TOO MANY TO JUDGE? MULTIDANCE: A MULTI-TASK BENCHMARK FOR MULTI-TARGET ACTION QUALITY ASSESSMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Action Quality Assessment (AQA) aims to quantify how well actions are performed, serving as a fine-grained video understanding task with broad applications. However, most studies evaluate a single athlete or team in isolation, overlooking competitions where multiple performers share the stage and interact. Such settings require target-specific assessment that accounts for competitive context and provides interpretable feedback. To address these issues, we introduce **Multi-Target AQA (MT-AQA)**, a task that jointly evaluates and compares multiple competing targets within the same video. We present **MultiDance**, the first large-scale benchmark for MT-AQA, comprising 1,548 long videos from 99 events with 14,498 annotated performer instances of diverse skill levels and annotations for classification, ranking, scoring, and coaching. We further propose **Multi-Target Assessment Coach (MultiAC)**, a multimodal large language model (MLLM) framework that unifies these tasks through domain knowledge prompts and coarse-to-fine dual-stage adaptation. MultiAC outperforms single-target AQA extensions and general MLLMs on MultiDance, where joint training reveals complementary benefits between single- and multi-target data. Results on EgoExo-Fitness further support coarse-to-fine multi-task adaptation beyond dance.

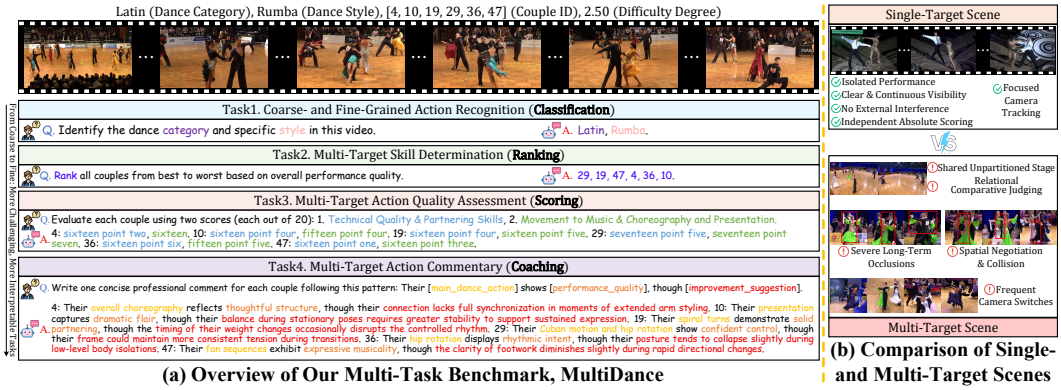
1 INTRODUCTION

Action Quality Assessment (AQA) automatically evaluates the quality of human actions in videos, providing numerical or semantic performance assessments. It has drawn growing attention due to broad applications in competitive sports (Xu et al., 2025c;a), healthcare (Ding et al., 2023; Yu et al., 2024), and skill determination (Luo et al., 2025; Xu et al., 2025b).

Existing AQA research has achieved promising progress in various single-target scenarios, such as diving (Xu et al., 2022b), figure skating (Xia et al., 2023), and gymnastics (Zeng et al., 2020). A few studies have also explored group performance assessment, like artistic swimming (Zhang et al., 2023) or group dancing (Xu et al., 2025b). However, these studies mostly follow an isolated evaluation paradigm, assessing one targeted individual or team at a time without external interference. In real-world competitions such as ballroom dance preliminaries and surfing heats, multiple contestants concurrently share a single, unpartitioned stage. As illustrated in Fig. 1(b), performers continuously negotiate space, avoid collisions, and adjust trajectories under frequent long-term occlusions and camera panning. These tangible spatial interactions inherently influence the final judges' scores. Consequently, simply "tracking, cropping, and scoring" each target is often brittle and overlooks crucial interactive context. This gap motivates **Multi-Target Action Quality Assessment (MT-AQA)**, which necessitates instance-aware evaluation, holistic perception, and relational comparison of multiple co-located targets within their shared competitive context.

Beyond multi-target perception, conventional scoring models (Xu et al., 2022b; Zhang et al., 2023) typically yield scalar scores from isolated features. However, in multi-target competitions, judges inherently perform *comparative judging*, where perceived quality depends on relative technical execution and presentation under the same stage conditions. Moreover, athletes remain uncertain about *why* they are ranked higher or lower. A promising direction is commentary-style evaluation (Li et al., 2026; Okamoto & Parmar, 2024; Zhang et al., 2024), which explains the reasoning. Thus,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Overview of Our Multi-Task Benchmark, MultiDance

(b) Comparison of Single- and Multi-Target Scenes

Figure 1: (a) Overview of our multi-task benchmark, MultiDance, a multi-target competitive dance video dataset with detailed annotations for numerous performers. It provides potential for analyzing coarse-to-fine cognitive processes in complex real-world scenarios, enabling the exploration of interpretable multi-target action quality assessment. (b) Comparing single-target and multi-target scenarios. Multi-target action assessment presents numerous unique real-world challenges.

our central motivation emerges: *How can a unified model simultaneously identify multiple performers, compare them to produce rankings, predict instance-specific scores, and generate interpretable commentary while retaining their shared competitive context?*

To address this challenge, we introduce **MultiDance**, the first multi-task benchmark for MT-AQA. It comprises videos from 99 events, averaging 107.5 seconds, with each multi-target video featuring at least six couples (i.e., 12+ dancers) performing concurrently. As illustrated in Fig. 1(a), each multi-target sample is annotated for four tasks: action classification, ranking, scoring, and coaching, forming a coarse-to-fine evaluation hierarchy. We also include a supplementary single-couple dance dataset with corresponding annotations (classification, scoring, coaching) to study the interplay between single- and multi-target AQA. MultiDance provides RGB, audio, optical flow, and text for multimodal AQA. Its shared competition scenes, official scores, and target-specific coaching support context-aware, interpretable assessment beyond isolated performances.

