Cross-view Multimodal Vision-Based Assessment Framework for Traditional Chinese Medicine Rehabilitation Training

Francis Xiatian Zhang*, *Member, IEEE,* Hao Yao*, Shengxuan Chen*, Hong Zhu, Hongxiao Jia, Sisi Zheng†, Hubert P. H. Shum†, *Senior Member, IEEE*

***Abstract*— Vision-based assessment can provide convenient and cost-effective evaluation in Traditional Chinese Medicine (TCM) rehabilitation training, where action quality assessment (AQA) from computer vision offers a promising solution. Existing automatic AQA frameworks for physical therapy typically rely on skeletal data captured from a single viewpoint, which is inefficient for TCM techniques such as acupuncture or Tuina that involve dense hand self-occlusion and complex hand–object interactions. To address these challenges, we propose CME-AQA, a cross-view, multimodal vision-based assessment framework that integrates visual–pose fusion to enhance understanding of environmental context and leverages both first-person and third-person videos during training to improve inference robustness. We collected two dual-view datasets, TCM-AQA61-A (Acupuncture) and TCM-AQA61-T (Tuina), each containing synchronized first- and third-person recordings of 61 subjects with expert annotations. Experimental results show that our approach achieves superior or comparable mean performance against competitive baselines, achieving over 10% relative improvement in weighted F1 over the best competing method on key rating tasks such as Needle Depth and Quick Needle Insertion, while also reducing mean absolute error in quantitative measures such as insertion time and manipulation frequency. Testing on a CPR dataset further demonstrates comparable performance on several posture-based criteria, suggesting applicability to related structured simulated clinical skill assessments where participant motion is central to evaluation. Overall, CME-AQA enhances assessment accuracy for structured TCM rehabilitation training and facilitates more convenient and effective training-oriented skill evaluation.**

F. X. Zhang and H. P. H. Shum are with the Department of Computer Science, Durham University, UK (e-mail: {xiatian.zhang, hubert.shum}@durham.ac.uk). F. X. Zhang is also with the Institute for Regeneration and Repair, The University of Edinburgh, UK (e-mail: francis.zhang@ed.ac.uk).

H. Yao is with Ningbo Hospital of Traditional Chinese Medicine, Ningbo, China (e-mail: 1242447006@qq.com).

S. Chen is with the Department of Rehabilitation Medicine, The Gulou Hospital of Traditional Chinese Medicine, China (e-mail: 13693012323@163.com).

S. Zheng, H. Zhu and H. Jia are with The National Clinical Research Center for Mental Disorders and Beijing Key Laboratory of Mental Disorders, Beijing Anding Hospital, China (e-mail: {zhengsisi, zhuhong, jhxlj}@ccmu.edu.cn). S. Zheng, H. Zhu and H. Jia are also with the Advanced Innovation Center for Human Brain Protection, Beijing Anding Hospital, China.

F. X. Zhang, H. Yao, and S. Chen share co-first authorship (*) of this paper.

†Corresponding authors: S. Zheng and H. P. H. Shum.

***Index Terms*— Action Quality Assessment, Machine Learning, Vision-based Rehabilitation Assessment, Traditional Chinese Medicine**

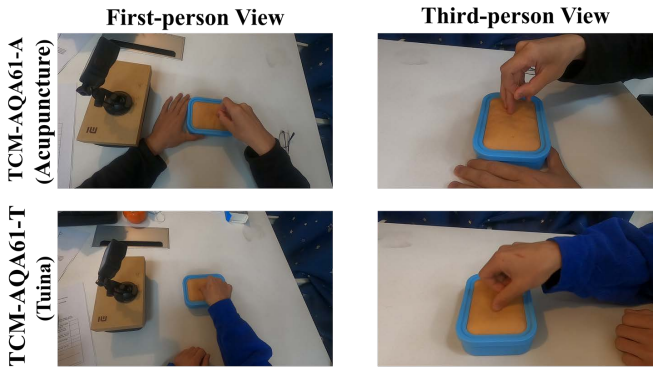

Fig. 1. Example frames from our two collected datasets: acupuncture (upper rows) and Tuina (lower rows). Each session was recorded with two cameras: a forehead-mounted first-person view (left column) and a separate third-person view of the hands (right column), providing complementary participant and observer perspectives.

I. Introduction

Traditional Chinese Medicine (TCM) rehabilitation therapies, such as acupuncture and Tuina (Chinese therapeutic massage), have demonstrated measurable effectiveness in treating motor and neurological disorders [1], [2]. Recent advancements in TCM rehabilitation training [3], [4], [5] have introduced automatic assessment tools to evaluate performance in structured TCM rehabilitation training settings. These tools enhance accessibility to training, especially in developing regions where experienced instructors are scarce. However, many current systems rely on specialized equipment, such as VR glasses [5] or robotic platforms [4], to assess practice behavior, which limits their practicality. To address this limitation, we propose a vision-based assessment framework that derives performance metrics directly from video, providing a cost-effective and portable solution.

In human-performance assessment [6], action quality assessment (AQA) is a widely adopted method for evaluating practice actions across multiple predefined aspects [7]. It applies computer vision techniques to quantify the quality of

movements performed during training and has proven effective in both sports and rehabilitation contexts [8], [6]. Most existing frameworks rely on pose-based skeletal data obtained through sensor-based or markerless motion capture to represent body configurations and dynamics, which helps reduce environmental noise and capture motion details accurately [9]. However, current AQA methods [10], [11], [12] mainly analyze full-body movements from a single-view perspective, which presents a major limitation for TCM rehabilitation tasks. These procedures require precise hand movements that are frequently obscured by self-occlusion [13]. Moreover, using only skeletal data omits key environmental cues that are critical for modeling hand–object interactions, for example, the manipulation of acupuncture needles on practice pads [14].

To design an effective assessment framework for TCM rehabilitation training, we identify three key challenges in existing AQA methods. First, current approaches [10], [11], [15] primarily depend on pose estimation data, which often fail to capture the environmental context where the hands interact with objects and surfaces during TCM procedures. This limitation reduces the capacity of AQA models to evaluate all clinically relevant aspects of rehabilitation performance. Second, existing AQA frameworks [16], [17], typically designed for single-view video inputs, struggle to handle frequent self-occlusions in which one hand or tool obscures another. Such occlusions can compromise the reliability and continuity of performance evaluation. Third, to the best of our knowledge, no publicly available dataset currently exists for AQA in TCM rehabilitation [7], making it difficult to establish a standardized benchmark and hindering progress in this research area.

To address these challenges, we propose the Cross-view Multimodality Enhanced Action Quality Assessment (CME-AQA) framework for Traditional Chinese Medicine rehabilitation training. The framework integrates both visual and skeletal information from multiple viewpoints to comprehensively capture participant movements and their interactions with the environment. Attention-based fusion [18] and multi-view learning [19] have been explored in prior AQA studies. CME-AQA advances TCM rehabilitation assessment through a task-specific integration of these mechanisms, together with dedicated dual-view benchmark datasets for this domain.

The CME-AQA framework comprises two main components. The first is the Attention-based Visual–Pose Fusion (AVPF) module, which fuses visual and pose features through an attention mechanism to generate a unified and informative representation of motion dynamics and contextual cues. The second is the Multiscale View Alignment (MVA) training strategy, which leverages both first-person and third-person videos during training to establish multi-view awareness while allowing single-view inference during deployment. In addition, to address the lack of publicly available datasets and to establish a benchmark for this task, we created two synchronized multi-view datasets, shown in Fig. 1: TCM-AQA61-A (Acupuncture) and TCM-AQA61-T (Tuina). Each dataset includes expert-annotated ratings from 61 subjects, recorded simultaneously from first-person and third-person viewpoints, providing a valuable resource for future research in TCM rehabilitation assessment.

With the proposed datasets, we conducted a comprehensive benchmark study comparing existing action quality assessment methods for physical therapy and validating the effectiveness of our CME-AQA framework. The experiments demonstrate that our approach achieves over 10% relative improvement in weighted F1 over the strongest competing baseline on key tasks such as Needle Depth and Quick Needle Insertion, while maintaining competitive performance across other metrics. These results highlight the capability of the CME-AQA framework to enhance TCM rehabilitation training through more precise and reliable measurement. Furthermore, our method reduced error by approximately 4% in metrological evaluations such as insertion time and manipulation frequency, demonstrating its effectiveness in both categorical and continuous assessment tasks. To further examine applicability beyond the proposed TCM tasks, we evaluated CME-AQA on a Cardiopulmonary Resuscitation (CPR) dataset [20], where it achieved measurement performance comparable to human experts on several posture-based criteria, suggesting applicability to related structured simulated clinical skill assessments where participant motion is central to evaluation.

The source code and proposed datasets are publicly available on GitHub[1]. For transparency, we note that parts of the Methodology section adapt content from the author's PhD thesis [21] and were also presented in the MICCAI 2025 Doctoral Consortium, with permission.

Our main contributions are summarized as follows:

1) We propose CME-AQA, a cross-view multimodal action quality assessment framework for TCM rehabilitation training, and introduce two synchronized dual-view benchmarks, TCM-AQA61-A (Acupuncture) and TCM-AQA61-T (Tuina), with expert annotations for both categorical and continuous skill indicators, enabling standardized evaluation of fine-grained hand-centric procedures in structured TCM rehabilitation training settings.
2) We design an attention-based visual-pose fusion module (AVPF) that performs modality-asymmetric pose-conditioned cross-attention, refining visual representations using pose as a stable conditioning signal to strengthen interaction-aware RGB-pose fusion for hand-object modeling under self-occlusion in fine-grained TCM procedures.
3) We introduce a multiscale view alignment (MVA) training strategy that enforces hierarchical cross-view consistency between egocentric and exocentric representations during training, while retaining single-view inference for practical deployment in routine rehabilitation training settings.

II. Related Works

A. Action Quality Assessment

Action Quality Assessment (AQA)evaluates action execution quality from video by predicting fine-grained performance scores [22]. Unlike action recognition, which predicts categorical labels, AQA requires modeling subtle performance variations among visually similar actions. Early methods relied on handcrafted motion descriptors or pose-based fea-

[1] https://github.com/FrancisXZhang/cme-aqa

tures [23], followed by convolutional and recurrent architectures for spatiotemporal representation learning [24]. Recent approaches adopt transformer-based architectures to capture long-range temporal dependencies and structured action dynamics. FineParser [25] introduces human-centric spatial parsing with step-level temporal decomposition and contrastive regression for score alignment. Uni-FineParser [26] incorporates mask-supervised attention to enhance human-region modeling while remaining video-only at inference. Complementary to geometry-aware parsing, PHI [27] addresses domain shift in long-term AQA through hierarchical feature flow modeling and instruction-guided adaptation, bridging action-recognition pretraining and quality assessment without explicit spatial supervision.

Beyond architectural adaptation, recent work explores richer spatial supervision and multimodal cues for AQA, particularly for geometric reasoning. AIFit [28] integrates multi-view RGB with 3D motion-capture and body-shape supervision to provide localized, interpretable feedback. PGT [18] incorporates pose-derived spatial priors and global–local feature decomposition for body-centric modeling. LucidAction [19] introduces a hierarchical multi-view and multimodal dataset, where multi-view signals are primarily leveraged for joint training rather than explicit cross-view alignment. However, most existing frameworks focus on global full-body analysis, whereas clinical TCM procedures require modeling localized hand–object interaction under frequent self-occlusion. In such settings, complementary views need to be explicitly aligned to enforce cross-view consistency. Without structured alignment, multi-view signals may be underutilized in tasks involving subtle manipulation and strong contextual dependency.

