a reliability cue to adjust aggregation weights, giving more influence to the reliable branch and improving stability and accuracy without manual tuning. (b) *Class-wise mean subtle-motion weight.* As shown in Fig. 4 (d), the class-wise mean subtle-motion weight exhibits a sensible pattern. head/hand-centric micro-actions (e.g., *"nodding"*, *"hands touching fingers"*) allocate higher subtle-motion weights, consistent with their highly localized, brief signals. In contrast, actions that require broader posture (e.g.,

Head	Action Top-1	Body Top-1	$F1_{mean}$
$\mathcal{G}_s$	67.49	82.56	70.60
$\mathcal{G}_h$	67.60	82.56	70.49
$\mathcal{G}_a$	**67.92**	**82.62**	**70.64**

Table 4: Comparison of classification heads on MA-52. $\mathcal{G}_s$: subtle-motion head; $\mathcal{G}_h$: holistic-motion head; $\mathcal{G}_a$: aggregated head.

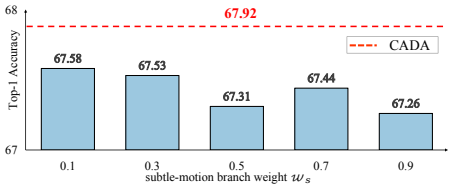

Fig. 5: Effect of the fixed subtle-motion branch weight w_s. The red dashed line denotes the CADA result.

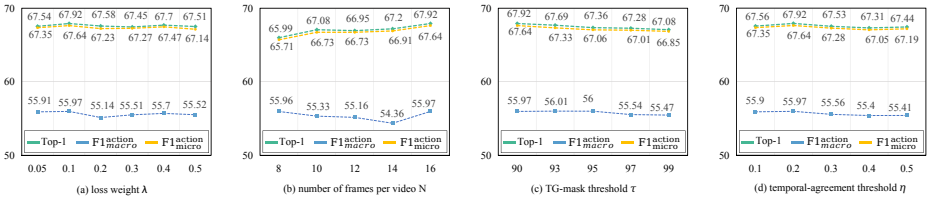

Fig. 6: Effect of hyperparameters at the action level on MA-52. (a) Varying the confidence-weight alignment $\mathcal{L}_{\text{CWA}}$ weight λ. (b) Varying the number of video frames. (c) Varying the TG-mask threshold τ. (d) Varying the temporal-agreement threshold η .

"scratching feet", "crossing legs") lower on the subtle-motion branch, reflecting the usefulness of holistic cues to disambiguate look-alike local movements.

Head-wise Analysis. As shown in Tab. 4, the aggregated head $\mathcal{G}_a$ achieves the best overall scores, with consistent gains over single-branch heads. The larger gain in Action Top-1 suggests that cross-branch interaction mainly benefits action-level discrimination, while single branches already saturate the coarser body categories. The closeness of $\mathcal{G}_s$ and $\mathcal{G}_h$ indicates each branch is competitive on its own, whereas the persistent advantage of $\mathcal{G}_a$ confirms the effectiveness of HSMA with CADA.

Fixed Branch Weight vs. CADA. In Fig. 5, we fix the branch weight w_s at several values and compare it with our CADA (red dashed line) at the action level. Across Top-1, every fixed setting underperforms CADA, indicating that static weights cannot accommodate the instance-dependent contributions of the two branches. Emphasizing holistic-motion branch under fixed weighting enhances accuracy (smaller w_s). Moreover, simple equal-weighting schemes ($w_s = 0.5$) performed poorly, further indicating that cross-branch contributions exhibit sample dependency rather than being fixed constants. In contrast, CADA dynamically derives weights from confidence signals, modulating cross-branch contributions on a per-sample basis and achieving superior overall performance without manual hyperparameter tuning.