Directly extending single-target AQA models yields a mean reported correlation of 0.256 (Sec. 5.2). Explicit tracking improves target-score association but remains vulnerable to occlusions, producing conflicting scores and inconsistent rankings. This motivates a key insight: target-specific assessment should distinguish individual performances while preserving their shared competitive context.

Guided by this insight, we propose **MultiAC** (**Multi**-Target **A**ssessment **C**oach), a multimodal large language model (MLLM) for assessment from full uncropped videos. MultiAC unifies the four tasks through instruction following instead of separate task heads (Lai et al., 2024; Zhang et al., 2024; Li et al., 2026). The shared system prompt encodes domain knowledge and comparative reasoning, with task-specific questions specifying the outputs. Coarse-to-fine dual-stage adaptation connects scene-level context and target associations with fine-grained scoring and coaching.

Extensive experiments on MultiDance show gains from unified assessment and joint single- and multi-target training. Evaluation on public EgoExo-Fitness (Li et al., 2024b) and a qualitative surfing case further examine adaptation beyond dance. Our main contributions are summarized as follows:

- This paper defines **Multi-Target Action Quality Assessment (MT-AQA)** for context-aware perception, evaluation, and comparison of competing performers within a shared video, capturing the target-specific and relational demands of real-world competitions.
- We present **MultiDance**, the first large-scale, multi-task, and multimodal benchmark for MT-AQA, with concurrent competing targets, diverse skill levels, official scores, and target-specific coaching annotations.
- We propose **MultiAC**, an MLLM-based framework that unifies classification, ranking, scoring, and coaching for coherent and explainable multi-target assessment through domain knowledge prompts and coarse-to-fine dual-stage adaptation. Experiments validate its effectiveness and the complementary benefits of single- and multi-target training.

2 Related Work

Action Quality Assessment. Action Quality Assessment (AQA) (Zhou et al., 2026a) has attracted significant attention for objective evaluation in sports (Xu et al., 2022b; 2024; 2025d;c;a), healthcare (Liu et al., 2021; Ding et al., 2023; Yu et al., 2024), and skill determination (Luo et al., 2025; Xu et al., 2025b). Most methods formulate AQA as score regression, evolving from holistic evaluation (Xu et al., 2020; Zeng et al., 2020; Xia et al., 2023; Parmar & Morris, 2019) to fine-grained sub-action understanding (Xu et al., 2022b; 2024; 2025d; Zhang et al., 2023; Ke et al., 2024) and grade pattern modeling (Xu et al., 2022a; Zhou et al., 2024a; Liu et al., 2025; Zhou et al., 2025; Xu et al., 2025a), leading to continuous performance gains. More recently, multimodal AQA paradigms have emerged, incorporating audio (Xia et al., 2023; Zeng & Zheng, 2024; Xu et al., 2025a; 2026b), optical flow (Zeng & Zheng, 2024), and language (Xu et al., 2024; 2025a;c; 2026a) to further improve accuracy. However, these methods typically output only a single score, offering limited insight into why certain scores are given or how to improve, hindering model interpretability. To address this, some works (Li et al., 2026; Okamoto & Parmar, 2024; Zhang et al., 2024) explore text-based feedback for more interpretable assessments. Nevertheless, existing methods focus on a single athlete or team, overlooking the common scenario of multiple performers competing simultaneously. In this work, we introduce a new task, Multi-Target AQA (MT-AQA), and a corresponding multi-task benchmark, MultiDance. Our goal is to foster a comprehensive understanding of multiple performers and provide coarse-to-fine, interpretable assessments and feedback.

Group Activity Understanding. Understanding group activities requires interpreting collective actions from multiple individuals and their interactions, making it a challenging task with practical value. Related video understanding methods explore multi-label action modeling (Zhu et al., 2024; Kung et al., 2024; Zhao et al., 2023) and hierarchical relations for group activity recognition (Ibrahim & Mori, 2018; Yan et al., 2023; Hu et al., 2020; Han et al., 2022). These approaches focus on recognition or detection without assessing performance quality. Recent works (Zhang et al., 2023; Xu et al., 2025b) suggest extending beyond high-level activity recognition to assess participant collaboration quality and fine-grained actions. Nevertheless, these methods still focus on analyzing a single target group and cannot handle more complex scenarios involving multiple interacting groups. Our work takes the first step toward multi-granularity understanding of multi-target group actions by establishing a benchmark for complex, real-world group scenarios.

3 The MultiDance Dataset

3.1 Dataset Construction

Collection. We source videos from the official YouTube channel of the World DanceSport Federation (WDSF). We select 99 top-tier events held from 2018 to 2025, all evaluated by more than ten WDSF-accredited international adjudicators. After manual verification, we trim each video to capture the dancers' complete performance, from their entrance to their exit. Detailed collection criteria and preprocessing are provided in Appendix C.2.

Annotation. MultiDance is divided into MT (multi-target) and ST (single-target) subsets. To facilitate coarse-to-fine, explainable AQA, the MT subset is annotated for four tasks: action classification, ranking, scoring, and coaching. Each couple forms one assessment target. The ST subset contains one couple per video and omits ranking. Each sample includes the action type, difficulty level, and three official scores: Technical Quality & Partnering Skills (TQ & PS, 0-20), Movement to Music & Choreography and Presentation (MM & CP, 0-20), and a total score (0-40), all sourced and verified from official results. MultiDance offers four modalities: video, audio, optical flow, and text commentary. For commentary, Qwen3-VL-235B (Bai et al., 2025a) generates initial drafts, followed by manual correction, model refinement, and verification of each sample by two of five certified ballroom dance experts. The entire dataset construction, annotation, and verification process took approximately 600 hours. Appendix C provides the dataset structure and annotation workflow.