B. Action Quality Assessment in Rehabilitation

Building upon general AQA frameworks, recent studies have extended assessment to rehabilitation training scenarios [10], [29], [30]. In the rehabilitation domain, several frameworks [10], [15], [12] employ AQA-based approaches that analyze human skeletal poses from video to assess movement quality and technique, thereby reducing reliance on specialized kinematic measurement devices [31], [32]. For example, Zheng *et al.* [12] compute a dot product matrix across frames to quantify joint rotations in skeletal sequences, enabling more precise modeling of rehabilitation quality. More recently, these works have inspired the development of AQA methods for TCM-related skills such as Tai Chi [33], [34] and Qigong [35], primarily based on human pose representations. For instance, Li *et al.* [33] reconstruct 3D Tai Chi motion through multi-view pose fusion and quantify motion deviations by comparing participants' movements against a reference template. Zhao *et al.* [34] leverage graph neural networks on skeletal representations for Tai Chi movement evaluation.

However, these pose-only frameworks [10], [15], [12] often overlook visual cues that are critical in complex TCM rehabilitation practices such as acupuncture, where key movements involve fine-grained interactions among the hand, needle, and practice pad [4]. This omission limits their ability to capture contextual information necessary for accurate inference. Moreover, reliance on single-view pose estimation makes them vulnerable to self-occlusion, a common challenge in TCM therapies characterized by intricate hand motions [36]. These limitations highlight the need for a tailored framework that integrates visual–pose fusion and multi-view awareness for reliable assessment of TCM rehabilitation performance.

TABLE I
SUMMARY OF EXISTING DATASETS FOR REHABILITATION AQA

Dataset	Subject	View	Therapy Type
UI-PRMD [41]	10	1	Common Rehabilitation Exercise
KIMORE [42]	78	1	Low-back Pain Exercise
IntelliRehabDS [43]	29	1	Common Rehabilitation Exercise
Keraal [44]	31	1	Low-back Pain Exercise
FineRehab [45]	50	2	Musculoskeletal exercises
TCM-AQA61-A	61	2	Acupuncture
TCM-AQA61-T	61	2	Tuina

C. Multi-view Clinical Video Analysis

To mitigate self-occlusion in complex clinical settings, recent video analysis frameworks for clinical applications, such as patient or clinician pose estimation [37], [38], surgical action recognition [39], and AQA [40], [20], increasingly employ multi-view camera setups to capture participant movements more comprehensively. These frameworks [37], [38], [39], [40], [20] typically extract visual or pose features from each video view and fuse them using learnable methods, such as weighted feature fusion across views [20], [39]. For example, Constable et al. [20] proposed a framework that first extracts the pose from each view of CPR training videos, then uses graph convolutions to learn joint relationships, and subsequently employs learnable weights to fuse the features of each view to determine the final rating label. However, these methods [37], [38], [39], [40], [20] depend on complex multi-view setups at inference, which hinders deployment in real-world training environments. Therefore, an effective AQA framework for TCM rehabilitation should be designed to operate with single-view input during inference while leveraging knowledge learned from multi-view training.

D. Existing Datasets for Rehabilitation AQA

Most existing rehabilitation AQA datasets rely on single-view recordings [41], [42], [43], [44], which limits the ability to train models with multi-view awareness. Table I summarizes several representative datasets. Although the recently introduced FineRehab dataset [45] includes multi-view recordings, it primarily focuses on full-body patient exercises rather than fine-grained therapist actions. Consequently, current datasets are well-suited for general rehabilitation exercises but do not capture the intricate hand movements and frequent self-occlusions typical of TCM practices such as acupuncture and Tuina [14]. To advance skill assessment in this domain, multi-view datasets specifically designed for TCM rehabilitation are essential, enabling models to learn from both participant hand motions and contextual interactions with tools and surfaces.

III. DATA COLLECTION

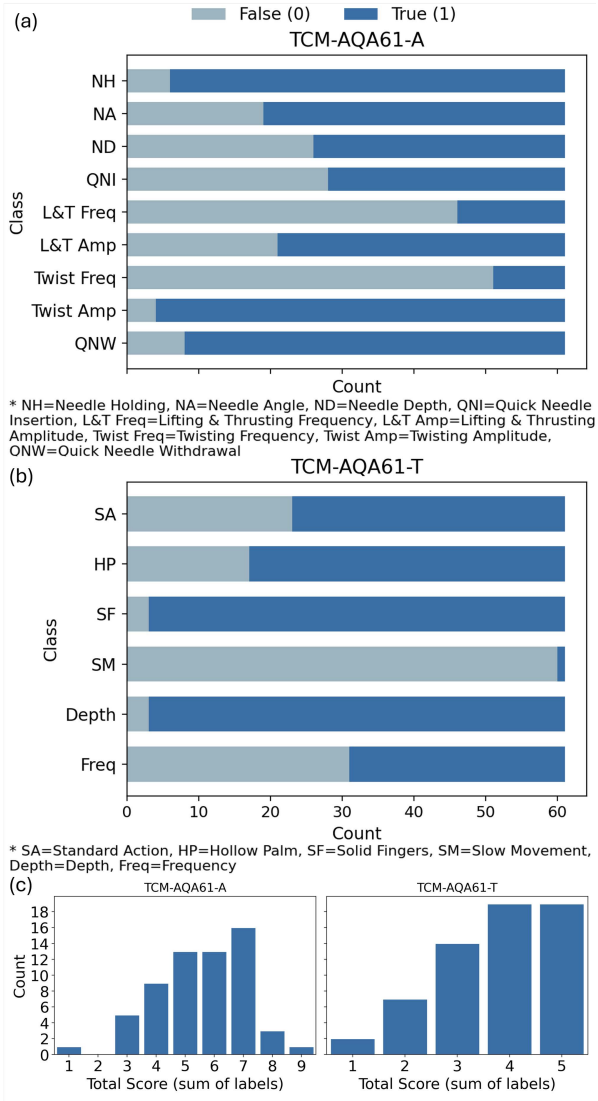

Fig. 2. Dataset statistics and label distribution for TCM-AQA61-A (Acupuncture) and TCM-AQA61-T (Tuina). (a)–(b) Binary distributions of negative (0) and positive (1) annotations for each assessment indicator across 61 subjects. Although some indicators exhibit moderate class imbalance (e.g., Twisting Frequency in Acupuncture and Slow Movement in Tuina), no task collapses to a single dominant class. (c) Distribution of composite skill scores (sum of binary labels per subject), showing moderate skew toward intermediate-to-high proficiency while preserving variability for multi-aspect evaluation and subgroup analysis.

A. Procedure

To the best of our knowledge, no publicly available multi-view video dataset exists for clinical TCM rehabilitation training. To address this gap, we introduce two intervention-specific datasets: TCM-AQA61-A (Acupuncture) and TCM-AQA61-T (Tuina). Each dataset contains recordings from 61 subjects, supporting robust model training and evaluation. Both datasets emphasize hand motions, hand–object interactions, and fine motor skills central to TCM rehabilitation, providing a foundation for vision-based skill assessment in traditional medicine.

The datasets provide synchronized dual-view recordings of hand movements (Fig. 1), including an egocentric (first-person) and an exocentric (third-person) perspective [46]. This configuration captures complementary motion cues from the participant's and observer's viewpoints. The egocentric view emphasizes in-plane hand motion and subtle manipulation dynamics, while the exocentric view better captures vertical displacement and depth progression. Compared to conventional fixed multi-camera configurations [45], this design preserves fine-grained hand–object interaction together with global spatial context. Such complementary perspectives facilitate cross-view representation learning in CME-AQA.

The data collection process was approved by the Ethical Committee of the Department of Computer Science at Durham University (Reference ID: COMP-2023-03-24T13_52_41-slxb76). It encompassed acupuncture and Tuina practice sessions on a simulated practice pad, conducted by 61 medical students from the Beijing University of Chinese Medicine, representing a structured teaching cohort for TCM rehabilitation skill assessment. Videos were recorded using two GoPro HERO8 cameras: one mounted on the subject's forehead to provide an egocentric perspective and another positioned to capture an exocentric view of the hands.

Assessment criteria were defined according to the Clinical Practice Guidelines of Traditional Chinese Medicine [47]. For TCM-AQA61-A, evaluated aspects included Needle Holding, Needle Angle, Needle Depth, Quick Needle Insertion, Lifting and Thrusting Frequency, Lifting and Thrusting Amplitude, Twisting Frequency, Twisting Amplitude, and Quick Needle Withdrawal. For TCM-AQA61-T, evaluated aspects included Sinking Shoulders, Dropping Elbows, Suspended Wrists, Hollow Palms, Solid Fingers, Elbow/Forearm Force, Depth, and Frequency. After data collection, two experienced TCM physical therapists independently assessed all subjects using synchronized dual-view recordings, followed by a consensus discussion to resolve discrepancies. Inter-rater reliability was computed on the first-round annotations prior to consensus using Cohen's κ and the two-way random-effects intraclass correlation coefficient (ICC(2,1)) across all indicators. The mean κ was 0.62, and the average ICC was 0.63, indicating moderate agreement [48]. These results demonstrate acceptable annotation consistency between experts and support the reliability of the dataset for subsequent model training and evaluation.

In addition to these classification-based assessments, four key aspects were manually annotated with continuous values to support metrological assessment: insertion time (from initial needle–surface contact until stabilization), withdrawal time (from initiation of withdrawal until the needle is fully removed), and manipulation frequency (per second) for both acupuncture and Tuina. These continuous annotations were designed to provide more detailed feedback and enhance the training utility of the system.

B. Dataset Statistics and Label Distribution

The label distributions of all binary assessment indicators are shown in Fig. 2. Each sub-skill includes 61 samples with varying proportions of negative (0) and positive (1) labels. For TCM-AQA61-A, positive samples range from 10 (Twisting Frequency) to 57 (Twisting Amplitude). Several aspects are relatively balanced (e.g., Needle Depth: 26 vs. 35; Quick

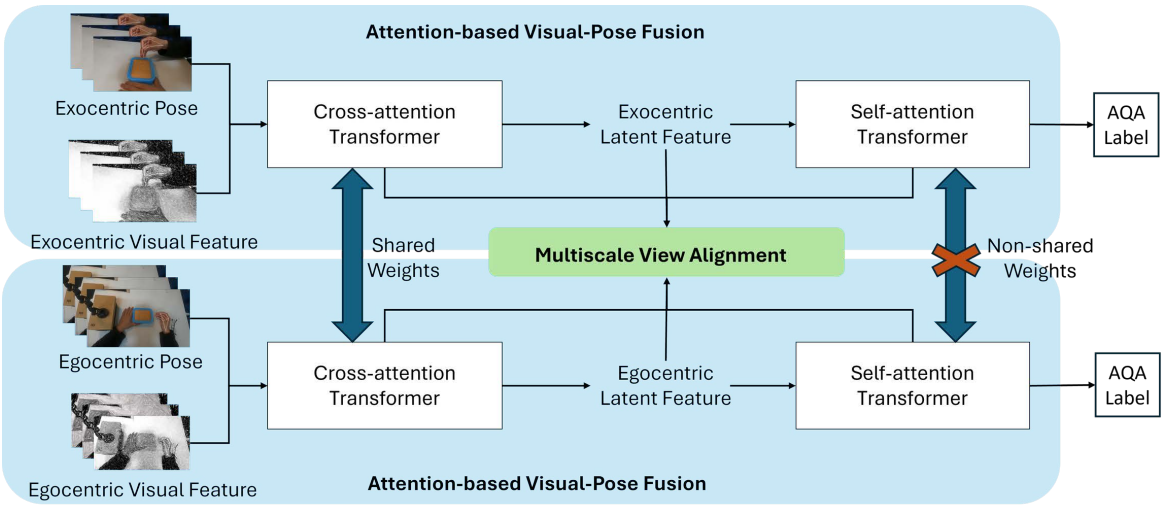

Fig. 3. Overview of the Cross-view Multimodality Enhanced Action Quality Assessment (CME-AQA) framework, which leverages multi-view and multimodal data for TCM rehabilitation assessment. It integrates two main components: (1) the Attention-based Visual–Pose Fusion (AVPF) module (Section IV-A), which fuses visual and pose features through cross- and self-attention, and (2) the Multiscale View Alignment (MVA) training strategy (Section IV-B), which aligns representations between egocentric and exocentric views to achieve multi-view awareness. During inference, only the exocentric (third-person) view is required.

