CLIP-MG: Guiding Semantic Attention with Skeletal Pose Features and RGB Data for Micro-Gesture Recognition on the iMiGUE Dataset

Santosh Patapati[1], Trisanth Srinivasan[1] and Amith Adiraju[1]

[1]Cyrion Labs, Texas, United States

Abstract

Micro-gesture recognition is a challenging task in affective computing due to the subtle, involuntary nature of the gestures and their low movement amplitude. In this paper, we introduce a Pose-Guided Semantics-Aware CLIP-based architecture, or CLIP for Micro-Gesture recognition (CLIP-MG), a modified CLIP model tailored for micro-gesture classification on the iMiGUE dataset. CLIP-MG integrates human pose (skeleton) information into the CLIP-based recognition pipeline through pose-guided semantic query generation and a gated multi-modal fusion mechanism. The proposed model achieves a Top-1 accuracy of 61.82%. These results demonstrate both the potential of our approach and the remaining difficulty in fully adapting vision-language models like CLIP for micro-gesture recognition.

Keywords

Micro-gesture recognition, Vision-language models, CLIP adaptation, Pose-guided fusion, Multi-modal deep learning, Semantic query generation, Human pose estimation, Affective computing

1. Introduction

Micro-gestures (MGs) are subtle, spontaneous body movements that can reveal hidden emotional states [1, 2], often occurring when people attempt to suppress their true feelings. Unlike overt actions or expressive gestures, micro-gestures involve minute motions (e.g., slight fidgeting, brief facial or limb movements) that are short in duration and low in amplitude, making them hard to detect and classify [3]. The analysis of micro-gestures has gained traction in affective computing and human behavior understanding because these involuntary cues provide valuable insight into a person's internal state. Automatic recognition of micro-gestures is therefore important for applications in psychology, human-computer interaction, and emotion analysis [4, 5].

In this paper, we present CLIP-MG, a novel multi-modal framework for micro-gesture classification on iMiGUE. Our approach builds upon previous work by incorporating pose (skeleton) data in a principled way. The main contributions are summarized as follows:

1. We develop a system that uses human pose cues to help guide the semantic query extraction from video frames. The skeleton information helps focus the CLIP visual encoder on the regions of subtle motion. This creates a semantic query embedding that is rich with features relevant to pose.
2. We introduce a gated fusion mechanism to combine visual and skeleton representations effectively. Our gated fusion learns to weight and integrate the two modalities. This allows pose features to adaptively modulate the visual features before and during the cross-attention process.
3. We extend CLIP to a multi-modal setting. The pose-based query is fed into the CLIP transformer for cross-attention over semantically significant visual token features. This limits the model to attend to parts relevant to gesture. This results in a fused representation that has both semantic and motion-specific information for classification.

MiGA@IJCAI25: International IJCAI Workshop on 3rd Human Behavior Analysis for Emotion Understanding, August 29, 2025, Guangzhou, China.

santosh@cyrionlabs.org (S. Patapati); trisanth@cyrionlabs.org (T. Srinivasan); aadiraju@cyrionlabs.org (A. Adiraju)
0009-0007-4780-2624 (S. Patapati); 0009-0009-7588-7498 (T. Srinivasan)

4. We evaluate CLIP-MG on the iMiGUE micro-gesture dataset. We additionally perform numerous ablation studies to quantify how much each proposed component improves performance. The results of our ablation studies provide insights for future researchers and future research directions.

2. Related Works

2.1. The iMiGUE Dataset

Recent progress in this area has been driven by the introduction of specialized datasets for micro-gesture understanding. In particular, iMiGUE is a large-scale video dataset introduced by Liu et al. [6] for identity-free micro-gesture understanding and emotion analysis. The iMiGUE dataset consists of video footage of tennis players during post-match interviews, with detailed frame-level annotations of various micro-gestures. The dataset contains 72 subjects (split into 37 for training and 35 for testing in a cross-subject protocol) and a total of 33 gesture classes.

One thing to note is that the class distribution is highly imbalanced. 28 of the 33 classes are tail classes with relatively few samples, meaning they collectively only make up less than 60% of the data. This long-tailed distribution, combined with the subtlety and high intra-class variability of micro-gestures, makes the recognition task extremely challenging [7].