3.2 Dataset Statistics

The MultiDance dataset contains 1,548 video samples from 99 events, averaging 107.5s per video (46.21h total), spanning 2 skill levels, 3 age groups, 3 score dimensions, 4 modalities, 4 annotation

Table 1: Comparison of MultiDance with existing datasets for long-term AQA or AQA with commentary (top) and dance performance (bottom). *MT*: Multi-Target. *Action*: Temporal boundaries of sub-actions. *Act.Label*: Multi-level action types. *Act.Mask*: Human-centric action masks. *Bbox*: Bounding boxes of actors. *Neatness*: Group action neatness. *Year*: Video source date.

Dataset	Publisher	MT	Group	Level	Sample	Performer	Event	Type	Length	Annotation	Domain	Year
Fis-V	TCSVT'20	✗	✗	✗	500	500	12	1	23.89h	Score	Sport	2010-2018
RG	ACMMM'20	✗	✗	✗	1,000	1,000	4	4	26.39h	Score	Sport	2018-2019
LOGO	CVPR'23	✗	✓	✗	200	1,600	26	12	11.34h	Score,Action,Formation	Sport	2018-2022
FS1000	AAAI'23	✗	✗	✗	1,247	1,733+	27	8	91.00h	Score	Sport	2018-2021
MTL-NAE	CVPR'24	✗	✗	✗	1,412	1,412	16	2	1.70h	Score,Commentary	Sport	2012-2016
EgoExo4D (Sports)	IJCV'25	✗	✗	✗	1,219	289	-	3	16.40h	Score,Commentary	Sport	2022-2023
MTL-AQA-HM	TPAMI'25	✗	✗	✗	1,412	1,412	16	2	1.70h	Score,Act.Mask,Commentary	Sport	2012-2016
MM-JDM	IJCV'26	✗	✗	✗	1,639	1,639	-	12	∼10h	Grade,Act.Label	Healthcare	2022-2024
CoffeeCraft	CVPRF'26	✗	✗	✓	610	610	11	4	32.20h	Rank,Score	Skill	2022-2025
AIST++	ICCV'21	✗	✗	✗	1,408	30	-	10	5.19h	N/A	Dance	2019
DanceNet	TOMM'22	✗	✗	✗	2	2	-	2	0.96h	N/A	Dance	2022
AIST-M	ACMMM'23	✗	✓	✗	340	1,020+	-	10	1.02h	N/A	Dance	2019
CDRG	IJAIN'23	✗	✗	✗	240	20	-	12	0.98h	Act.Label,Rank	Dance	2023
ImperialDance	AAAI'24	✗	✗	✓	100	5	-	5	19.25h	Score,Act.Label	Dance	2022-2023
EgoExo4D (Dance)	IJCV'25	✗	✗	✗	728	93	-	1	31.55h	Score,Commentary	Dance	2022-2023
Dancer Parts	AAAI'25	✗	✓	✗	65	260+	-	6	0.13h	Bbox,Part Tracking	Dance	2023-2024
DNV	AAAI'25	✗	✓	✗	50	150+	-	2	0.05h	Neatness	Dance	2023-2024
MultiDance (Ours)	This Paper	✓	✓	✓	1,548	14,498	99	10	46.21h	Score,Act.Label,Rank,Commentary	Sport&Dance	2018-2025

types, and 10 dance styles. Both subsets are split into training and testing sets at a 4:1 ratio by events, with no event overlap between splits. The action types and score distributions are shown in Fig. 7 in Appendix C.1. Table 1 compares MultiDance with existing AQA datasets (Xu et al., 2020; Zeng et al., 2020; Zhang et al., 2023; Xia et al., 2023; Zhang et al., 2024; Grauman et al., 2025; Xu et al., 2025d; Zhou et al., 2026b; Yu et al., 2026) and dance datasets (Li et al., 2021; Zhuang et al., 2022; Yao et al., 2023; Hipiny et al., 2023; Zhong & Demiris, 2024; Xu et al., 2025b). MultiDance combines multi-target competition scenes with multi-task annotations based on standardized scoring protocols. Retaining concurrent performers and their shared competitive context enables target-specific assessment and within-video comparison. Its complementary MT and ST subsets further support systematic study of shared assessment knowledge across target configurations.

Unique Realistic Challenges. Performers dynamically alter their trajectories to secure the judge's visual attention or to dodge collisions. These explicit interactions result in complex inter-target occlusions and truncated visibility. Therefore, evaluating a single target in MultiDance requires spatial-temporal context from the entire visible group. Appendix B provides further discussion.

4 Approach

In this section, we formalize the new **M**ulti-**T**arget **A**ction **Q**uality **A**ssessment (**MT-AQA**) task and introduce our **Multi**-Target **A**ssessment **C**oach (MultiAC), a multimodal large language model (MLLM) that progressively understands and evaluates multi-target actions, as illustrated in Fig. 2.

4.1 Problem Formulation

Most existing AQA methods target isolated single-target evaluation. They can be formulated as $\hat{s} = \mathcal{R}(\mathcal{F}(X))$, where X is the input video, $\mathcal{F}$ extracts deep features, and $\mathcal{R}$ regresses a score. For scenes with N performers, an intuitive approach is to build a multiple single-target extension:

$$\hat{s}_i = \mathcal{R}\big(\mathcal{F}(X_{\text{crop},i})\big), \;\; i = 1, \ldots, N, \tag{1}$$

where $X_{\text{crop},i}$ is an instance-specific crop or mask obtained by detection/tracking.