Needle Insertion: 28 vs. 33), whereas others are moderately skewed due to structured training (e.g., Twisting Frequency: 51 vs. 10). For TCM-AQA61-T, positive counts range from 1 (Slow Movement) to 58 (Solid Fingers and Depth). Frequency remains nearly balanced (31 vs. 30), while Slow Movement is near-saturated (60 vs. 1), reflecting the consistency of this criterion among participants.

Fig. 2(c) further illustrates the distribution of composite skill scores (sum of binary labels per subject). For TCM-AQA61-A, total scores span 1 to 9, with most concentrated between 4 and 7, indicating moderate to high but non-saturated proficiency. For TCM-AQA61-T, scores range from 1 to 5, with most subjects scoring between 3 and 5. Although the distributions are moderately skewed toward intermediate to high performance, consistent with the structured training background of medical students, they retain meaningful variability across skill levels.

Overall, although some sub-skills show moderate imbalance, the dataset is not dominated by a single label, and composite scores retain sufficient spread for multi-aspect evaluation and subgroup analysis. To reduce potential bias, we report both accuracy and class-sensitive metrics such as weighted F1 score in subsequent experiments.

IV. Methodology

Fig. 3 illustrates CME-AQA. At inference, the model takes a single exocentric (third-person) video [10], [12] and predicts aspect-wise scores. We extract visual features $F_V \in \mathbb{R}^{T\times C}$ and hand pose features $F_P \in \mathbb{R}^{T\times C}$ per frame, where T denotes the number of frames and C the feature dimension, and fuse them with Attention-based Visual–Pose Fusion (AVPF) via pose-conditioned cross-attention followed by temporal self-attention. The cross-attention module refines visual representations using geometric cues from hand pose, supporting occlusion-aware refinement when visual cues are available, while temporal self-attention models long-range dependencies necessary for capturing fine-grained procedural dynamics. The resulting representation is then mapped to task outputs by a prediction head.

During training, Multiscale View Alignment (MVA) uses synchronized FPV and TPV pairs to align multi-scale latent features across views. Multi-scale alignment encourages view-invariant representations at different abstraction levels, enabling robust single-view inference at deployment.

A. Attention-based Visual–Pose Fusion (AVPF)Module

Conventional AQA methods for rehabilitation primarily rely on skeletal representations [10], [15], [12], which are insufficient for modeling fine-grained hand–object interaction in TCM procedures such as acupuncture [4]. Recent multimodal AQA approaches [18] integrate pose and visual features through pose-guided attention and global–local body modeling, where pose cues primarily bias spatial attention during visual feature extraction. Because pose influences representations only through attention modulation, its geometric information may not be strongly preserved across layers. For hand-centric procedures, where pose defines interaction geometry, a more explicit pose–vision coupling is therefore required. To this end, AVPF performs pose-conditioned refinement by using pose features as keys and values to update visual representations while keeping pose features fixed as a geometric reference.

As shown in Fig. 4, AVPF employs a multi-layer cross-attention transformer [49] to condition visual features on hand pose. Cross-attention enables structured feature-level interaction beyond simple concatenation [50]. In our formulation, pose features act as keys and values, while only visual features are updated at each layer. This asymmetric update allows iterative refinement of visual representations while preserving stable pose structure.

For implementation, visual features are extracted using a pretrained backbone (e.g., ResNet [51]) to obtain $F_V \in \mathbb{R}^{T\times C}$.

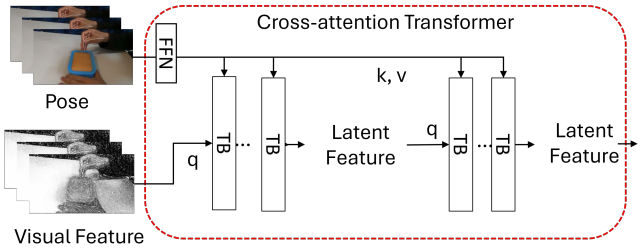

Fig. 4. Architecture of the cross-attention transformer in our AVPF module. Visual features ($\boldsymbol{Q}$) are refined with pose features ($\boldsymbol{K}, \boldsymbol{V}$), while pose features remain fixed. This design enables the model to attend to pose cues relevant to the visual context, supporting interaction-aware feature fusion for AQA.

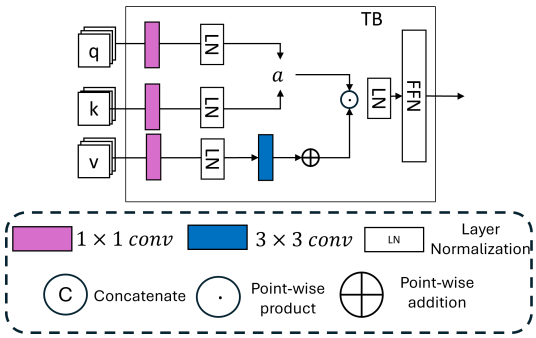

Fig. 5. The architecture of our layer-normalization-enhanced attention transformer block. To mitigate potential noise caused by pose estimation inaccuracies, an additional normalization layer is applied after the 1D convolution to enhance robustness.

Hand pose features are obtained using a single-view 3D pose estimator (MediaPipe Hands [52], [53]) and projected to the same dimensionality, producing $F_P \in \mathbb{R}^{T\times C}$. Each cross-attention block, indexed by i, updates F_V^i using pose features F_P through a layer-normalized attention formulation:

$$\text{Attn}^i(Q,K,V) = \text{softmax}\left(\frac{LN(Q)LN(K)^T}{\sqrt{d_k}}\right)LN(V), \tag{1}$$

where Q is derived from F_V^i, and K, V are derived from F_P via 1D convolutions. As illustrated in Fig. 5, an additional layer normalization is applied after the 1D convolution to enhance robustness against pose estimation noise. The attended features are then passed through an FFN to produce the updated F_V^{i+1}. Stacking I layers yields the fused representation F_V^I, while pose features remain unchanged throughout.

The fused representation F_V^I is further processed by a stack of layer-normalized temporal self-attention blocks (Fig. 6) to model intra-sequence dependencies:

$$\text{Attn}^j(Q',K',V') = \text{softmax}\left(\frac{LN(Q')LN(K')^T}{\sqrt{d_{k'}}}\right)LN(V'), \tag{2}$$

where Q', K', V' are derived from F_V^j via 1D convolutions. Stacking J layers produces the refined representation F_V^J.

Finally, F_V^J is passed through a two-layer fully connected network to predict aspect-wise action quality scores.

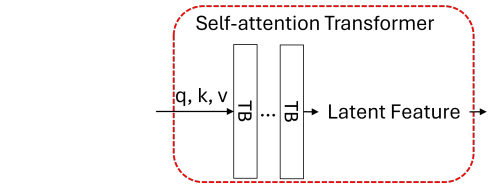

Fig. 6. Architecture of the self-attention transformer in our AVPF module. Fused features act as $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$, allowing the model to refine representations by focusing on the most relevant information for AQA inference.

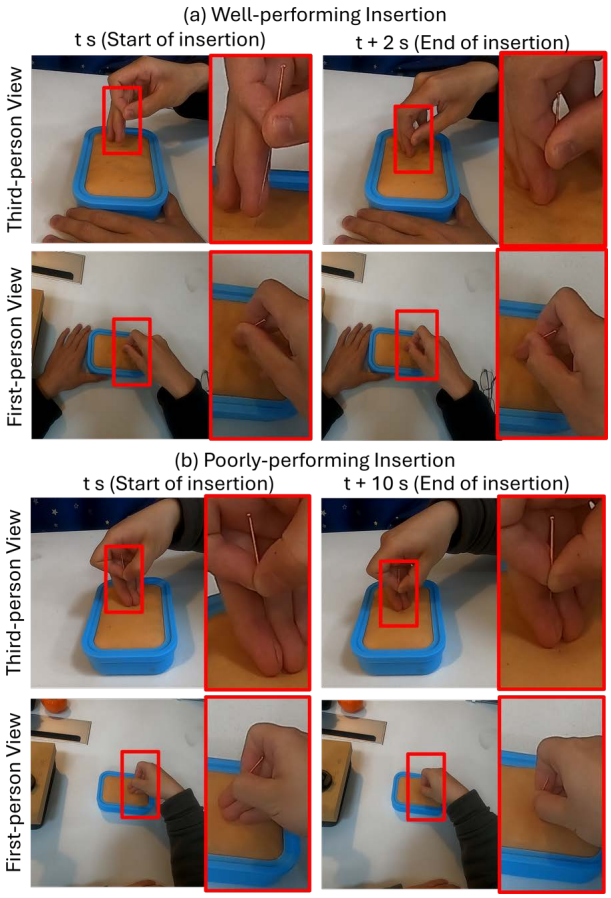

Fig. 7. Illustration of complementary motion cues captured by third-person (TPV) and first-person (FPV) views in TCM-AQA61-A during needle insertion. (a) Well-performed trial where the needle is inserted stably and quickly; (b) poorly performed trial where the needle is inserted irregularly and slowly. In each case, the upper row shows TPV frames and the lower row shows FPV frames. TPV better exposes the vertical insertion trajectory relative to the practice pad, while FPV highlights in-plane adjustments of the needle. This cross-view complementarity motivates the proposed MVA strategy for aligning representations across views during training.

B. Multiscale View Alignment (MVA) Training Strategy

Conventional AQA methods for TCM rehabilitation rely on a single exocentric view during both training and inference [10], [15], [12]. However, hand-centric procedures often exhibit view-dependent motion cues [54]. As shown in Fig. 7, TPV captures vertical insertion and depth progression, whereas FPV highlights in-plane adjustments and subtle needle rotations. Such complementary yet view-specific information cannot be fully represented by a single viewpoint. While multi-view AQA can mitigate occlusion, existing approaches

typically treat synchronized views as independent samples or fuse them via feature aggregation [19], [20]. These strategies emphasize information accumulation rather than enforcing cross-view consistency and are not inherently designed for single-view deployment.

To address this limitation, MVA formulates multi-view learning as a cross-view representation alignment problem. During training, synchronized FPV–TPV pairs are processed through shared cross-attention modules, and latent representations are aligned at multiple hierarchical stages within AVPF.