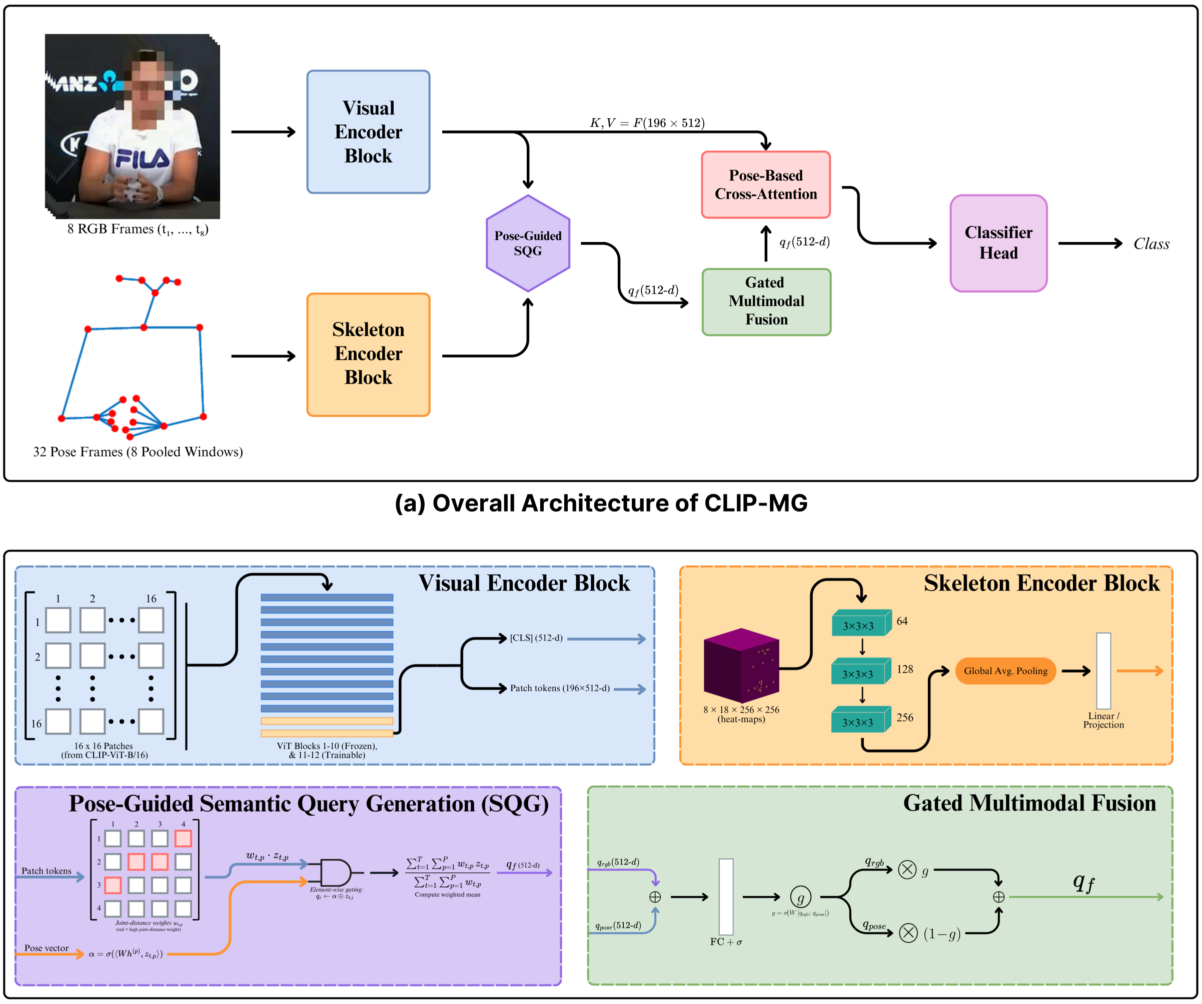
2.2. Signal Processing & Machine Learning Techniques for Micro-Gesture Recognition

To tackle these challenges, the research community has organized the Micro-Gesture Analysis (MiGA) challenges in recent years [7, 8]. These competitions have spurred the development of novel multi-modal approaches that leverage both video (RGB) and skeleton (pose) modalities for micro-gesture recognition. In the 2024 MiGA Challenge, for example, all top-performing methods integrated pose information alongside RGB frames. The winning entry by Chen et al. introduced a prototype-based learning approach with a two-stream 3D CNN (PoseConv3D) backbone for RGB and pose, cross-modal attention fusion, and a prototypical refinement component to calibrate ambiguous samples [9]. This method achieved a Top-1 accuracy of 70.25% on the iMiGUE test set, substantially outperforming earlier approaches. The second-place method by Huang et al. proposed a multi-scale heterogeneous ensemble network (M2HEN) combining a 3D convolutional model and a Transformer for feature diversity, reaching 70.19% accuracy [10]. Another notable approach by Wang et al. leveraged the vision-language model CLIP: they used a frozen CLIP as a teacher network for RGB frames and injected CLIP-derived text embeddings into a pose-based model, achieving 68.9% accuracy with an ensemble of RGB, joint, and limb pose streams. These efforts demonstrate that multi-modal fusion and semantic knowledge transfer are highly important in improving micro-gesture recognition [11].