However, this independent formulation is fundamentally brittle for MT-AQA. *First*, in dense competitive scenes, producing continuous instance-specific tracks ($X_{\text{crop},i}$) is fragile under long-term occlusions and camera panning. *Second*, performance assessment is often context-dependent: judges compare targets under the same stage conditions, and cues such as spatial negotiation and presentation emerge only in the shared context. Discarding the full video context X removes critical comparative signals. Therefore, we define **Multi-Target AQA (MT-AQA)** as context-aware assessment and comparison over the holistic input:

$$\hat{s}_i = \mathcal{S}(X, \mathcal{I}_i), \;\; i = 1, \ldots, N, \tag{2}$$

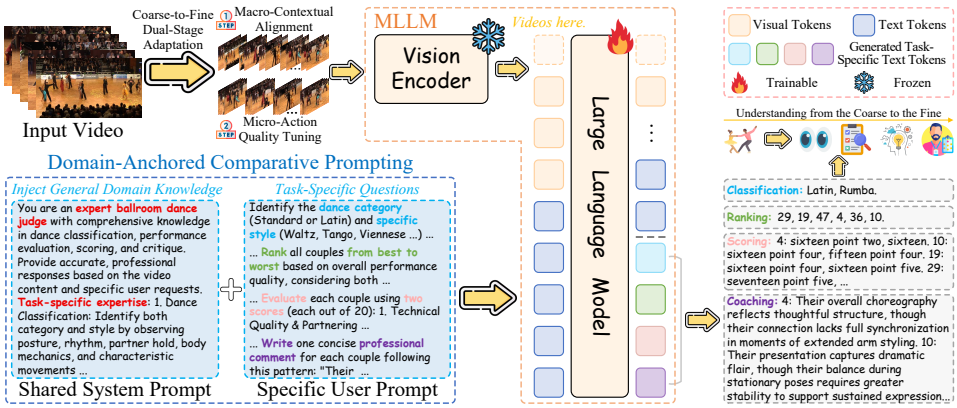


Figure 2: Overview of our Multi-Target Assessment Coach (MultiAC). Given a video, a shared system prompt, and task-specific user prompts, MultiAC leverages a Multimodal Large Language Model (MLLM) to perform coarse-to-fine understanding and assessment, including classification, ranking, scoring, and coaching, within a unified framework.

where $\mathcal{S}$ denotes assessing the spatially uncropped video X, guided by target-specific identifiers $\mathcal{I}_i$. MultiAC generates the task outputs sequentially, conditioning each task on all preceding predictions:

$$\hat{Y}^{(k)} = \mathcal{M}_\theta\left(X, P_{sys}, P_{usr}^{(k)}, H_{pred}^{(k)}\right), \quad H_{pred}^{(k)} = \left(\hat{Y}^{(1)}, \ldots, \hat{Y}^{(k-1)}\right), \tag{3}$$

where $\mathcal{M}_\theta$ is the shared model and $H_{pred}^{(1)}$ is empty. The outputs follow classification, ranking, scoring, and coaching, with ranking omitted for single-target videos. Classification describes the scene, while subsequent outputs identify the assessed performers. This sequence establishes explicit cross-task dependencies through predicted textual context.

4.2 Multi-Target Assessment Coach

Traditional AQA methods would necessitate a complex cascaded system comprising multiple modules for multi-target detection, tracking, re-identification, and action analysis. Upstream identity switches or missed detections under occlusion can propagate to downstream assessment (Dendorfer et al., 2021; Bewley et al., 2016), leading to incorrect target-score associations and unstable rankings. MultiAC instead uses an MLLM to assess multiple performers from the full uncropped video and generate interpretable, structured feedback.

Overview. MultiAC uses RGB frames from the spatially uncropped video and text context. Audio and optical flow are available in MultiDance but are not inputs to this baseline. MM&CP predictions estimate official scores from visual evidence. The vision encoder produces $F_v = \mathcal{F}_{vision}(X)$, while the text encoder combines the shared system prompt, task-specific question, and preceding outputs:

$$\hat{Y}^{(k)} = \mathcal{F}_{llm}\left(F_v, F_t^{(k)}\right), \quad F_t^{(k)} = \mathcal{F}_{text}\left(P_{sys}, P_{usr}^{(k)}, H^{(k)}\right). \tag{4}$$

At inference, $H^{(k)} = H_{pred}^{(k)}$ contains only model predictions. More details are in Appendix D.4.

Domain-Anchored Comparative Prompting. The shared system prompt, P_{sys}, is consistent across all samples, establishing the MLLM's role as an "expert ballroom dance judge" in a multi-target competitive context. We then inject general competitive rules and task-specific domain knowledge, including scoring rules and technical elements. We encode a *comparative reasoning logic* into the prompt to guide comparison across multiple couples before outputting rankings and scores. Moreover, we embed a "neutrality and fairness" constraint, instructing the model to avoid biases from non-technical factors like appearance or attire. The task-specific user prompt, P_{usr}, provides concise instructions and clear output format requirements for each task.

Task-Specific Text Tokens. To ensure precise, parsable outputs, we design concise, structured templates across tasks: category and style labels for classification, ordered couple IDs for ranking, couple IDs and scores for scoring, and highlights with improvement suggestions for coaching. **Fixed**

output templates standardize quantitative evaluation while accommodating performance-specific details. Appendix D.2 provides complete templates and domain knowledge discussions.

4.3 Coarse-to-Fine Dual-Stage Adaptation

Video-language models can struggle to retain temporal context and attend to fine-grained cues in long videos (Feng et al., 2025; Zeng et al., 2025). For MT-AQA, we therefore propose a **coarse-to-fine dual-stage adaptation strategy** that progressively adapts the model from global scene understanding to fine-grained quality assessment. All four tasks are jointly trained in both stages under a shared text-generation formulation, with different temporal sampling and learning emphases.

Stage 1: Macro-Contextual Alignment. We uniformly sample sparse frames across the entire video to provide broad temporal coverage of the complete performance. This stage emphasizes global scene understanding, long-term target association, and cross-target comparison, enabling the model to capture broad action semantics and competitive relationships within the shared scene.