Specifically, we define three alignment scales: (i) early cross-attention features, where visual and pose representations are initially fused; (ii) AVPF output features that capture integrated multimodal embeddings; and (iii) self-attention features that provide view-refined representations. At each scale, the corresponding features are projected through a two-layer fully connected network with a nonlinear activation function prior to alignment. The alignment loss is defined as:

$$L_{\text{Align}} = \sum_{m=1}^{n} \lambda_m \left\| F_{V,\text{exo}}^{m} - F_{V,\text{ego}}^{m} \right\|_1, \tag{3}$$

where $F_{V,\text{exo}}^{m}$ and $F_{V,\text{ego}}^{m}$ denote exocentric and egocentric latent features at the m-th scale, and λ_m balances contributions across scales. This regularization enforces cross-view consistency while preserving hierarchical feature structure.

During training, cross-attention weights are shared across views to capture view-invariant geometric correlations between visual and pose features, whereas self-attention weights remain view-specific to preserve task-adaptive refinement. The overall objective is defined as

$$L = \alpha L_{\text{Align}} + \beta L_{\text{AQA}}, \tag{4}$$

where L_{AQA} denotes the classification or regression loss for quality prediction, and α, β balance alignment and task supervision. This formulation enables view-invariant representation learning while retaining discriminative capacity for AQA.

V. Experiment

A. Experimental Design

We conducted experiments on our collected datasets, TCM-AQA61-A and TCM-AQA61-T, as well as on an external multi-view dataset of 40 CPR training practices proposed by Constable et al. [20]. The TCM-AQA61 datasets were used for cross-validation and model optimization, while the CPR dataset was used to assess generalization across different practices and view configurations.

1) Experiments on TCM-AQA61-A and TCM-AQA61-T: Our main experiments were conducted on the proposed TCM-AQA61-A and TCM-AQA61-T datasets. Each dataset was split into five folds for cross-validation. Consistent with recent automatic rehabilitation assessment studies [10], we report results as mean $\pm$ standard deviation across folds to reflect performance variability for each assessment indicator. ResNet [51] was employed as the backbone for visual feature extraction, and MediaPipe Hands [52], [53] was used for markerless skeletal extraction, as both are widely adopted in AQA and provide standard visual/skeleton features [55].

The network was configured with the following hyperparameters: $J = 2$, $I = 2$, $\alpha = 0.5$, and $\beta = 0.5$. We set a batch size of 2 and trained the model for 50 epochs using the Adam optimizer with a learning rate of 1×10^{-4}. All experiments were conducted on an NVIDIA GeForce RTX 4090 GPU.

For metrological assessments such as insertion time, withdrawal time, and manipulation frequency, we used Mean Absolute Error (MAE)to quantify the deviation between predicted values and expert annotations. Mean Absolute Error (MAE) is a standard metric for regression tasks and offers a straightforward measure of prediction accuracy in continuous-valued AQA [20]. Together with classification metrics, it enables a comprehensive evaluation of both categorical and continuous outcomes, supporting precise and actionable feedback in TCM rehabilitation training.

For classification assessments, each aspect was evaluated using accuracy and weighted F1 score. Accuracy reflects overall prediction correctness, while the weighted F1 score balances precision and recall under class imbalance, providing a more informative metric for clinical data [56]. These evaluation criteria are well-suited to clinical applications of AQA, where reliability and interpretability are essential for effective TCM rehabilitation assessment [22].

2) Experiments on CPR Dataset: To validate the generalization of our method across different practices and view configurations, we evaluated our framework on a multi-view dataset of 40 CPR training videos proposed by Constable et al. [20]. The evaluation of our CME-AQA framework followed the evaluation protocol of Constable et al. [20]. This dataset includes individual evaluations from two experts as well as the consensus evaluation of the CPR practice, covering aspects such as Hand Position, Arm Position, Shoulder Position, Depth, Rate, and Compression Release.

Our CME-AQA framework was retrained on this dataset. Since the ratings for each subject range from 0 to 4 in this dataset, the AQA task in this context is treated as a regression task. To align with this, we modified the final layer of our framework to output a regression prediction. Unlike our collected datasets, which include first-person and third-person views, the CPR dataset contains only front-view and side-view configurations, as egocentric views were not available. Because large portions were blurred for privacy, reliable visual features were unavailable; we therefore used skeleton data only. As a result, only skeletal data could be used as input in our CME-AQA framework. To facilitate a fair comparison with expert evaluation, data from both front and side views were utilized during training (as shown in Fig. 8), as they are the most commonly chosen views for evaluation in this dataset, while only the front view was used during inference.

The network hyperparameters were configured to match those used in our experiments on our collected datasets to verify the generalization ability. We set a batch size of 2 and trained the model for 50 epochs using the Adam optimizer with a learning rate of 5×10^{-5}. All experiments were conducted on an NVIDIA GeForce RTX 4090 GPU.

For training and testing, we used a fivefold cross-validation protocol, allocating 80% of the data for training and 20% for testing. Following Constable et al. [20], the Mean Absolute

Fig. 8. Example frames of the views used. For fair comparison, both the front and side views were included during training, as these were the main viewpoints used by experts, while only the front view was used for inference.

Error (MAE) metric was used to quantify the average error between the predicted scores and the expert-agreed scores.

B. Metrological Assessment Performance

TABLE II
METROLOGICAL ASSESSMENT PERFORMANCE COMPARISON ON THE TCM-AQA61 DATASET. RESULTS ARE REPORTED AS MAE (MEAN±SD OVER FOLDS). FOR EACH METRIC, THE METHOD WITH THE LOWEST MEAN IS HIGHLIGHTED IN BOLD; WHEN MEANS ARE IDENTICAL, THE METHOD WITH THE LOWER SD IS SELECTED.

Method	**Acupuncture**			**Tuina**
	Insertion Time (s) (↓)	Withdrawal Time (s) (↓)	Frequency (Hz) (↓)	Frequency (Hz) (↓)
STGCN [57]	3.51±0.91	1.05±0.22	0.56±0.14	0.75±0.14
STNN [10]	3.26±1.06	0.94±0.15	0.53±0.05	0.65±0.08
STGCN-LSTM [11]	3.55±0.79	0.94±0.17	0.51±0.05	0.65±0.05
STGCN-RI [12]	6.98±4.25	1.71±0.78	0.93±0.59	0.68±0.06
FineParser [25]	3.44±1.61	1.08±0.12	0.61±0.22	0.69±0.12
Uni-FineParser [26]	3.43±0.71	**0.94**±0.12	0.50±0.05	0.64±0.07
PGT [18]	3.45±0.63	0.96±0.11	0.50±0.06	0.66±0.08
PHI [27]	3.13±1.15	0.99±0.21	0.52±0.06	**0.63**±0.07
Ours	**3.12**±1.04	0.94±0.14	**0.49**±0.05	0.64±0.05

Baselines. We conducted a benchmarking study on our dataset to compare the proposed framework with representative baselines. Baselines 1–4 correspond to classical pose-based methods commonly used in rehabilitation-oriented action quality assessment. Specifically, (1) STGCN [57] is a spatial–temporal graph convolutional network for skeleton-based action modeling; (2) STNN [10] is an early LSTM-based framework that leverages pose sequences for rehabilitation exercise assessment; (3) STGCN-LSTM [11] extends STGCN with an LSTM module to capture longer temporal dependencies; and (4) STGCN-RI [12] incorporates joint rotation matrices into an STGCN backbone to enhance motion representation for rehabilitation skill evaluation. All four baselines were retrained on our dataset under a unified evaluation protocol.

Baselines 5–8 represent more recent transformer-based methods originally developed for general action quality assessment. These include (5) FineParser [25], which uses transformer-based spatio-temporal parsing to align human-centric action steps between query and exemplar videos; (6) Uni-FineParser [26], a unified variant that introduces mask-guided spatial attention and a temporal transformer with learnable queries for fine-grained contrastive score regression; (7) PGT [18], a pose-guided transformer framework that integrates pose heatmaps into spatial attention and models global–local body dynamics for fine-grained AQA and (8) PHI [27], which applies long-term transformer modeling for absolute score regression from single-view RGB videos with progressive domain adaptation. For FineParser and Uni-FineParser, we approximate the foreground using hand poses and replace the original classification head with a two-layer fully connected regression head to enable metrological prediction, as the original implementations are classification-oriented.

Results. As shown in Table II, our method achieves the lowest MAE for acupuncture insertion time and frequency, and matches the best-performing baselines for withdrawal time. Compared to the second-best baselines, our framework reduces the MAE for acupuncture insertion time from 3.26 s to 3.12 s (4.3% improvement) and for acupuncture frequency from 0.51 Hz to 0.49 Hz (3.9%). Notably, these improvements are achieved over recent transformer-based AQA frameworks such as PGT, PHI, and Uni-FineParser, which model long-term temporal dynamics but rely on single-view global representations. For withdrawal time, our method matches the best-performing baseline with an MAE of 0.94 s, while exhibiting a comparable standard deviation (± 0.14) to other strong methods. For Tuina frequency, PHI achieves the lowest MAE (0.63 Hz), while our method attains 0.64 Hz with reduced standard deviation (± 0.05 vs. ± 0.07), indicating improved stability.

Table II further shows that these performance gains are accompanied by stable prediction behavior. In particular, our method exhibits lower standard deviation for Tuina frequency (± 0.05) compared to PHI (± 0.07), and comparable standard deviation for withdrawal time, indicating that the observed improvements are not driven by outliers. The lower mean errors than most recent methods suggest that explicitly modeling localized hand-centric interactions and structured multi-view cues could contribute to favorable metrological performance in TCM rehabilitation. Although the absolute prediction errors remain non-trivial due to the inherent variability and complexity of fine-grained hand motions in TCM procedures, our approach still achieves the best or comparable mean performance across the evaluated metrological indicators, demonstrating its potential to provide more precise and reliable metrological assessments in clinical training applications.

C. Classification-Based Assessment Performance

Baselines. We employed the same baseline methods as in the metrological assessment experiments. However, PHI [27] and PGT [18] were originally formulated for continuous score regression. To enable fair comparison in the classification-based evaluation, we replaced their regression heads with categorical prediction layers while keeping the backbone architectures unchanged. All other network configurations remained consistent with their original implementations.

Results. As practical metrological assessments may not fully reflect practice quality, we also conducted classification-based evaluations. The results presented in Tables III, IV and V demonstrate that our proposed CME-AQA framework achieves competitive or superior performance relative to the strongest competing baselines across both acupuncture and Tuina tasks. Importantly, these comparisons include recent transformer-based AQA frameworks such as Uni-FineParser, PGT, and PHI, which employ global spatio-temporal modeling but do

TABLE III

PERFORMANCE COMPARISON OF ACUPUNCTURE SKILLS ON THE TCM-AQA61-A DATASET (PART I: NEEDLE HANDLING AND INSERTION). RESULTS ARE REPORTED AS ACCURACY/WEIGHTED F1 (%, MEAN±SD OVER FOLDS). FOR EACH METRIC, THE METHOD WITH THE HIGHEST MEAN IS HIGHLIGHTED IN BOLD; WHEN MEANS ARE IDENTICAL, THE METHOD WITH THE LOWER SD IS SELECTED.