2.3. CLIP-Based Video Understanding

Meanwhile, in the broader action recognition field, researchers have explored using pre-trained vision-language models like CLIP for video understanding. CLIP (Contrastive Language-Image Pre-training) [12] has shown powerful visual feature representations aligned with semantics via natural language supervision. However, straightforward fine-tuning of CLIP on video data can neglect smaller semantic information. To address this, Quan et al. proposed Semantic-Constrained CLIP (SC-CLIP) [13]. SC-CLIP adapts CLIP to video by generating a compact semantic query from dense visual tokens and using cross-attention to refocus the model on those action-relevant semantics. This ”constrains” CLIP's attention to discriminative features and yields stronger zero-shot and fine-grained recognition.

SC-CLIP demonstrates that directing attention to semantically meaningful regions can improve fine-grained video understanding. Micro-gestures, despite being low in their extent of movement, still take place with subtle visual semantics. Skeleton key-points give precise spatial-temporal anchors (hands, face, shoulders) that show where and when these cues take place. Thus, we design a pose-guided


8 RGB Frames (t₁, ..., t₈)
32 Pose Frames (8 Pooled Windows)
Visual Encoder Block
Skeleton Encoder Block
Pose-Guided SQG
$K, V = F(196 \times 512)$
Pose-Based Cross-Attention
$q_f(512\text{-}d)$
Gated Multimodal Fusion
$q_f(512\text{-}d)$
Classifier Head
Class
Visual Encoder Block
16 x 16 Patches (from CLIP-ViT-B/16)
ViT Blocks 1-10 (Frozen), & 11-12 (Trainable)
[CLS] (512-d)
Patch tokens (196×512-d)
Skeleton Encoder Block
8 × 18 × 256 × 256 (heat-maps)
3×3×3
64
3×3×3
128
3×3×3
256
Global Avg. Pooling
Linear / Projection
Pose-Guided Semantic Query Generation (SQG)
Patch tokens
Joint-distance weights $w_{t,p}$ (red = high joint-distance weight)
$w_{t,p} \cdot z_{t,p}$
Element-wise gating: $q_i \leftarrow \alpha \odot z_{t,i}$
$\frac{\sum_{t=1}^{T}\sum_{p=1}^{P} w_{t,p}\, z_{t,p}}{\sum_{t=1}^{T}\sum_{p=1}^{P} w_{t,p}}$
Compute weighted mean
q_f (512-d)
Pose vector
$\alpha = \sigma(\langle W h^{(p)}, z_{t,p}\rangle)$
Gated Multimodal Fusion
$q_{rgb}(512\text{-}d)$
$q_{pose}(512\text{-}d)$
FC + σ
$g = \sigma(W[q_{rgb}; q_{pose}])$
q_{rgb}
$\otimes\ g$
q_{pose}
$\otimes\ (1-g)$
q_f

(a) Overall Architecture of CLIP-MG

(b) Components Within CLIP-MG

Figure 1: Overview of the proposed architecture for micro-gesture classification. The CLIP-MG pipeline integrates a frozen CLIP ViT-B/16 visual encoder and an OpenPose-based skeleton encoder to generate pose-guided semantic queries. This focuses attention on image patches relevant to gestures. A gated multimodal fusion and cross-attention mechanism then blend the visual and pose features before a simple classification head predicts the micro-gesture label.

semantic attention mechanism that uses skeletal cues to steer CLIP towards where the gesture is taking place. This creates a query that captures the subtle semantics of micro-gestures.

3. Methodology

Our model (illustrated in Figure 1) has several components working in sequence: (1) a visual encoder (based on CLIP's vision transformer) processes the RGB frames, (2) a skeleton encoder processes the pose sequence, (3) pose-based semantic query generation producse semantic queries from visual features guided by pose features, (4) a gated fusion and semantics-based cross-attention fuses the modalities and improves the representation, and (5) a classification head outputs the predicted gesture label. In the following, we detail each component and the overall pipeline.

3.1. Visual Encoder

We adopt the OpenAI CLIP ViT-B/16 image tower [14] with the standard 224×224 input resolution and 16×16 patching, which yields $P = 196$ patch tokens plus one [CLS] token per frame. The internal transformer width is 768 dimensions, while CLIP's projection head maps the final [CLS] embeddings to a 512-dimensional space. From each micro-gesture clip we uniformly sample $T' = 8$ frames. Formally, for frame t we obtain the token sequence:

$$Z_t = \{\, z_{t,\mathrm{CLS}},\ z_{t,1},\ \ldots,\ z_{t,196} \,\}, \qquad z_{t,\mathrm{CLS}} \in \mathbb{R}^{768}\,.$$

During training we freeze the first 10 of the 12 ViT blocks and fine-tune only the last two blocks together with our added components [15]. Temporal information is aggregated by average-pooling the eight [CLS] embeddings to produce the per-clip visual feature.