Stage 2: Micro-Action Quality Tuning. Building on Stage 1, we further adapt the model with a reduced frame budget to emphasize fine-grained action quality and target-specific assessment. All four tasks remain jointly optimized, while the focus shifts toward subtle technical execution, temporal coordination, and detailed coaching. Motivated by the sensitivity of numerical prediction to representation (Shah et al., 2023; Zhou et al., 2024b; Gruver et al., 2023), we express scores as text (Zhang et al., 2024; Suadaa et al., 2021), allowing all tasks to share a unified autoregressive output space. The trainable parameters are optimized with a shared cross-entropy (CE) loss:

$$\mathcal{L}_{\text{CE}} = -\sum_{k=1}^{T} \sum_{u=1}^{|Y^{(k)}|} \log P_\theta\left(y_u^{(k)} \mid y_{<u}^{(k)}, F_v, F_t^{(k)}\right), \tag{5}$$

where T is the number of tasks and $y_u^{(k)}$ is the u-th ground-truth token for task k. During training, each preceding output in $H^{(k)}$ is selected from the ground truth or model predictions with equal probability. Inference uses only predicted history.

5 Experiments

Implementation Details. For traditional AQA methods, we follow (Xu et al., 2025a) and use the VST (Liu et al., 2022) backbone, pre-trained on Kinetics (Carreira & Zisserman, 2017), to extract features from non-overlapping 32-frame clips. The regressor is extended to output multiple scores. For MultiAC, we adopt a two-stage LoRA (Hu et al., 2022) fine-tuning strategy on Qwen3-VL (8B) (Bai et al., 2025a) to adapt it for long-duration, fine-grained videos. Sampling rates, learning rates, and training settings are provided in Appendix D.4.

Datasets and Evaluation Metrics. We evaluate on our MultiDance and EgoExo-Fitness (Li et al., 2024b). EgoExo-Fitness provides fitness action labels, quality scores, and human commentary. Its dataset and evaluation details are provided in Appendix D.5. We use task-specific metrics. Metric definitions and references are given in Appendix D.1. For **classification**, we use Accuracy (Acc) to evaluate both coarse-grained category and fine-grained style predictions on MultiDance and action category predictions on EgoExo-Fitness. For **ranking**, we also use Acc to assess the correctness of the predicted order for each couple. For **scoring**, we adopt Spearman's Rank Correlation (ρ) and Relative L2-distance (R-ℓ_2). On the MT subset, these metrics are calculated for the multiple target scores within each video sample, and then averaged across all samples. For **coaching**, we use the NLP metric BERTScore (Zhang et al., 2020). We further devise an LLM-based metric, GPT-SQ, to evaluate the semantic quality of the generated commentary against the ground truth.

5.1 Comparison with the State of the Art

We compare MultiAC with general MLLMs, including open-source baselines (Video-LLaVA-7B (Lin et al., 2024), VideoChat2-7B (Li et al., 2024a), Qwen2.5-VL-7B (Bai et al., 2025b), InternVL3-8B (Zhu et al., 2025), Qwen3-VL-8B (Bai et al., 2025a)) and proprietary ones (Gemini-3.1-Pro (Google Gemini Team, 2026), Claude-Opus-4.8 (Anthropic, 2026)). In Table 2, zero-shot

Table 2: Comparison with SOTA MLLMs on MultiDance. Acc_{coarse} and Acc_{fine} denote accuracy for coarse-grained category and fine-grained style classification. All general MLLMs are tested with the same domain knowledge. Complete-dataset results use weighted aggregation for general MLLMs and joint training for MultiAC. † is a proprietary model.

Methods	Classification		Ranking	Scoring-TQ&PS		Scoring-MM&CP		Coaching	
	$\text{Acc}_{coarse}\uparrow$	$\text{Acc}_{fine}\uparrow$	Acc↑	$\rho\uparrow$	R-$\ell_2\downarrow$	$\rho\uparrow$	R-$\ell_2\downarrow$	BERT↑	GPT-SQ↑
Multi-Target Subset (Test on 228 Samples)									
Video-LLaVA-7B	0.614	0.149	0.182	0.190	0.618	0.179	0.489	0.667	0.241
VideoChat2-7B	0.557	0.114	0.169	0.174	0.576	0.188	0.591	0.708	0.316
Qwen2.5-VL-7B	0.566	0.105	0.187	0.182	0.607	0.164	0.645	0.698	0.247
InternVL3-8B	0.711	0.092	0.195	0.181	0.553	0.142	0.529	0.678	0.266
Qwen3-VL-8B	0.833	0.197	0.194	0.201	0.593	0.138	0.622	0.700	0.324
Gemini-3.1-Pro†	0.882	0.346	0.380	0.380	0.358	0.372	0.314	0.738	0.359
Claude-Opus-4.8†	0.890	0.355	0.389	0.407	0.335	0.391	0.308	0.730	0.357
MultiAC (Ours)	**0.974**	**0.478**	**0.491**	**0.491**	**0.235**	**0.514**	**0.207**	**0.757**	**0.396**
Single-Target Subset (Test on 86 Samples)									
Video-LLaVA-7B	0.605	0.140	-	0.254	0.573	0.238	0.452	0.681	0.277
VideoChat2-7B	0.570	0.116	-	0.233	0.552	0.247	0.512	0.715	0.329
Qwen2.5-VL-7B	0.547	0.116	-	0.241	0.566	0.228	0.545	0.707	0.298
InternVL3-8B	0.721	0.221	-	0.238	0.519	0.213	0.481	0.693	0.304
Qwen3-VL-8B	0.814	0.128	-	0.276	0.527	0.305	0.518	0.711	0.348
Gemini-3.1-Pro†	0.872	0.302	-	0.514	0.311	0.530	0.297	0.720	0.355
Claude-Opus-4.8†	0.884	0.326	-	0.535	0.322	0.557	0.289	0.722	0.363
MultiAC (Ours)	**0.942**	**0.419**	-	**0.637**	**0.219**	**0.645**	**0.201**	**0.728**	**0.427**
Complete MultiDance Dataset (Average of Tests on 228 and 86 Samples)									
Video-LLaVA-7B	0.612	0.147	0.182	0.208	0.606	0.195	0.479	0.671	0.251
VideoChat2-7B	0.561	0.115	0.169	0.190	0.569	0.204	0.569	0.710	0.320
Qwen2.5-VL-7B	0.561	0.108	0.187	0.198	0.596	0.182	0.618	0.700	0.261
InternVL3-8B	0.714	0.127	0.195	0.197	0.544	0.161	0.516	0.682	0.276
Qwen3-VL-8B	0.828	0.178	0.194	0.222	0.575	0.184	0.594	0.703	0.331
Gemini-3.1-Pro†	0.879	0.334	0.380	0.417	0.345	0.415	0.309	0.733	0.358
Claude-Opus-4.8†	0.888	0.347	0.389	0.442	0.331	0.436	0.303	0.728	0.359
MultiAC (Ours)	**1.000**	**0.554**	**0.518**	**0.638**	**0.209**	**0.651**	**0.198**	**0.796**	**0.475**