	Needle Holding (↑)	Needle Angle (↑)	Needle Depth (↑)	Quick Needle Insertion (↑)
STGCN [57]	0.90±0.09/**0.85**±0.08	0.62±0.11/0.55±0.11	0.57±0.13/0.41±0.15	0.57±0.18/0.51±0.16
STNN [10]	0.90±0.08/0.85±0.12	**0.68**±0.05/0.56±0.07	0.57±0.10/0.41±0.16	0.54±0.17/0.37±0.17
STGCN-LSTM [11]	0.90±0.08/0.85±0.12	0.68±0.08/0.56±0.11	0.57±0.07/0.41±0.07	0.42±0.10/0.42±0.10
STGCN-RI [12]	0.86±0.13/0.83±0.10	0.59±0.05/0.54±0.06	0.52±0.08/0.52±0.09	0.49±0.17/0.47±0.16
FineParser [25]	0.88±0.08/0.85±0.13	0.63±0.09/0.57±0.14	0.52±0.06/0.46±0.07	0.55±0.07/0.51±0.07
Uni-FineParser [26]	0.90±0.10/0.85±0.14	0.68±0.14/0.56±0.18	0.59±0.07/0.48±0.11	0.50±0.10/0.47±0.16
PGT [18]	0.90±0.09/0.85±0.13	0.66±0.10/0.60±0.15	0.47±0.08/0.39±0.14	0.52±0.08/0.48±0.11
PHI [27]	**0.90**±0.06/0.85±0.09	0.68±0.10/0.56±0.12	0.55±0.14/0.42±0.16	0.53±0.15/0.44±0.20
Ours	0.88±0.08/0.85±0.13	0.63±0.12/**0.61**±0.14	**0.60**±0.08/**0.60**±0.08	**0.63**±0.12/**0.63**±0.11

TABLE IV

PERFORMANCE COMPARISON OF ACUPUNCTURE SKILLS ON THE TCM-AQA61-A DATASET (PART II: MANIPULATION AND WITHDRAWAL). RESULTS ARE REPORTED AS ACCURACY/WEIGHTED F1 (%, MEAN±SD OVER FOLDS). FOR EACH METRIC, THE METHOD WITH THE HIGHEST MEAN IS HIGHLIGHTED IN BOLD; WHEN MEANS ARE IDENTICAL, THE METHOD WITH THE LOWER SD IS SELECTED.

	Lifting & Thrusting Frequency (↑)	Lifting & Thrusting Amplitude (↑)	Twisting Frequency (↑)	Twisting Amplitude (↑)	Quick Needle Withdrawal (↑)
STGCN [57]	0.68±0.10/**0.65**±0.07	0.70±0.19/0.63±0.22	0.83±0.08/0.76±0.11	0.93±0.11/0.90±0.10	0.85±0.08/0.80±0.11
STNN [10]	0.75±0.11/0.64±0.14	0.65±0.18/0.51±0.24	**0.83**±0.05/0.76±0.07	**0.93**±0.06/**0.90**±0.09	0.86±0.10/0.80±0.14
STGCN-LSTM [11]	**0.75**±0.10/0.64±0.11	0.65±0.14/0.51±0.17	**0.83**±0.05/0.76±0.07	**0.93**±0.06/**0.90**±0.09	0.86±0.10/0.80±0.14
STGCN-RI [12]	0.62±0.13/0.60±0.12	0.49±0.20/0.47±0.18	0.77±0.04/0.72±0.04	0.93±0.10/0.90±0.11	**0.88**±0.08/**0.84**±0.12
FineParser [25]	0.70±0.13/0.64±0.16	0.59±0.21/0.58±0.19	0.82±0.12/0.76±0.18	**0.93**±0.06/**0.90**±0.09	0.78±0.13/0.78±0.16
Uni-FineParser [26]	0.75±0.11/0.65±0.14	0.75±0.11/0.73±0.13	0.83±0.14/0.77±0.19	**0.93**±0.06/**0.90**±0.09	0.85±0.10/0.80±0.15
PGT [18]	**0.75**±0.10/0.65±0.14	0.62±0.09/0.63±0.08	0.83±0.13/0.77±0.19	**0.93**±0.06/**0.90**±0.09	0.80±0.08/0.78±0.14
PHI [27]	0.75±0.18/0.65±0.23	**0.77**±0.09/**0.74**±0.10	0.82±0.06/0.75±0.09	**0.93**±0.06/**0.90**±0.09	0.86±0.11/0.81±0.16
Ours	0.65±0.09/0.60±0.08	0.75±0.08/**0.74**±0.10	0.83±0.12/**0.78**±0.17	**0.93**±0.06/**0.90**±0.09	0.85±0.11/0.80±0.16

TABLE V

PERFORMANCE COMPARISON OF TUINA SKILLS ON THE TCM-AQA61-T DATASET. RESULTS ARE REPORTED AS ACCURACY/WEIGHTED F1 (%, MEAN±SD OVER FOLDS). FOR EACH METRIC, THE METHOD WITH THE HIGHEST MEAN IS HIGHLIGHTED IN BOLD; WHEN MEANS ARE IDENTICAL, THE METHOD WITH THE LOWER SD IS SELECTED.

	Standard Action (↑)	Hollow Palm (↑)	Solid Fingers (↑)	Slow Movement (↑)	Depth (↑)	Frequency (↑)
STGCN [57]	0.68±0.05/0.66±0.06	0.66±0.14/0.64±0.15	0.94±0.06/0.92±0.07	**0.98**±0.03/**0.97**±0.05	0.82±0.05/0.85±0.05	0.47±0.25/0.47±0.26
STNN [10]	0.61±0.07/0.46±0.12	0.71±0.05/0.60±0.05	0.94±0.04/0.92±0.06	**0.98**±0.03/**0.97**±0.05	0.94±0.04/0.92±0.06	0.42±0.11/0.36±0.15
STGCN-LSTM [11]	0.50±0.14/0.46±0.20	0.71±0.13/0.60±0.18	0.94±0.04/0.92±0.06	**0.98**±0.03/**0.97**±0.05	0.94±0.04/**0.97**±0.06	0.43±0.10/0.43±0.14
STGCN-RI [12]	0.56±0.18/0.56±0.21	0.68±0.20/0.60±0.21	0.80±0.06/0.84±0.06	0.98±0.04/**0.97**±0.05	0.84±0.04/0.86±0.05	0.49±0.06/0.45±0.06
FineParser [25]	0.66±0.09/0.64±0.10	0.66±0.12/0.64±0.13	0.94±0.04/0.92±0.06	0.98±0.04/**0.97**±0.05	0.94±0.07/0.92±0.10	0.59±0.08/0.58±0.06
Uni-FineParser [26]	0.59±0.10/0.58±0.11	0.64±0.12/0.65±0.09	0.91±0.09/0.90±0.08	0.96±0.04/0.96±0.05	0.91±0.08/0.90±0.09	**0.63**±0.04/**0.62**±0.04
PGT [18]	0.66±0.07/0.66±0.07	0.70±0.18/0.67±0.21	0.94±0.04/0.92±0.06	**0.98**±0.03/**0.97**±0.05	0.93±0.08/0.91±0.10	0.61±0.10/0.56±0.10
PHI [27]	0.60±0.13/0.51±0.19	0.72±0.10/0.61±0.13	**0.95**±0.07/0.92±0.10	**0.98**±0.03/**0.97**±0.05	**0.95**±0.04/0.92±0.06	0.62±0.11/0.55±0.11
Ours	**0.71**±0.15/**0.70**±0.15	**0.73**±0.14/**0.69**±0.18	0.94±0.02/**0.92**±0.05	**0.98**±0.03/**0.97**±0.05	0.94±0.01/0.92±0.10	0.59±0.05/0.59±0.04

not explicitly incorporate structured multi-view alignment or localized hand-centric fusion.

For acupuncture, our method delivers improvements in several key skills. For example, in Needle Depth, our framework achieves an F1 score of 0.60, representing a relative gain of approximately 15% compared with the second-best method (0.52). This improvement is particularly notable given that transformer-based methods such as PGT and PHI achieve lower F1 scores (0.39 and 0.42, respectively), suggesting that global temporal modeling alone is insufficient for fine-grained depth discrimination. In Quick Needle Insertion, we obtain $0.63/0.63$ in accuracy/F1 score, which is about 6% higher in accuracy and 12% higher in F1 score than the second-best baseline ($0.57/0.51$). This margin is maintained over both graph-based and transformer-based competitors, indicating stronger modeling of rapid hand-object interactions. For Twisting Frequency, our model matches the highest accuracy (0.83) while slightly improving the F1 score (0.78 vs. 0.76). These gains are accompanied by moderate fold-level variability. In Needle Depth, our method achieves 0.60 ± 0.08 F1, exhibiting substantially lower standard deviation than PHI (0.42 ± 0.16) and comparable stability to the strongest graph-based baselines. Similarly, for Quick Needle Insertion, our F1 standard deviation (± 0.11) is lower than that of STGCN (± 0.16), suggesting that the observed mean-level advantage is less likely to be driven by a single outlier split.

In Tuina assessment, our method also outperforms not only classical pose-based baselines but also advanced transformer models on key indicators such as Standard Action and Hollow Palm. For example, compared with PHI (0.51 F1) and Uni-FineParser (0.58 F1) in Standard Action, our method achieves 0.70 ± 0.15 F1, representing a substantial relative improvement. Similarly, in Hollow Palm, our framework achieves 0.69 ± 0.18 F1, exceeding PHI (0.61) and PGT (0.67). This performance is supported by the fusion of visual and skeletal features and the integration of egocentric-view cues, which provide complementary information for fine-grained motion modeling and occlusion mitigation when visual hand-object

cues are available in the TCM setting. These findings suggest that explicitly modeling localized hand-centric interactions and cross-view complementary information is associated with favorable mean performance over purely global transformer modeling strategies on several fine-grained TCM rehabilitation indicators. These results further highlight the potential of our CME-AQA framework to provide convenient yet accurate feedback for structured TCM rehabilitation training.

We further examine whether the observed improvements could be attributed to label imbalance, as most participants are medical students and several indicators exhibit moderate skew (Fig. 2). Notably, the performance gains of CME-AQA are not confined to near-saturated indicators. For relatively balanced skills such as Needle Depth (26/35), Quick Needle Insertion (28/33), and Tuina Frequency (31/30), our method achieves superior or comparable mean performance relative to the strongest baselines (Tables III-V). Meanwhile, for highly imbalanced indicators such as Twisting Amplitude (4/57) and Slow Movement (60/1), our method achieves performance comparable to or slightly better than competing methods rather than exhibiting artificially inflated accuracy. These results suggest that the observed improvements are not confined to label-skewed indicators and could be associated with improved modeling of fine-grained interactions.

TABLE VI
OVERALL PAIRED COMPARISON BETWEEN CME-AQA (OURS) AND PHI ACROSS FIVE CROSS-VALIDATION FOLDS ON TCM-AQA61-A (ACUPUNCTURE) AND TCM-AQA61-T (TUINA). REPORTED VALUES ARE FOLD MEANS OVER ALL CLASSES (MEAN±SD; HIGHER IS BETTER ↑); p-VALUES ARE FROM PAIRED FOLD-LEVEL TESTS (n=5).

Dataset	Metric	n	PHI	Ours	p
TCM-AQA61-A	Accuracy	5	**0.75**±0.03	**0.75**±0.03	0.73
	Weighted F1	5	0.68±0.05	**0.72**±0.04	0.35
TCM-AQA61-T	Accuracy	5	0.80±0.04	**0.81**±0.04	0.19
	Weighted F1	5	0.74±0.05	**0.79**±0.04	0.08

D. Overall Fold-Level Statistical Comparison

Table VI presents an overall paired comparison between CME-AQA (Ours) and PHI [27], which achieves the strongest or second-best performance among the competing baselines in the classification-based evaluation. For each cross-validation fold, Accuracy and F1 scores are averaged across all classes to obtain a single fold-level metric, and results are reported as mean ± standard deviation over folds. Statistical significance is assessed using a paired two-sided t-test at the fold level to evaluate performance differences across splits. Given the limited number of folds ($n = 5$), the resulting p-values are treated as exploratory.