3.2. Skeleton Encoder

We use the OpenPose format [16] to extract skeleton features. Given a clip, we first sample $T' = 32$ pose frames:

$$X^{(p)} = \{x_1^{(p)},\ \ldots,\ x_{32}^{(p)}\}, \qquad x_i^{(p)} \in \mathbb{R}^{18\times 2}\,.$$

To stay time-aligned with the eight RGB frames, the 32 pose frames are grouped into eight non-overlapping four-frame windows centered on $t_1, \ldots, t_8$. The heat-maps of each window are average-pooled along the temporal axis. This results in eight pose volumes that correspond one-to-one with the RGB inputs.

Each joint is then rasterized into a 256 × 256 canvas as a 2D Gaussian [17]:

$$H_{n,i,j}^{(t)} \ = \ \exp\!\left(-\frac{(i - x_n^{(t)})^2 + (j - y_n^{(t)})^2}{2\,\sigma^2} \right), \qquad \sigma = 2.5\,\mathrm{px},$$

where $(x_n^{(t)},\ y_n^{(t)})$ is the n-th joint of frame t. Stacking all joints and time-steps yields a 4D tensor

$$\mathscr{H} \ \in \ \mathbb{R}^{8\times 18\times 256\times 256}\,.$$

Utilizing an implementation similar to Tessa [18], we then employ a three-stage 3x3x3 convolutional network (with channel depths of 64, 128, and 256) to encode subtle pose dynamics [19]. We apply downsampling only in the spatial dimensions to preserve the motion details over time.

Global average pooling over (T', H, W) produces a 256-dimensional clip descriptor $h^{(p)}$. A linear projection:

$$W_p \ \in \ \mathbb{R}^{512\times 256}$$

maps it to $d = 512$ so the pose feature matches the CLIP visual dimension $D = 512$:

$$\tilde{h}^{(p)} \ = \ W_p\, h^{(p)} \ \in \ \mathbb{R}^{512}\,.$$

3.3. Pose-Guided Semantic Query Generation

The proposed semantic-query approach extracts a representation of the video's most important cues with the support of the skeleton features. It does so (1) spatially, by concentrating on visual tokens that are near body parts exhibiting motion, and (2) temporally, by giving higher weight to frames where the pose dynamics show that a micro-gesture is taking place.

Let $z_{t,p}$ be the set of patch embeddings from all selected frames (excluding the global tokens). We first identify a subset of these visual tokens that are relevant to the micro-gesture. ”Pose guidance” is applied by using the skeleton features to weight or select visual tokens:

- We compute an attention mask over image patches based on the distance of each patch to the nearest skeletal joint position. If a patch lies close to a joint that is moving significantly, it receives a higher weight. For example, if $j_{t,k}$ are the coordinates of joint k in frame t, we can define a relevance score:

$$w_{t,p} = \exp(-\min_{k} |\mathrm{pos}(z_{t,p}) - j_{t,k}|^2/\sigma^2)$$

where $\mathrm{pos}(z_{t,p})$ is the spatial location of patch p and σ controls the spatial scale. This yields weights $w_{t,p} \in [0,1]$ that highlight patches near active joints.

- Additionally, we leverage the skeleton encoder's output $h^{(p)}$ as a global descriptor of the motion. We project $h^{(p)}$ to the same dimension D and use it to modulate the visual tokens via a simple gating:

  $$\tilde{z}_{t,p} = \alpha\, z_{t,p},$$
  $$\text{where} \quad \alpha = \sigma\big(\langle W h^{(p)},\, z_{t,p}\rangle\big).$$

  Here α is the sigmoid of the dot-product between the projected pose feature $Wh^{(p)}$ and the visual token $z_{t,p}$, so it down-weights any token not well aligned with the pose direction.

After computing the pose-based weights $w_{t,p}$, we flatten the full set of visual tokens

$$\big\{\, z_{t,p} \mid t = 1,\dots,T,\ \ p = 1,\dots,P \big\},$$

with $T = 8$, $P = 196$, and $N = T \times P = 1568$. We then aggregate them into a single D-dimensional semantic query $q \in \mathbb{R}^{512}$ by weighted mean pooling:

$$q \;=\; \frac{\displaystyle\sum_{t=1}^{T}\sum_{p=1}^{P} w_{t,p}\, z_{t,p}}{\displaystyle\sum_{t=1}^{T}\sum_{p=1}^{P} w_{t,p}}\,.$$

This pose-weighted query q thus encapsulates the most relevant gesture semantics and is passed to the cross-attention component to guide the final feature fusion and classification.