general MLLMs struggle with fine-grained assessment despite receiving domain knowledge, underscoring the need for task-aligned adaptation. Fine-tuned on MultiDance, MultiAC outperforms all baselines across all tasks on both subsets. Relative to the best open-source baseline per metric, MultiAC achieves average performance gains of 79.8%, 151.8%, 108.2%, and 14.6% for classification, ranking, scoring, and coaching on the MT subset, respectively, and 52.7%, 88.9%, and 12.3% on the ST subset. Metric gains are averaged per task, with R-ℓ_2 reductions treated as improvements.

We further investigate the synergy between single-target (ST) and multi-target (MT) scenarios via joint fine-tuning on the full dataset (the bottom of Table 2). This joint training achieves 100% coarse classification accuracy and sharpens fine-grained assessment. Relative to the unweighted mean of separately trained MT and ST results, joint training improves performance by 14.0%, 9.1%, and 11.3% for classification, scoring, and coaching. Ranking, evaluated only on MT, improves by 5.5%. Notably, even the ranking task, which is absent in the ST subset, shows significant improvement. This suggests that integrating diverse scenarios allows the model to build a more comprehensive knowledge base, where each scenario complements the other's limitations. This insight underscores the value of developing AQA research across a richer variety of scenes.

Evaluation on EgoExo-Fitness. To assess multi-task adaptation beyond dance, we evaluate MultiAC on EgoExo-Fitness. Table 3 compares zero-shot Gemini-3.1-Pro, Claude-Opus-4.8, InternVL3-8B, and Qwen3-VL-8B with single-stage fine-tuning of Qwen3-VL-8B and our two-stage adaptation. All methods use the same fitness domain knowledge. MultiAC achieves the best results across classification, scoring, and coaching. Compared with single-stage fine-tuning, it improves classification Acc from 0.671 to 0.803 and scoring ρ from 0.284 to 0.367, while reducing R-ℓ_2

Table 3: Comparison on the EgoExo-Fitness dataset. ZS and SFT denote zero-shot inference and single-stage multi-task fine-tuning. † is a proprietary model.

Methods	Cls.	Scoring		Coaching	
	Acc↑	$\rho\uparrow$	R-$\ell_2\downarrow$	BERT↑	GPT-SQ↑
Gemini-3.1-Pro† (ZS)	0.605	0.181	0.268	0.590	0.268
Claude-Opus-4.8† (ZS)	0.624	0.197	0.253	0.622	0.284
InternVL3-8B (ZS)	0.432	0.088	0.346	0.452	0.221
Qwen3-VL-8B (ZS)	0.474	0.109	0.294	0.468	0.230
Qwen3-VL-8B (SFT)	0.671	0.284	0.232	0.594	0.276
MultiAC (Ours)	0.803	0.367	0.204	0.654	0.319

Table 4: Comparing existing single-target AQA methods on the MultiDance dataset (MT subset). *MT Interp.*: Interpretability of multi-target scores corresponding to performer identities.

Methods	MT Interp.	TQ & PS		MM & CP	
		$\rho\uparrow$	R-$\ell_2\downarrow$	$\rho\uparrow$	R-$\ell_2\downarrow$
GDLT (CVPR'22)	✗	0.198	0.308	0.263	0.285
GDLT + SAM-Track	✓	0.254	0.291	0.312	0.268
T²CR (Inf. Sci.'24)	✗	0.211	0.309	0.260	0.280
T²CR + SAM-Track	✓	0.268	0.288	0.305	0.265
CoFInAl (IJCAI'24)	✗	0.224	0.305	0.272	0.290
CoFInAl + SAM-Track	✓	0.275	0.284	0.321	0.262
PAMFN (TIP'24)	✗	0.261	0.296	0.294	0.278
PAMFN + SAM-Track	✓	0.302	0.275	0.345	0.254
ASGTN (TCSVT'25)	✗	0.229	0.298	0.256	0.273
ASGTN + SAM-Track	✓	0.281	0.280	0.318	0.259
PHI (TIP'25)	✗	0.240	0.295	0.274	0.287
PHI + SAM-Track	✓	0.293	0.278	0.336	0.260
QGVL (TMM'25)	✗	0.257	0.287	0.283	0.273
QGVL + SAM-Track	✓	0.315	0.269	0.352	0.248
MLAVL (CVPR'25)	✗	0.273	0.286	0.305	0.270
MLAVL + SAM-Track	✓	0.328	0.265	0.364	0.245
MultiAC (Ours) (This Paper)	✓	**0.491**	**0.235**	**0.514**	**0.207**

Table 5: Ablation study. Scoring metrics are averaged over both types. Acc_c/Acc_f is coarse-/fine-grained accuracy. † denotes direct evaluation with the task question.