Overall, CME-AQA shows higher descriptive mean F1 scores on both acupuncture and Tuina datasets, with a larger mean difference observed in Tuina assessment. The overall fold-level pattern is in line with the per-indicator AQA results reported in Tables III, IV, and V. Given that statistical significance is not reached in the paired fold-level tests, these results are best interpreted as descriptive trends rather than statistically confirmed improvements, motivating further validation with larger cohorts.

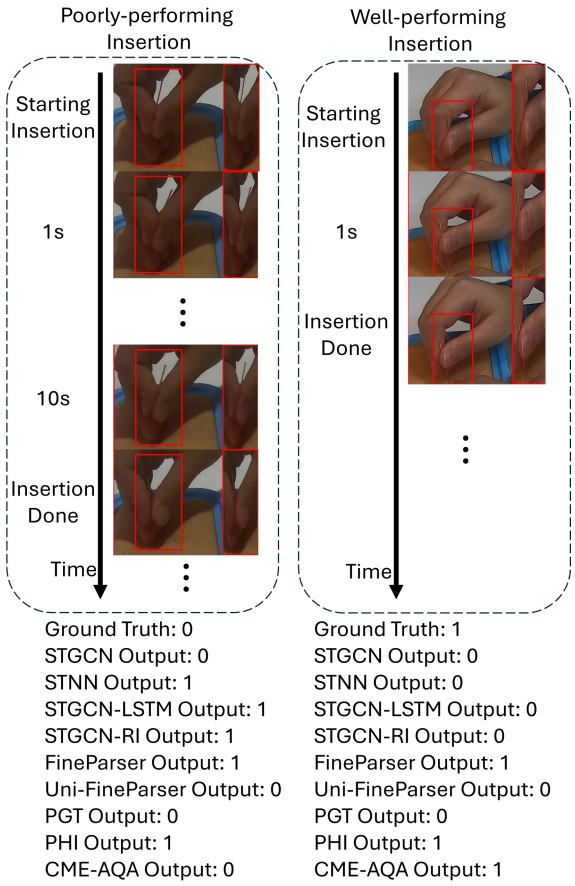

Fig. 9. Critical case comparison for Quick Needle Insertion. The red box highlights the hand–needle interaction. The well-performing trial completes insertion within 3s, whereas the poorly performing trial lasts about 10s with slower motion. CME-AQA correctly differentiates the two cases, while strong baselines misclassify at least one instance. See supplementary videos for full sequences.

E. Qualitative Analysis

Critical Case Comparison. To further examine the effectiveness of CME-AQA under challenging conditions, we present a critical case comparison of Quick Needle Insertion in Fig. 9. This aspect requires rapid and precise hand–needle coordination to achieve therapeutic effects while minimizing discomfort [58]. The well-performing trial completes insertion within ∼3s, whereas the poorly performing trial lasts ∼10s with slower execution. CME-AQA correctly distinguishes the two cases, while strong baselines (FineParser, Uni-FineParser, and PHI) misclassify at least one instance.

Temporal Attention Visualization. Figure 10 visualizes the learned temporal attention patterns for the same trials. The cross-attention weights are aggregated across query positions to yield a global temporal importance curve. In the well-performing case, attention exhibits a sharp peak tightly aligned with the annotated insertion window, indicating focused modeling of the brief hand–needle interaction. In contrast, the poorly performing case shows a broader, double-peaked pattern spanning the prolonged insertion period, reflecting extended and less decisive motion. This structured but temporally diffused response suggests that CME-AQA captures key events

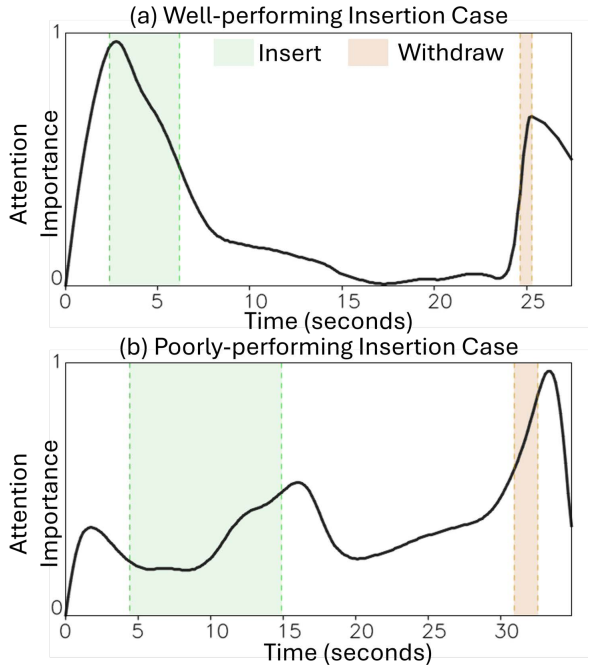

Fig. 10. Temporal attention patterns learned by CME-AQA for the same well-performing (a) and poorly performing (b) trials. Shaded regions denote insertion (green) and withdrawal (brown) phases. In the well-performing case, attention is sharply concentrated within the brief insertion window, whereas in the poorly performing case it is temporally dispersed across the prolonged insertion period. This behavior indicates that the model captures time-sensitive hand–needle interaction dynamics rather than relying on coarse temporal aggregation.

while remaining sensitive to variations in execution dynamics rather than relying on coarse temporal cues.

F. Ablation Study

To validate the effectiveness of the proposed components, we conducted a comprehensive ablation study across (i) view/modality settings, (ii) module design, and (iii) weight-sharing strategies; see Table VII.

View and Modality Ablation. We removed each view (egocentric/FPV or exocentric/TPV) and each modality (visual or pose) to assess their contributions. Integrating both views and both modalities yields the strongest results. Removing FPV leads to noticeable drops in accuracy and F1 for fine hand-movement aspects (e.g., needle depth), underscoring the role of multiple viewpoints in mitigating self-occlusion. Omitting pose features degrades tasks that require precise spatial reasoning (e.g., Needle Depth). These findings indicate that visual and pose cues are complementary.

Framework Design Ablation. Replacing the cross-attention in AVPF with fully connected layers reduces performance, especially for Needle Angle. This indicates that cross-modal attention is necessary to fuse pose and visual information effectively. Removing MVA and aligning only at the final output also degrades performance, with the largest drop on temporally structured aspects such as Twisting Frequency. These results support the benefit of multi-scale view alignment within the feature hierarchy.

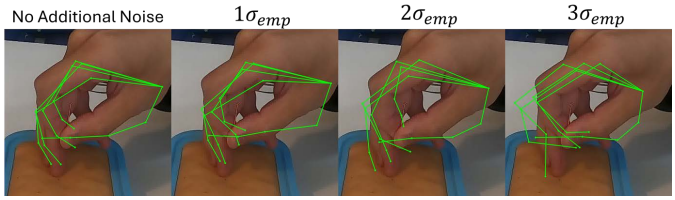

Fig. 11. Illustration of progressive pose perturbation. From left to right: original pose (no additional noise), and injected Gaussian jitter with $\boldsymbol{\sigma} = \mathbf{1}\boldsymbol{\sigma}_{\text{emp}}, \mathbf{2}\boldsymbol{\sigma}_{\text{emp}}, \mathbf{3}\boldsymbol{\sigma}_{\text{emp}}$. Noise is added independently to each joint coordinate in normalized space. As $\boldsymbol{\sigma}$ increases, joint locations exhibit progressively larger spatial deviations, leading to increasing geometric distortion of finger alignment and hand contour while preserving overall pose structure at lower perturbation levels.

Weight-Sharing Strategies. We compared three strategies: (1) share weights in *self-attention only*, (2) *no* weight sharing, and (3) share weights in *both* cross- and self-attention. Our default design—sharing weights in cross-attention while using view-specific self-attention—achieves the best overall accuracy and F1. This suggests cross-attention benefits from shared parameters to capture view-invariant correlations, whereas self-attention benefits from view-specific adaptation to refine task-relevant temporal features.

G. Sensitivity Analysis

To better understand the mechanisms of AVPF and MVA, we conduct a sensitivity analysis examining two factors: (i) pose quality and (ii) the scale of multi-view feature alignment.

Pose noise. Using the fully trained model, we inject zero-mean Gaussian noise into pose coordinates to simulate landmark jitter, a common data augmentation strategy [59], defining $p' = p + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$. The empirical median frame-to-frame pose displacement across all sequences is $\sigma_{\text{emp}} = 0.01$ in normalized coordinates. Relative to the median hand size in our dataset, this corresponds to $\sim$7% of hand width (4% of hand height). As illustrated in Fig. 11, we evaluate $\sigma \in \{1, 2, 3\}\sigma_{\text{emp}}$ to simulate different noise levels.

As shown in Table VIII, performance remains relatively stable at $1\sigma_{\text{emp}}$, while larger perturbations lead to gradual declines in both metrics. This relative stability under the tested perturbations could be related to the design of AVPF, where pose features serve as fixed geometric references in cross-attention and are not iteratively updated, thereby reducing the risk of noise amplification across layers. This behavior is further illustrated in Fig. 11, where stronger perturbations progressively distort joint configurations, weakening the spatial priors used for visual refinement. The smooth degradation suggests that the CME-AQA framework remains relatively stable within the simulated lower-to-moderate pose perturbation range tested here, while larger or systematic pose estimation errors may still propagate to the final AQA output.

Multi-view alignment scale. We evaluate different alignment configurations within MVA by retraining each variant while keeping all other components fixed. The three scales correspond to early cross-attention features, AVPF outputs, and self-attention output features (Sec. IV-B). As shown in Table VIII, performance generally improves as more scales are aligned,

TABLE VII
ABLATION STUDY ON THE TCM-AQA61-A DATASET (MEASURED BY ACCURACY/F1 SCORE). HIGHER VALUES INDICATE BETTER PERFORMANCE.