3.4. Gated Multi-modal Fusion

Before feeding the query into the CLIP transformer, we further integrate the pose information via a gated fusion mechanism inspired by Arevalo et al. [20]. The goal here is to merge the skeleton representation with the visual representation in a way that the model can selectively attend to one or the other modality as needed. We implement gated fusion at two points in the pipeline:

- We fuse the skeleton encoder output $h^{(p)}$ with the semantic query q. First we compute a gating vector

  $$g = \sigma\big(W_g\, h^{(p)}\big),$$

  where $W_g \in \mathbb{R}^{D \times d}$ is a learned projection and σ is the sigmoid. We then modulate the query by

  $$q_f = q \odot g \;+\; q \odot (1 - g).$$

  In our implementation g is element-wise, so each feature of q is scaled into $[0,1]$, allowing pose-aligned dimensions to be amplified or suppressed.
- Similarly, we fuse the pose descriptor $h^{(p)}$ into the CLIP encoder's intermediate token sequence. Let

  $$F = \{\, f_1,\, f_2,\, \dots,\, f_N\}$$

  be the set of visual features from CLIP's penultimate layer (these serve as keys and values in cross-attention). We then compute a second gating vector

  $$u = \sigma\big(W_u\, h^{(p)}\big) \;\in\; \mathbb{R}^{D},$$

and apply it element-wise:

$$\tilde{f}_i = f_i \odot u, \quad i = 1, \ldots, N.$$

This global gate highlights or suppresses certain channels based on pose.

These gating operations are learned end-to-end and ensure that the multi-modal information is blended before the cross-attention step. The gating is soft (continuous values between 0 and 1), so the model can learn to rely on pose heavily in some scenarios or ignore it in others. This adaptability is important because pose data can sometimes be noisy or incomplete (e.g., occluded joints), so a static fusion might hurt performance if pose is trusted blindly. Our gated fusion allows the network to fall back to visual cues when pose is uncertain, and vice versa.

3.5. Cross-Attention with Semantic Query

Next, we apply a cross-attention mechanism driven by our pose-guided query. We insert the query vector q_f as an extra token into the final transformer layer of the CLIP visual encoder, allowing it to attend over the gated visual token set F. This focused attention refines the representation by pooling features most relevant to the detected micro-gesture.

Consider the transformer architecture of CLIP's visual encoder. Let

$$F = \{\tilde{f}_1, \ldots, \tilde{f}_N\}, \quad K = V = F, \quad \tilde{f}_i \in \mathbb{R}^D, \quad Q \in \mathbb{R}^{1 \times D}.$$

We then insert our pose-guided query q_f into the final layer and compute cross-attention:

$$q_{\text{out}} \;=\; A(Q, K, V) \;=\; \text{softmax}\Big(\frac{QK^\top}{\sqrt{D}}\Big) V,$$

which yields $q_{\text{out}} \in \mathbb{R}^{1 \times D}$.

The cross-attention computes an output query embedding q_{out} that is a weighted sum of the values V, with weights determined by the compatibility of Q with keys K. Mathematically, if we denote Q (1×D), K (N×D), V (N×D), the attention is:

$$A(Q, K, V) \;=\; \text{softmax}\Big(\frac{QK^\top}{\sqrt{D}}\Big) V.$$

where the softmax produces a $1 \times N$ vector of attention weights. The resulting q_{out} (of dimension 1×D) is effectively a semantic-aware video representation that has "pooled" information from the visual tokens, biased by the semantic content of Q and thereby by the pose cues we injected. In other words, q_{out} should ideally encode the crucial features needed to distinguish the micro-gesture class.

This unique attention mechanism forces the model to concentrate on what is important for the gesture. It acts as a form of feature selection: among the many visual features of a scene (some possibly irrelevant background or person identity cues), it emphasizes those that correlate with the action semantics. In our case, because Q was guided by pose, the attention is further narrowed to regions of actual motion or posture change.

After cross-attention, we obtain q_{out} which we consider as the fused video representation for the whole clip.

3.6. Classification and Training Objective

The final stage is the classification of the micro-gesture. We feed the fused representation q_{out} (dimension D) into a classifier head, implemented as a simple two-layer MLP followed by softmax. This yields a probability distribution $\hat{y} \in \mathbb{R}^C$ over the C gesture classes (here $C = 33$ for iMiGUE). We train the model using a supervised classification objective. The primary loss is the cross-entropy between the predicted distribution and the ground-truth label. Given a training sample i with true class label y_i (represented as a one-hot vector) and predicted probabilities $\hat{y}_i$, the loss is:

Table 1
Comparison of prior micro-gesture classifiers.