Hyperparameters Analysis. (**I**) *Confidence-Weight Alignment Loss weight* λ (Fig. 6(a)): performance remains stable within a moderate range and peaks at $\lambda = 0.1$. A too small value weakens cross-branch consistency, while an overly large value suppresses discriminative cues. (**II**) *Number of frames per video*

Table 5: Module-wise efficiency analysis on MA-52. Params and GFLOPs are measured per clip under the same input setting as the main experiments.

Metric	VideoMAE	TGLA	HSMA	CADA	A^3-MAE
Params (M)	86.47	0	14.40	< 0.01	100.87
GFLOPs	180.50	< 0.01	0.12	< 0.01	180.62

N (Fig. 6(b)): increasing N improves temporal coverage, with the best results achieved at $N = 16$. Short clips (e.g., $N = 8$) often miss subtle motion phases, whereas moderate lengths ($N = 10\sim14$) already provide stable gains. (**III**) *TG-mask threshold* τ (Fig. 6(c)): lower thresholds retain more motion evidence, achieving the best overall performance at $\tau = 90$. Slightly higher settings (e.g., τ=93) can marginally favor $\mathrm{F1}^{\mathrm{action}}_{\mathrm{macro}}$ by further filtering low-amplitude noise and emphasizing rarer subtle patterns. Pushing τ beyond this discards too many informative regions and degrades performance. (**IV**) *Temporal-agreement threshold* η (Fig. 6(d)): the best performance is obtained at $\eta = 0.2$. Smaller values allow more noisy regions, whereas larger values (e.g., $\eta \geq 0.4$) over-constrain temporal agreement and degrade performance.

Computational Efficiency. We further analyze the computational overhead introduced by each component in Tab. 5. Compared with VideoMAE, A^3-MAE increases the parameters from 86.47M to 100.87M, mainly due to HSMA, while TGLA is parameter-free and CADA adds negligible cost. For computation, A^3-MAE only increases the cost from 180.50 to 180.62 GFLOPs, indicating that the performance gains are achieved with marginal overhead and that the main cost remains dominated by the VideoMAE backbone.

5 Conclusion

In this paper, we presented A^3-MAE, a VideoMAE-based holistic-subtle framework that unifies Amplification, Aggregation, and Adjustment for micro-action recognition. By amplifying subtle motion cues and adaptively integrating them with holistic motion patterns, A^3-MAE establishes a robust amplify-aggregate-adjust paradigm. Extensive experiments demonstrate its superior performance, setting new state-of-the-art results on MA-52 and iMiGUE benchmarks.

Acknowledgements

This work was funded by the National Natural Science Foundation of China (No. 62572166 & No. 62402157) and the Fundamental Research Funds for the Central Universities (No. JZ2025HGTB0219). The computation is completed on the HPC Platform of Hefei University of Technology.

References

1. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: Proceedings of the International Conference on Machine Learning (2021)
2. Bilen, H., Fernando, B., Gavves, E., Vedaldi, A., Gould, S.: Dynamic image networks for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2016)
3. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
4. Chen, Y., Li, Y., Zhang, C., Zhou, H., Luo, Y., Hu, C.: Informed patch enhanced hypergcn for skeleton-based action recognition. Information Processing & Management (2022)
5. Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W.: Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
6. Contributors, M.: Openmmlab's next generation video understanding toolbox and benchmark (2020)
7. Do, J., Kim, M.: Skateformer: skeletal-temporal transformer for human action recognition. In: European Conference on Computer Vision (2024)
8. Duan, H., Zhao, Y., Chen, K., Lin, D., Dai, B.: Revisiting skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
9. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE International Conference on Computer Vision (2019)
10. Gong, F., Chen, J., Zhu, J., Bao, Q., Gao, F., Gu, R., Xu, G.: Micro-action recognition via hierarchical fusion and inference. In: Proceedings of the 32nd ACM International Conference on Multimedia (2024)
11. Gu, J., Li, K., Wang, F., Wei, Y., Wu, Z., Fan, H., Wang, M.: Motion matters: Motion-guided modulation network for skeleton-based micro-action recognition. arXiv preprint arXiv:2507.21977 (2025)
12. Guo, D., Li, K., Hu, B., Zhang, Y., Wang, M.: Benchmarking micro-action recognition: Dataset, method, and application. IEEE Transactions on Circuits and Systems for Video Technology (2024)
13. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision (2017)
14. Heilbron, F.C., Escorcia, V., Ghanem, B., Niebles, J.C.: Activitynet: A large-scale video benchmark for human activity understanding. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (2015)
15. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
16. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: Proceedings of the IEEE International Conference on Computer Vision (2011)
17. Kullback, S., Leibler, R.A.: On information and sufficiency. The annals of mathematical statistics (1951)
18. Lee, J., Lee, M., Lee, D., Lee, S.: Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)