		Needle Holding	Needle Angle	Needle Depth	Quick Needle Insertion	Lifting & Thrusting Frequency	Lifting & Thrusting Amplitude	Twisting Frequency	Twisting Amplitude	Quick Needle Withdrawal
Modality & View	Ours	**0.88/0.85**	**0.63/0.61**	**0.60/0.60**	0.63/0.63	0.65/0.60	**0.75/0.74**	0.83/0.78	**0.93/0.90**	**0.85/0.80**
	(w/o TPV)	0.85/0.82	0.62/0.52	0.59/0.58	0.59/0.58	0.68/0.61	0.68/0.68	0.80/0.76	0.91/0.89	**0.85**/0.79
	(w/o FPV)	0.85/0.82	0.55/0.55	0.55/0.54	**0.68**/0.63	0.73/0.60	0.73/0.73	**0.85/0.81**	**0.93/0.90**	0.83/0.79
	(w/o Visual)	0.86/0.83	0.59/0.58	**0.60/0.60**	0.65/**0.65**	0.67/0.62	0.73/0.72	**0.85**/0.79	**0.93/0.90**	0.83/0.79
	(w/o Pose)	0.85/0.82	0.59/0.54	0.55/0.54	0.55/0.55	**0.75/0.64**	0.63/0.62	0.83/0.76	**0.93/0.90**	0.83/0.79
Design	Ours	**0.88/0.85**	0.63/0.61	**0.60/0.60**	**0.63/0.63**	0.65/0.60	**0.75/0.74**	**0.83/0.78**	**0.93/0.90**	**0.85/0.80**
	(w/o AVPF)	**0.88/0.85**	**0.65**/0.61	0.54/0.53	0.49/0.48	**0.72/0.63**	0.72/0.72	0.81/0.75	**0.93/0.90**	0.80/0.77
	(w/o MVA)	**0.88/0.85**	**0.65/0.63**	0.44/0.43	0.60/0.60	0.65/**0.63**	0.73/0.73	**0.83/0.78**	**0.93/0.90**	0.83/0.79
Share Weight	Ours	**0.88/0.85**	**0.63/0.61**	**0.60/0.60**	**0.63/0.63**	0.65/**0.60**	0.75/0.74	**0.83/0.78**	**0.93/0.90**	**0.85/0.80**
	Late	**0.88/0.85**	0.62/0.60	0.59/0.58	0.57/0.56	0.63/**0.60**	**0.78/0.77**	**0.83/0.78**	0.93/0.90	0.83/0.79
	No share	0.86/0.83	**0.63**/0.60	0.57/0.54	0.52/0.51	0.63/**0.60**	0.75/0.75	0.81/0.77	0.93/0.90	0.83/0.79
	All share	0.86/0.83	0.62/0.56	0.59/0.58	0.54/0.53	**0.67/0.60**	0.77/0.74	**0.83**/0.76	0.93/0.90	0.83/0.79

TABLE VIII
SENSITIVITY ANALYSIS OF OVERALL PERFORMANCE ON TCM-AQA61-A (ACUPUNCTURE). REPORTED VALUES ARE FOLD MEANS OVER ALL ASSESSMENT INDICATORS (MEAN±SD; HIGHER IS BETTER ↑ ; n=5). POSE NOISE DENOTES ADDITIVE GAUSSIAN JITTER ON JOINT COORDINATES, AND MULTI-VIEW SCALE REFERS TO THE NUMBER OF ALIGNED FEATURE SCALES IN MVA.

Setting	Accuracy	Weighted F1
Pose noise (Gaussian jitter), $\epsilon \sim \mathcal{N}(0, \sigma^2)$		
No additional noise	**0.75**±0.03	**0.72**±0.04
$1\sigma_{\text{emp}}$	0.74±0.05	0.70±0.05
$2\sigma_{\text{emp}}$	0.74±0.05	0.69±0.05
$3\sigma_{\text{emp}}$	0.72±0.05	0.67±0.06
Multi-view alignment scale (number of aligned feature scales in MVA)		
1-scale (early only)	0.74±0.03	0.65±0.03
1-scale (mid only)	0.74±0.03	0.64±0.04
1-scale (late only)	0.73±0.03	0.70±0.04
2-scale (early+mid)	0.74±0.03	0.65±0.04
2-scale (early+late)	0.74±0.03	0.67±0.05
2-scale (mid+late)	**0.75**±0.03	0.67±0.05
3-scale (all scales, default MVA)	**0.75**±0.03	**0.72**±0.04

and the full configuration achieves the best results. Notably, late-only alignment, performed on self-attention features, already provides competitive performance. This trend reflects the rationale of MVA, indicating that high-level semantic consistency in late features is the primary driver of multi-view robustness. Early and mid-level alignment operate on the shared multimodal backbone features and mainly provide geometric regularization. Aligning all scales combines geometric stabilization with semantic alignment, resulting in more robust view-invariant embeddings.

TABLE IX
MEAN ABSOLUTE ERROR (HUMAN EXPERTS VS. OUR CME-AQAFRAMEWORK). LOWER VALUES INDICATE BETTER PERFORMANCE.

Aspect	Expert 1	Expert 2	MV-STGCN [20]	Ours
Hand Position	1.62	1.08	**0.33**	**0.33**
Arm Position	0.70	0.15	**0.07**	**0.07**
Shoulder Position	0.40	0.34	**0.13**	**0.13**
Depth	0.49	**0.30**	0.69	0.69
Rate	0.89	**0.11**	1.67	1.78
Compression Release	1.04	**0.98**	1.00	1.00

H. Generalization Analysis for CPR Skill Assessment

The generalization results are reported in Table IX. Expert 1 and Expert 2 denote the mean absolute error (MAE) of each expert relative to the consensus scores. Compared with the multi-view baseline MV-STGCN [20], our single-view framework achieves comparable errors across several aspects: it matches the baseline on Hand Position (0.33), Arm Position (0.07), Shoulder Position (0.13), and Compression Release (1.00); it is identical on Depth (0.69) and slightly higher on Rate (1.78 vs. 1.67). Relative to the human experts, our errors are substantially lower on Hand Position, Arm Position, and Shoulder Position. Our error is comparable to that of Expert 2 on Compression Release (1.00 vs. 0.98). Errors are higher than those of Expert 2 on Depth and Rate. Notably, although our framework uses only single-view input at inference, it achieves performance comparable to MV-STGCN, which relies on multi-view data at inference.

The higher errors in Rate relative to MV-STGCN [20] and the higher errors in Depth and Compression Release compared with human experts therefore expose boundary conditions of the proposed method under restricted input modalities. In the CPR dataset, inference was conducted using pose information only, as privacy blurring removed reliable RGB cues, thereby limiting the visual input for AVPF and the interaction cues required for cross-view alignment in MVA. Therefore, the reduced interaction modeling in CPR reflects a missing-modality boundary condition, rather than negating the intended role of AVPF in TCM procedures where visual hand-object cues are available. This restriction prevents the framework from fully exploiting its multi-view interaction modeling capability under single-view input, which partly explains why our Rate is lower than that of MV-STGCN, which directly leverages multi-view observations as input. Moreover, the key distinction between CPR and the TCM task lies not in motion frequency, but in the type of interaction. CPR compression depth and recoil depend on the deformation of an external compliant object (i.e., the manikin chest) [60], rather than on hand pose alone. Such deformable contact dynamics are not directly observable from skeletal sequences, which partly explains why our errors are higher than those of human experts when the relevant interaction dynamics cannot be inferred from pose alone. These findings indicate that the CME-AQA framework shows preliminary transferability to pose-dominant criteria, but its interaction understanding and view alignment capabilities are inherently constrained when critical interaction cues are absent.

VI. Discussion and Conclusion

In this paper, we presented the Cross-view Multimodality Enhanced Action Quality Assessment (CME-AQA) framework, a vision-based approach that integrates pose–visual feature fusion and cross-view learning for TCM rehabilitation AQA. To our knowledge, this is the first framework to combine these elements for fine-grained assessment in traditional medicine training, providing a practical tool to support skill development for learners in structured TCM rehabilitation training. We also introduced two multi-view datasets, TCM-AQA61-A (Acupuncture) and TCM-AQA61-T (Tuina), which include recordings from 61 subjects performing representative TCM rehabilitation procedures. Quantitative experiments on our datasets demonstrated that our approach achieved over 10% higher F1 scores in key tasks such as Needle Depth and Quick Needle Insertion. Evaluation on the CPR dataset [20] further showed performance comparable to human experts on several posture-based criteria, suggesting applicability to related structured simulated clinical skill assessments where participant motion is central to evaluation.

Several limitations of CME-AQA remain. One important consideration is the dependency of our framework on the accuracy of initial pose estimation from raw video data. Although the sensitivity analysis in Table VIII shows relative stability within the simulated lower-to-moderate perturbation range, this finding does not remove the framework's dependency on pose-estimation quality; larger or systematic pose errors could still propagate to the final AQA output. Future work could integrate more robust pose estimation techniques, such as temporal-based pose estimation [61]. This integration would help pose estimation align more closely with the demands of our AQA tasks, particularly in ensuring consistent pose estimation for complex hand movements in TCM therapy scenarios.

Another challenge is how objects are semantically represented in the model. Although our CME-AQA framework introduces visual features to provide a more comprehensive understanding of the environment, this may distract the focus from motion itself. The AVPF module is designed to strengthen interaction-aware fusion when visual hand-object cues are available, while settings with missing or unreliable RGB input may require additional sensing or explicit object representations. Recent advancements in interaction detection, such as the use of bounding box detection [62], segmentation methods [63], and dynamic graph representations [64] to highlight objects during video analysis, could help address this issue. These methods could emphasize object interactions and provide a more robust representation in our future work.

Finally, although the current dataset provides a valuable benchmark for fine-grained AQA, it reflects a structured teaching cohort in which most participants are medical students. Consequently, certain advanced or rare skill indicators exhibit label imbalance, while intermediate proficiency dominates overall score distributions. Although our analysis shows that the favorable mean-level trends are not confined to near-saturated indicators, the present validation should be interpreted within structured TCM rehabilitation training rather than as evidence of generalization across the full spectrum of expertise levels. In addition, the overall fold-level comparisons should be regarded as descriptive rather than statistically confirmed improvements. Future work will expand data collection to include more experienced practitioners and a wider range of error patterns, enabling more comprehensive evaluation across diverse skill levels and stronger statistical validation with larger cohorts.

Acknowledgement

This research is supported in part by Capital's Funds for Health Improvement and Research (ref: 2024-4-21211), Training Plan for High Level Public Health Technical Talents Construction Project (ref: TL-02-40), National Natural Science Foundation of China (ref: 8240152532), and the EPSRC NortHFutures project (ref: EP/X031012/1).

Reference

[1] W.-W. Tao, H. Jiang, X.-M. Tao, P. Jiang, L.-Y. Sha, and X.-C. Sun, "Effects of acupuncture, tuina, tai chi, qigong, and traditional chinese medicine five-element music therapy on symptom management and quality of life for cancer patients: a meta-analysis," *Journal of pain and symptom management*, vol. 51, no. 4, pp. 728–747, 2016.

[2] X. Yu, B. Gong, H. Yang, Z. Wang, G. Qi, J. Sun, Y. Fang, and X. Fan, "Effect of acupuncture treatment on cortical activation in patients with tinnitus: a functional near-infrared spectroscopy study," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 729–737, 2023.

[3] A. de Sá Ferreira, "Evidence-based practice of chinese medicine in physical rehabilitation science," *Chinese journal of integrative medicine*, vol. 19, no. 10, pp. 723–729, 2013.

[4] G. Du, Y. Li, K. Su, C. Li, and P. X. Liu, "A mobile natural human–robot interaction method for virtual chinese acupuncture," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–10, 2022.

[5] Q. Sun, J. Huang, H. Zhang, P. Craig, L. Yu, and E. G. Lim, "Design and development of a mixed reality acupuncture training system," in *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2023, pp. 265–275.

[6] S. Rahman, S. Sarker, A. N. Haque, M. M. Uttsha, M. F. Islam, and S. Deb, "Ai-driven stroke rehabilitation systems and assessment: a systematic review," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 192–207, 2022.

[7] J. Liu, H. Wang, K. Stawarz, S. Li, Y. Fu, and H. Liu, "Vision-based human action quality assessment: A systematic review," *Expert Systems with Applications*, p. 125642, 2024.

[8] F. Wu, Q. Wang, J. Bian, N. Ding, F. Lu, J. Cheng, D. Dou, and H. Xiong, "A survey on video action recognition in sports: Datasets, methods and applications," *IEEE Transactions on Multimedia*, 2022.

[9] X. Zhang, F. Liang, C. T. Lau, J. C. Chan, N. Wang, J. Deng, J. Wang, Y. Ma, L. L. Zhong, C. Zhao *et al.*, "Standards for reporting interventions in clinical trials of tuina/massage (strictotm): Extending the consort statement," *Journal of Evidence-Based Medicine*, vol. 16, no. 1, pp. 68–81, 2023.