Reference	Method	Top-1 (%)
[21]	GCN + Skeleton (ST-GCN)	46.38
[22]	Multi-scale GCN + Skeleton (MS-G3D)	52.00
[23]	Temporal Relational + RGB (TRN)	55.24
[24]	Temporal Shift + RGB (TSM)	58.77
[19]	3D CNN + Skeleton Heatmaps (PoseConv3D)	61.11
[25]	Vision Transformer + RGB (Video Swin-B)	61.73
[26]	Dense-Sparse Fusion + RGB+Skeleton (DSCNet)	62.50
[11]	CLIP Distillation + Skeleton 3DCNN	68.90
[10]	Multi-scale Ensemble + RGB+Skeleton (M2HEN)	70.19
[9]	Prototype-based GCN + Skeleton	70.25
—	**Pose-guided Semantic Attention + CLIP + Skeleton (CLIP-MG, ours)**	**61.82**

$$L_{\text{cls}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\,\log\big(\hat{y}_{i,c}\big)\,.$$

where N is the number of training examples in a batch and $y_{i,c} \in 0, 1, \sum_c y_{i,c} = 1$. We minimize this loss with respect to the parameters of the skeleton encoder, the fusion components, the classifier, and the parts of the CLIP encoder we allow to be fine-tuned.

4. Experiments

4.1. Dataset and Evaluation Protocol

We conduct experiments on the iMiGUE dataset, focusing on the micro-gesture classification task. As described earlier, iMiGUE contains 33 micro-gesture classes collected from interview videos of tennis players. These gestures include subtle body-language cues such as pressing lips or touching one's jaw. We set aside a portion of the training data (20%) to serve as a local validation set for our experiments. This local validation is used for model selection and ablation studies due to the unavailability of a separate testing environment at the time of experimentation. However, final results on the test set are referenced for comparison with other approaches [8, 7].

4.2. Results and Comparisons

As shown in Table 1, CLIP-MG achieves a Top-1 accuracy of 61.82%, outperforming a range of single-modality baselines [21, 22, 23, 24]. Notably, CLIP-MG even performs on par with standalone architectures such as Video Swin-B [25] and PoseConv3D [19] despite using a largely frozen CLIP backbone and a compact pose encoder. This shows that steering CLIP's attention with pose-guided semantic queries yields more discriminative features for fine-grained micro-gestures. However, the proposed architecture does not set a new state-of-the-art in this area. It comes close in performance to the dense-sparse fusion network DSCNet [26] (62.50%), but falls behind architectures presented in previous editions of the MiGA challenge [9, 10, 11]. These results motivate future work to explore richer query adaptation and improved temporal fusion to close the gap with top models.

4.3. Ablation Study

We performed comprehensive ablation experiments to validate the contribution of each component in CLIP-MG. Table 2 reports Top-1 accuracy on our validation split.

Table 2
Ablation study on CLIP-MG components

Variant	Description	Top-1 (%)	Δ (pp)
w/o Pose branch	Visual-only cross-attention	45.30	−16.52
w/o Pose guidance	Visual query + cross-attention	51.23	−10.59
w/o Cross-attention	Concat(CLIP CLS, pose)	53.17	−8.65
w/o Gated fusion	Pose query + cross-attn without gating	60.08	−1.74
Full model	**Pose-guided Semantic Attention + CLIP + Skeleton**	**61.82**	—

4.3.1. Without Pose Branch

Here, we completely eliminate the pose branch to see the benefit of adding pose at all. The model gave 45.30% (−16.52 pp) accuracy. Thus, adding the pose branch (with our fusion and guidance) yields a 16.52% gain, which demonstrates that skeleton data carries complementary information for the task.

4.3.2. Without Pose Guidance

In this variant, we remove the pose influence from the semantic query generation. The query is generated purely by clustering visual tokens without using skeleton data. The semantics-based cross-attention still operates, but only on the visual-based query. We found that the accuracy dropped to 51.23% (−10.59 pp). This confirms that pose guidance is essential and provides a significant boost. This makes sense because, in theory, without pose the model may attend to irrelevant semantics or background context. This misses subtle gesture cues.

4.3.3. Without Semantic Cross-Attention

Here, we skip SCCA. Instead, we simply concatenate the global visual [CLS] embedding with the pose feature and feed that to a classifier. This essentially tests a late-fusion approach without our semantic query mechanism. The accuracy was 53.17% (−8.65 pp). This indicates that the semantic query and cross-attention are effective at focusing on important features that a flat concatenation would miss.

4.3.4. Without Gated Fusion

In this ablation, we disable the gating in both the query generation and the visual token modulation. We still generate a query using pose (via simple concatenation of average visual token and pose feature) and perform cross-attention. The accuracy achieved we 60.08% (−1.74 pp), a modest drop. This shows that gating helps but is not as critical as the presence of pose info or semantics-based cross-attention. The gating mostly fine-tunes the balance between modalities.