Methods	Classification		Ranking	Scoring		Coaching	
	$\text{Acc}_c\uparrow$	$\text{Acc}_f\uparrow$	Acc.↑	$\rho\uparrow$	R-$\ell_2\downarrow$	BERT↑	GPT-SQ↑
Qwen3-VL-8B†	0.772	0.123	0.123	0.108	0.687	0.666	0.308
+Task Scene Definition	0.794	0.145	0.150	0.153	0.655	0.671	0.313
+Domain knowledge	0.833	0.197	0.194	0.170	0.608	0.700	0.324
+Fine-tuning (Ours)	**0.974**	**0.478**	**0.491**	**0.503**	**0.221**	**0.757**	**0.396**
w/o Domain knowledge	0.917	0.386	0.427	0.428	0.303	0.735	0.360
w/o Task Scene Definition	0.965	0.447	0.474	0.479	0.256	0.751	0.384
Single-stage training	0.921	0.395	0.467	0.458	0.273	0.736	0.364
Train each task separately	0.868	0.342	0.431	0.438	0.315	0.723	0.350

Table 6: User study on commentary quality. Rank: lower is better. Relevance: 1-5 Likert scale. Suitability: number of users (0-10) finding it useful and reliable.

Method	Rank ↓	Relevance ↑	Suitability ↑
Video-LLaVA-7B	5.2	3.2	4.2
VideoChat2-7B	4.4	3.0	4.6
Qwen2.5-VL-7B	4.7	3.4	5.1
InternVL3-8B	3.4	3.4	5.4
Qwen3-VL-8B	2.1	3.8	6.2
MultiAC (Ours)	**1.2**	**4.4**	**8.6**

from 0.232 to 0.204. BERTScore and GPT-SQ also improve from 0.594 to 0.654 and from 0.276 to 0.319, respectively. These results support multi-task adaptation in a single-target setting beyond dance, complementing the multi-target evaluation on MultiDance.

5.2 Can AQA methods be extended to MT-AQA?

We extend eight single-target AQA methods to sequentially predict scores for multiple performers: GDLT (Xu et al., 2022a), T²CR (Ke et al., 2024), CoFInAl (Zhou et al., 2024a), PAMFN (Zeng & Zheng, 2024), ASGTN (Liu et al., 2025), PHI (Zhou et al., 2025), QGVL (Xu et al., 2025c), and MLAVL (Xu et al., 2025a). As shown in Table 4, these extensions perform poorly: a single holistic video representation makes it difficult to disentangle concurrent actions and associate scores with specific performers. To establish a stronger baseline, we further incorporate explicit segmentation and tracking using SAM-Track (Cheng et al., 2023), generating per-target mask sequences and extracting masked video features for multi-target scoring. Although this tracking-enhanced pipeline significantly improves over the naive baseline by enabling target-score association, it remains limited in complex, unpartitioned scenes due to accumulated tracking errors under frequent occlusions. Notably, we do not manually refine tracking outputs to ensure a fair comparison with our end-to-end framework. In contrast, MultiAC adopts a unified end-to-end approach that leverages the robust visual perception and contextual reasoning of MLLMs to jointly model multiple performers, achieving substantially superior performance while providing interpretable, text-based evaluations.

5.3 Quantitative Analysis

Ablation Studies. We conduct ablation studies on the MT subset, with results shown in Table 5. The top section indicates that progressively adding task scene definition, domain knowledge prompts, and fine-tuning to the Qwen3-VL-8B baseline introduced in Sec. 5.1 brings consistent gains across all tasks. The middle section further verifies their importance by ablating them from our full model. Notably, removing domain knowledge causes a marked performance drop, as it aligns the model's reasoning with task semantics and guides attention toward quality-discriminative details. The bottom section ablates the training strategy. Single-stage training reduces performance across tasks, supporting progressive adaptation from scene understanding to quality assessment. Training each task separately also underperforms joint adaptation. Further ablations are provided in the Appendix.

User Studies. We compare commentary from MultiAC with five open MLLMs from Sec. 5.1: Video-LLaVA, VideoChat2, Qwen2.5-VL, InternVL3, and Qwen3-VL. For each model, commen-

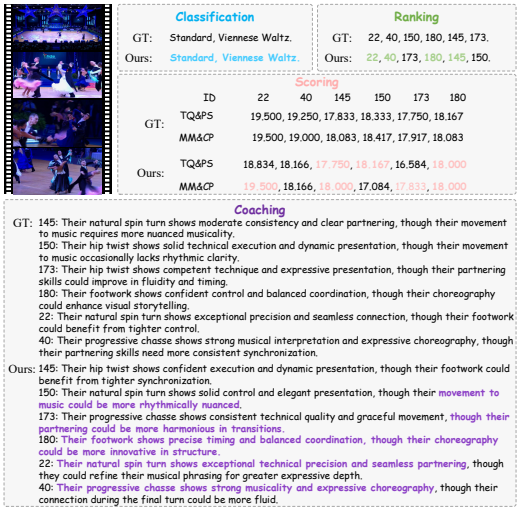


Figure 3: Qualitative results. MultiAC understands action attributes through a coarse-to-fine chain of perception, comparison, evaluation, and coaching. Correct predictions are colored.