[10] Y. Liao, A. Vakanski, and M. Xian, "A deep learning framework for assessing physical rehabilitation exercises," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 2, pp. 468–477, 2020.

[11] S. Deb, M. F. Islam, S. Rahman, and S. Rahman, "Graph convolutional networks for assessment of physical rehabilitation exercises," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 410–419, 2022.

[12] K. Zheng, J. Wu, J. Zhang, and C. Guo, "A skeleton-based rehabilitation exercise assessment system with rotation invariance," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023.

[13] G. Karvounas, N. Kyriazis, I. Oikonomidis, and A. Argyros, "Dynamic multiview refinement of 3d hand datasets using differentiable ray tracing," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3156–3166.

[14] J. E. Jang, Y. S. Lee, W. S. Jang, W. S. Sung, E.-J. Kim, S. D. Lee, K. H. Kim, and C. Y. Jung, "Trends in acupuncture training research: focus on practical phantom models," 2022.

[15] L. Yao, Q. Lei, H. Zhang, J. Du, and S. Gao, "A contrastive learning network for performance metric and assessment of physical rehabilitation exercises," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023.

[16] J. Xu, Y. Rao, X. Yu, G. Chen, J. Zhou, and J. Lu, "Finediving: A fine-grained dataset for procedure-aware action quality assessment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 2949–2958.

[17] S. Zhang, W. Dai, S. Wang, X. Shen, J. Lu, J. Zhou, and Y. Tang, "Logo: A long-form video dataset for group action quality assessment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 2405–2414.

[18] Y. Zhang, X. Li, W. Chai, C. Yan, W. Wang, and G. Wang, "Pose-guided transformer for fine-grained action quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[19] L. Dong, W. Wang, Y. Qiao, and X. Sun, "Lucidaction: A hierarchical and multi-model dataset for comprehensive action quality assessment," *Advances in neural information processing systems*, vol. 37, pp. 96 468–96 482, 2024.

[20] M. D. Constable, F. X. Zhang, T. Conner, D. Monk, J. Rajsic, C. Ford, L. J. Park, A. Platt, D. Porteous, L. Grierson *et al.*, "Advancing healthcare practice and education via data sharing: demonstrating the utility of open data by training an artificial intelligence model to assess cardiopulmonary resuscitation skills," *Advances in Health Sciences Education*, pp. 1–21, 2024.

[21] X. Zhang, "Clinical video analysis with geometric feature enhanced deep learning," Doctoral Thesis, Durham University, Durham, UK, 2025, available under Creative Commons Attribution 3.0 (CC BY). [Online]. Available: https://etheses.dur.ac.uk/16017/

[22] S. Wang, D. Yang, P. Zhai, Q. Yu, T. Suo, Z. Sun, K. Li, and L. Zhang, "A survey of video-based action quality assessment," in *2021 International conference on networking systems of AI (INSAI)*. IEEE, 2021, pp. 1–9.

[23] H. Pirsiavash, C. Vondrick, and A. Torralba, "Assessing the quality of actions," in *European conference on computer vision*. Springer, 2014, pp. 556–571.

[24] P. Parmar and B. T. Morris, "What and how well you performed? a multitask learning approach to action quality assessment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 304–313.

[25] J. Xu, S. Yin, G. Zhao, Z. Wang, and Y. Peng, "Fineparser: A fine-grained spatio-temporal action parser for human-centric action quality assessment," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2024, pp. 14 628–14 637.

[26] J. Xu, S. Yin, and Y. Peng, "Human-centric fine-grained action quality assessment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[27] K. Zhou, H. P. Shum, F. W. Li, X. Zhang, and X. Liang, "Phi: Bridging domain shift in long-term action quality assessment via progressive hierarchical instruction," *IEEE Transactions on Image Processing*, 2025.

[28] M. Fieraru, M. Zanfir, S. C. Pirlea, V. Olaru, and C. Sminchisescu, "Aifit: Automatic 3d human-interpretable feedback models for fitness training," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9919–9928.

[29] S. Zahan, G. M. Hassan, and A. Mian, "Learning sparse temporal video mapping for action quality assessment in floor gymnastics," *IEEE Transactions on Instrumentation and Measurement*, 2024.

[30] Z. Wang, J. Li, J. Wang, H. Zhao, S. Qiu, N. Yang, and X. Shi, "Inertial sensor-based analysis of equestrian sports between beginner and professional riders under different horse gaits," *IEEE Transactions on Instrumentation and Measurement*, vol. 67, no. 11, pp. 2692–2704, 2018.

[31] A. Alamri, M. Eid, R. Iglesias, S. Shirmohammadi, and A. El Saddik, "Haptic virtual rehabilitation exercises for poststroke diagnosis," *IEEE transactions on instrumentation and measurement*, vol. 57, no. 9, pp. 1876–1884, 2008.

[32] M. Borghetti, E. Sardini, and M. Serpelloni, "Sensorized glove for measuring hand finger flexion for rehabilitation purposes," *IEEE Transactions on Instrumentation and Measurement*, vol. 62, no. 12, pp. 3308–3314, 2013.

[33] J. Li, S. Mo, and Y. Shen, "Taichi action capture and performance analysis with multi-view rgb cameras," *arXiv preprint arXiv:2306.14490*, 2023.

[34] X. Zhao, "Ai-driven tai chi mastery using deep learning framework for movement assessment and personalized training," *Scientific Reports*, vol. 15, no. 1, p. 31700, 2025.

[35] M. Baldinger, K. Lippmann, G. Lisca, and V. Senner, "Development of a motion capture and feedback system for qigong," *Sports Engineering*, vol. 28, no. 1, p. 23, 2025.

[36] A. Al-Bedah, G. Ali, T. Abushanab, and N. Qureshi, "Tui na (or tuina) massage: a minireview of pertinent literature, 1970-2017," *Journal of Complementary and Alternative Medical Research*, vol. 3, no. 1, pp. 1–14, 2017.

[37] W. Xu, D. Xiang, G. Wang, R. Liao, M. Shao, and K. Li, "Multiview video-based 3-d pose estimation of patients in computer-assisted rehabilitation environment (caren)," *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 2, pp. 196–206, 2022.

[38] B. G. Gerats, J. M. Wolterink, and I. A. Broeders, "3d human pose estimation in multi-view operating room videos using differentiable camera projections," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 11, no. 4, pp. 1197–1205, 2023.

[39] A. Schmidt, A. Sharghi, H. Haugerud, D. Oh, and O. Mohareri, "Multi-view surgical video action detection via mixed global view attention," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24*. Springer, 2021, pp. 626–635.

[40] A. E. Abdelaal, A. Avinash, M. Kalia, G. D. Hager, and S. E. Salcudean, "A multi-camera, multi-view system for training and skill assessment for robot-assisted surgery," *International journal of computer assisted radiology and surgery*, vol. 15, no. 8, pp. 1369–1377, 2020.

[41] A. Vakanski, H.-p. Jun, D. Paul, and R. Baker, "A data set of human body movements for physical rehabilitation exercises," *Data*, vol. 3, no. 1, p. 2, 2018.

[42] M. Capecci, M. G. Ceravolo, F. Ferracuti, S. Iarlori, A. Monteriu, L. Romeo, and F. Verdini, "The kimore dataset: Kinematic assessment of movement and clinical scores for remote monitoring of physical rehabilitation," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 7, pp. 1436–1448, 2019.

[43] A. Miron, N. Sadawi, W. Ismail, H. Hussain, and C. Grosan, "Intellirehabds (irds)—a dataset of physical rehabilitation movements," *Data*, vol. 6, no. 5, p. 46, 2021.

[44] M. Devanne, O. R. Neris, M. Lempereur, A. Thepaut *et al.*, "A medical low-back pain physical rehabilitation database for human body movement analysis," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–8.

[45] J. Li, J. Xue, R. Cao, X. Du, S. Mo, K. Ran, and Z. Zhang, "Finerehab: A multi-modality and multi-task dataset for rehabilitation analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3184–3193.

[46] S. Ardeshir and A. Borji, "An exocentric look at egocentric actions and vice versa," *Computer Vision and Image Understanding*, vol. 171, pp. 61–68, 2018.

[47] C. A. of Chinese Medicine, "Evaluation specification of clinical practice guidelines of traditional chinese medicine," China Association of Chinese Medicine, Tech. Rep., January 2021, draft version.

[48] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *biometrics*, pp. 159–174, 1977.

[49] C.-F. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 357–366.

[50] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, "Early vs late fusion in multimodal convolutional neural networks," in *2020 IEEE 23rd international conference on information fusion (FUSION)*. IEEE, 2020, pp. 1–6.

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[52] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee *et al.*, "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.

[53] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, "Mediapipe hands: On-device real-time hand tracking," *arXiv preprint arXiv:2006.10214*, 2020.

[54] S. Han, P.-c. Wu, Y. Zhang, B. Liu, L. Zhang, Z. Wang, W. Si, P. Zhang, Y. Cai, T. Hodan *et al.*, "Umetrack: Unified multi-view end-to-end hand tracking for vr," in *SIGGRAPH Asia 2022 conference papers*, 2022, pp. 1–9.

[55] K. Zhou, R. Cai, L. Wang, H. P. Shum, and X. Liang, "A comprehensive survey of action quality assessment: Method and benchmark," *arXiv preprint arXiv:2412.11149*, 2024.
[56] R. J. Chen, T. Ding, M. Y. Lu, D. F. Williamson, G. Jaume, A. H. Song, B. Chen, A. Zhang, D. Shao, M. Shaban *et al.*, "Towards a general-purpose foundation model for computational pathology," *Nature medicine*, vol. 30, no. 3, pp. 850–862, 2024.
[57] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.
[58] C. S. Yin, J.-H. Kim, and H.-J. Park, "High-velocity insertion of acupuncture needle is related to lower level of pain," *The Journal of Alternative and Complementary Medicine*, vol. 17, no. 1, pp. 27–32, 2011.
[59] Y. Liu, R. Liu, Y. Hu, M. Wu, W. Xin, Q. Miao, S. Wu, and L. Li, "A systematic review of skeleton-based action recognition: Methods, challenges, and future directions," *IEEE Transactions on Neural Networks and Learning Systems*, 2025.
[60] V. Krasteva, I. Jekova, and J.-P. Didon, "An audiovisual feedback device for compression depth, rate and complete chest recoil can improve the cpr performance of lay persons during self-training on a manikin," *Physiological measurement*, vol. 32, no. 6, pp. 687–699, 2011.
[61] Y. Wen, H. Pan, L. Yang, J. Pan, T. Komura, and W. Wang, "Hierarchical temporal transformer for 3d hand pose estimation and action recognition from egocentric rgb videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 243–21 253.
[62] T. Qiao, Q. Men, F. W. Li, Y. Kubotani, S. Morishima, and H. P. Shum, "Geometric features informed multi-person human-object interaction recognition in videos," in *European Conference on Computer Vision*. Springer, 2022, pp. 474–491.
[63] M. Zhu, E. S. Ho, S. Chen, L. Yang, and H. P. Shum, "Geometric features enhanced human-object interaction detection," *IEEE Transactions on Instrumentation and Measurement*, 2024.
[64] F. X. Zhang, J. Deng, R. Lieck, and H. P. Shum, "Adaptive graph learning from spatial information for surgical workflow anticipation," *IEEE Transactions on Medical Robotics and Bionics*, 2024.