4.3.5. Discussion

These ablations show that each component of CLIP-MG plays a supporting role, although certain components are more important than others. Dropping the entire pose branch drives accuracy down to 45.30% (−16.52 pp). This demonstrates how much discriminative information there is within the skeletal signal. Removing pose guidance lowers accuracy from 61.82% to 51.23% (−10.59 pp), showing that skeletal information is very important for localizing subtle joint motions, as they act as an attention prior [27] that focuses the visual stream towards the regions where micro-gestures occur. Eliminating cross-attention drops accuracy to 53.17% (−8.65 pp), which indicates that without a mechanism to selectively pool pose-weighted tokens, the model may struggle to tell apart very similar gestures. Finally, disabling gated fusion yields 60.08% (−1.74 pp), which indicates that adaptively balancing pose and visual information slightly improves the robustness of the architecture.

Taken together, these results show how the different components effectively complement each other. Pose cues localize the gesture, cross-attention extracts the relevant semantics, and gating balances both streams. We find the highest performance when all the components are combined.

5. Conclusions and Future Works

We introduced CLIP-MG, a pose-guided, multi-modal CLIP architecture for micro-gesture recognition on the iMiGUE benchmark. By guiding CLIP's visual attention with skeleton-based spatial priors, generating compact semantic queries, and fusing pose and appearance via a learnable gate, CLIP-MG extracts subtle, discriminative features that simple RGB or pose-only models cannot recognize. Our model achieves 61.82% Top-1 accuracy, outperforming most single-modality baselines and performing on par with strong 3D-CNN and vision-transformer approaches. Extensive ablation studies confirm that each component provides a measurable benefit. The experiments provide insights into how each component interacts with and complements others, which highlights important design patterns that can inform future model development in micro-gesture classification and similar fine-grained recognition tasks. Our findings demonstrate the value of integrating multimodal and semantic information to address challenging visual recognition problems.

Our future work will explore richer temporal approaches and data strategies to close the gap between CLIP-MG and more recent state-of-the-art models [9]. First, integrating sequence models (temporal transformers or recurrent layers over cross-attention outputs) should capture patterns that static sampling loses. Second, video motion magnification [28] could amplify imperceptible movements. This would help with pose tracking and visual encoding. Third, joint pre-training on related action-gesture datasets and weakly- or self-supervised learning could improve feature robustness [29]. Finally, regarding accuracy, we plan to incorporate uncertainty-aware gating for noisy skeletons and class-balanced or prototype-based calibrations to address long-tail imbalance [30]. To improve and better evaluate the explainability of the model, we will incorporate gradient-weighted class activation mapping (Grad-CAM) [31] and more recent attention-aware token-filtering approaches [32]. Currently, the proposed architecture suffers heavily due to its relatively low speed on commodity hardware. To address this issue, we plan to experiment with several multimodal compression and optimization algorithms for more efficient computing [33, 34, 35, 36]. We plan to train an improved version of our architecture for downstream tasks on the DAIC-WoZ dataset [5, 37, 38] for low-level mental health analysis. These different research directions are promising in pushing pose-guided CLIP models closer to (and beyond) human-level understanding of the subtlest gestures.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4o to edit the paper, checking for grammar and spelling mistakes. GPT-4o was also utilized to revise the draft for brevity and improved flow. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] J. F. Cohn, P. Ekman, Observing and coding facial expression of emotion, Handbook of Emotion (2000).

[2] M. Funes, et al., Micro-expression recognition: A survey, in: FG, 2019.

[3] M. Pantic, Affective multimedia databases: Affective video databases, Handbook of Affective Computing (2009).

[4] A. Kapoor, R. Picard, Automatic prediction of human behavior in social settings, in: IUI, 2007.

[5] S. V. Patapati, Integrating large language models into a tri-modal architecture for automated depression classification, 2024. arXiv:2407.19340v5, preprint.

[6] X. Liu, H. Shi, H. Chen, Z. Yu, X. Li, G. Zhao, imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis, in: CVPR, 2021.

[7] C. Haoyu, et al., The 2nd challenge on micro-gesture analysis for hidden emotion understanding (miga) 2024: Dataset and results, in: MiGA 2024: Proceedings of IJCAI 2024 WorkshopChallenge on Micro-gesture Analysis for Hidden Emotion Understanding (MiGA 2024) co-located with 33rd International Joint Conference on Artificial Intelligence (IJCAI 2024), 2024.

[8] G. Zhao, et al., The workshop challenge on micro-gesture analysis for hidden emotion understanding (miga), in: MiGA 2023: Proceedings of IJCAI 2023 WorkshopChallenge on Micro-gesture Analysis for Hidden Emotion Understanding (MiGA 2023) co-located with 32nd International Joint Conference on Artificial Intelligence (IJCAI 2023), 2023.