19. Li, K., Guo, D., Chen, G., Fan, C., Xu, J., Wu, Z., Fan, H., Wang, M.: Prototypical calibrating ambiguous samples for micro-action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence (2025)
20. Li, K., Liu, P., Guo, D., Wang, F., Wu, Z., Fan, H., Wang, M.: Mmad: Multi-label micro-action detection in videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2025)
21. Li, K., Wang, Y., Zhang, J., Gao, P., Song, G., Liu, Y., Li, H., Qiao, Y.: Uniformer: Unifying convolution and self-attention for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
22. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. In: Proceedings of the IEEE International Conference on Computer Vision (2019)
23. Liu, J., Chen, C., Liu, M.: Multi-modality co-learning for efficient skeleton-based action recognition. In: Proceedings of the 32nd ACM international conference on multimedia (2024)
24. Liu, X., Shi, H., Chen, H., Yu, Z., Li, X., Zhao, G.: 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 (2021)
25. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
26. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in Neural Information Processing Systems (2019)
27. Noroozi, F., Corneanu, C.A., Kaminska, D., Sapinski, T., Escalera, S., Anbarjafari, G.: Survey on emotional body gesture recognition. IEEE Transactions on Affective Computing (2021)
28. Shao, H., Qian, S., Liu, Y.: Temporal interlacing network. In: Proceedings of the AAAI Conference on Artificial Intelligence (2020)
29. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Proceedings of Advances in Neural Information Processing Systems (2014)
30. Soomro, K., Zamir, A.R., Shah, M.: UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
31. Tong, Z., Song, Y., Wang, J., Wang, L.: VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in Neural Information Processing Systems (2022)
32. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision (2015)
33. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems (2017)
34. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks for action recognition in videos. IEEE Transactions on Pattern Analysis and Machine Intelligence (2018)
35. Wang, M., Xing, J., Jiang, B., Chen, J., Mei, J., Zuo, X., Dai, G., Wang, J., Liu, Y.: A multimodal, multi-task adapting framework for video action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence (2024)
36. Wang, M., Xing, J., Liu, Y.: Actionclip: A new paradigm for video action recognition. arXiv preprint arXiv:2109.08472 (2021)

37. Wang, X., Xu, X., Mu, Y.: Neural koopman pooling: Control-inspired temporal dynamics encoding for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2023)
38. Xiao, J., Jing, L., Zhang, L., He, J., She, Q., Zhou, Z., Yuille, A., Li, Y.: Learning from temporal gradient for semi-supervised action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
39. Yang, J., Dong, X., Liu, L., Zhang, C., Shen, J., Yu, D.: Recurring the transformer for video action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
40. Yu, L., Cheng, Y., Sohn, K., Lezama, J., Zhang, H., Chang, H., Hauptmann, A.G., Yang, M.H., Hao, Y., Essa, I., et al.: Magvit: Masked generative video transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
41. Zhang, Y., Cheng, L., Wang, Y., Zhong, Z., Wang, M.: Towards micro-action recognition with limited annotations: An asynchronous pseudo labeling and training approach. In: International Joint Conference on Artificial Intelligence (2025)
42. Zhou, H., Liu, Q., Wang, Y.: Learning discriminative representations for skeleton based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2023)