[9] G. Chen, et al., Prototype learning for micro-gesture classification, in: MiGA 2024: Proceedings of IJCAI 2024 WorkshopChallenge on Micro-gesture Analysis for Hidden Emotion Understanding (MiGA 2024) co-located with 33rd International Joint Conference on Artificial Intelligence (IJCAI 2024), 2024.

[10] H. Huang, et al., Multi-modal micro-gesture classification via multi-scale heterogeneous ensemble network, in: MiGA 2024: Proceedings of IJCAI 2024 WorkshopChallenge on Micro-gesture Analysis for Hidden Emotion Understanding (MiGA 2024) co-located with 33rd International Joint Conference on Artificial Intelligence (IJCAI 2024), 2024.

[11] Y. Wang, et al., A multimodal micro-gesture classification model based on clip, in: MiGA 2024: Proceedings of IJCAI 2024 WorkshopChallenge on Micro-gesture Analysis for Hidden Emotion Understanding (MiGA 2024) co-located with 33rd International Joint Conference on Artificial Intelligence (IJCAI 2024), 2024.

[12] A. Radford, et al., Learning transferable visual models from natural language supervision, ICML (2021).

[13] Z. Quan, et al., Semantic matters: A constrained approach for zero-shot video action recognition, in: Pattern Recognition, 2025.

[14] A. Dosovitskiy, et al., An image is worth 16x16 words: Transformers for image recognition at scale, in: ICLR, 2021.

[15] X. H., et al., Videoclip: Learning video representations from text and clips, arXiv (2021).

[16] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, in: CVPR, 2017.

[17] X. Zhou, et al., On heatmap representation for 6d pose estimation, in: ICCV, 2019.

[18] S. V. Patapati, T. Srinivasan, H. Musku, A. Adiraju, A framework for eca-based psychotherapy, 2025.

[19] J. Zhang, Z. Huang, Y. Chen, Poseconv3d: Revisiting skeleton-based action recognition, in: European Conference on Computer Vision (ECCV), 2020.

[20] J. Arevalo, et al., Gated multimodal units for information fusion, in: ICLR, 2020.

[21] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: AAAI Conference on Artificial Intelligence, 2018.

[22] S. Liu, et al., MS-G3D: Multi-scale graph convolution for skeleton-based action recognition, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[23] B. Zhou, A. Andonian, A. Oliva, A. Torralba, Temporal relational reasoning in videos, in: European Conference on Computer Vision (ECCV), 2018.

[24] J. Lin, C. Gan, S. Han, TSM: Temporal shift module for efficient video understanding, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[25] Z. Liu, et al., Video swin transformer: Hierarchical vision transformer for video recognition, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2022.

[26] Q. Cheng, et al., DSCNet: Dense-sparse complementary network for human action recognition, Expert Systems with Applications (2024).

[27] H. Zhang, et al., Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition, CVPR (2019).

[28] N. Wadhwa, et al., Eulerian video magnification for revealing subtle changes in the world, in:

SIGGRAPH, 2013.

[29] T. Han, et al., Self-supervised video representation learning with neighborhood context aggregation, ECCV (2020).

[30] B. Kang, et al., Decoupling representation and classifier for long-tail recognition, in: ICLR, 2020.

[31] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, 2017, pp. 618–626.

[32] T. Naruko, H. Akutsu, Speed-up of vision transformer models by attention-aware token filtering, 2025. arXiv:2506.01519v1, preprint.

[33] Y. Omri, P. Shroff, T. Tambe, Token sequence compression for efficient multimodal computing, 2025. arXiv:2504.17892v1.

[34] L. Lei, J. Gu, X. Ma, C. Tang, J. Chen, T. Xu, Generic token compression in multimodal large language models from an explainability perspective, 2025. arXiv:2506.01097.

[35] X. Tan, P. Ye, C. Tu, J. Cao, Y. Yang, L. Zhang, D. Zhou, T. Chen, Tokencarve: Information-preserving visual token compression in multimodal large language models, 2025. arXiv:2503.10501.

[36] J. Cao, P. Ye, S. Li, C. Yu, Y. Tang, J. Lu, T. Chen, Madtp: Multimodal alignment-guided dynamic token pruning for accelerating vision-language transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, p. –. doi:10.1109/CVPR.2024.00XX.

[37] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo, L.-P. Morency, The distress analysis interview corpus of human and computer interviews, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC), European Language Resources Association (ELRA), 2014, pp. 3123–3128.

[38] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, S. Song, S. Liu, Z. Zhao, A. Mallol-Ragolta, Z. Ren, M. Soleymani, M. Pantic, Avec 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition, in: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, AVEC '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 3–12. URL: https://doi.org/10.1145/3347320.3357688. doi:10.1145/3347320.3357688.