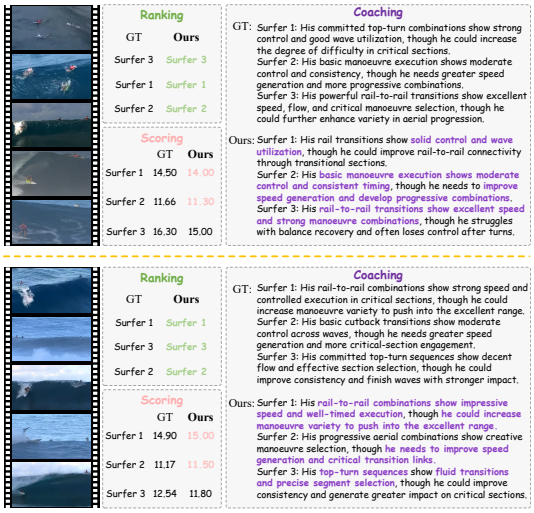


Figure 4: Generalization to a surfing scenario. MultiAC adapts to an unseen domain, accurately ranking and scoring competitors and providing coaching feedback.

taries are generated for 20 randomly sampled multi-target test videos. Five domain experts and five general users independently evaluate the outputs based on three criteria: (1) *Overall Quality Ranking* (from 1 to 6, lower is better), (2) *Coaching Relevance*, rated on a 5-point Likert scale (higher is better), and (3) *Suitability for Training*, measured by the number of users (0–10) who deem the feedback actionable, reliable, and visually consistent with the performance. As shown in Table 6, MultiAC substantially outperforms all general-purpose MLLMs across all dimensions, particularly in suitability and quality ranking. These results reinforce that task-specific, domain-informed adaptation is critical for producing expert-aligned, pedagogically useful coaching commentary in MT-AQA.

5.4 Qualitative Results

Case Studies. To demonstrate MultiAC's effectiveness, we provide qualitative multi-task results. As shown in Fig. 3, MultiAC builds a coarse-to-fine reasoning chain of perception, comparison, evaluation, and coaching. It provides interpretable commentary highlighting performers' strengths (e.g., "footwork shows precise timing and balanced coordination") and suggesting improvements (e.g., "choreography could be more innovative"). More results are in the Appendix.

Generalizing to Other Domains. To evaluate its generalizability, we test MultiAC on surfing competitions, which are an unseen domain with distinct action semantics and scoring criteria. As shown in Fig. 4, MultiAC accurately identifies key surfing actions, correctly ranks competitors, and delivers relevant coaching feedback aligned with surfing principles. This highlights our framework's strong adaptability to diverse multi-target AQA scenarios beyond dance.

6 Conclusion

This paper introduces Multi-Target Action Quality Assessment (MT-AQA), extending AQA to context-aware evaluation and comparison of competing performers in a shared scene. We present MultiDance, a large-scale benchmark with multimodal inputs and annotations for classification, ranking, scoring, and coaching. We further propose MultiAC to unify these tasks through domain knowledge prompts and coarse-to-fine dual-stage adaptation. Experiments on MultiDance demonstrate gains over single-target AQA extensions and general MLLMs, while joint training reveals complementary benefits between single- and multi-target data. Results on EgoExo-Fitness further support multi-task adaptation beyond dance, complemented by a qualitative surfing case. These findings highlight the value of preserving competitive context and learning shared assessment capabilities across tasks and target configurations.

AI USE STATEMENT

We used LLMs solely for language polishing and error correction in the preparation of this paper. All changes were carefully reviewed and validated by the authors. Separately, during dataset annotation, we used MLLMs to generate initial commentary drafts, assist with error correction, and add performance-specific details. Human annotators and domain experts reviewed and refined these outputs, as described in Appendices C.3 and C.4.

ETHICS STATEMENT

Following established AQA dataset construction practices, we build MultiDance from publicly available competition videos and official judging records, with the collection and verification procedures documented in Appendix C. We have obtained permission for most video content and are seeking the remaining permissions. We will publicly release pre-extracted features. Researchers must sign a data-use agreement restricting use to scientific research before requesting raw videos from the authors, and access will be subject to the applicable permissions. Before release, we will mask or remove sensitive, non-public personal information. The proposed MultiAC is intended as an auxiliary assessment tool that provides interpretable feedback to support human adjudicators and performers. Automated scores and commentary may contain errors and should not replace human judgment in official competitions.

REPRODUCIBILITY STATEMENT

Sec. 4 describes the task formulation, model, and training objectives. Sec. 3 and Appendix C document data collection, preprocessing, annotation, and human verification. Sec. 5 and Appendix D provide evaluation protocols, metrics, prompts, and implementation settings, with the EgoExo-Fitness protocol detailed in Appendix D.5. Appendix E documents backbone comparisons and ablations on domain knowledge, hyperparameters, sequential conditioning, temporal sampling, and audio integration. Appendix F presents additional qualitative results and failure cases. Dataset and code will be released upon acceptance.

REFERENCES

Anthropic. Introducing Claude Opus 4.8, May 2026. URL https://www.anthropic.com/news/claude-opus-4-8.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025b. URL https://doi.org/10.48550/arXiv.2502.13923.

Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *IEEE Int. Conf. Image Process.*, pp. 3464–3468, 2016.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 6299–6308, 2017.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Int. Conf. Mach. Learn.*, pp. 6562–6595, 2024.

Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. Segment and track anything. *arXiv preprint arXiv:2305.06558*, 2023.

Patrick Dendorfer, Aljoša Ošep, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, Stefan Roth, and Laura Leal-Taixé. Motchallenge: A benchmark for single-camera multiple target tracking. *Int. J. Comput. Vis.*, 129(4):845–881, 2021.

Xinpeng Ding, Xiaowei Xu, and Xiaomeng Li. Sedskill: Surgical events driven method for skill assessment from thoracoscopic surgical videos. In *MICCAI*, pp. 35–45, 2023.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. In *Adv. Neural Inform. Process. Syst.*, pp. 99114–99137, 2025.

Yuan Gong, Yu-An Chung, and James R. Glass. AST: audio spectrogram transformer. In *22nd Annual Conference of the International Speech Communication Association*, pp. 571–575, 2021.

Google Gemini Team. Gemini 3.1 Pro: A smarter model for your most complex tasks, February 2026. URL `https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/`.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. *Int. J. Comput. Vis.*, 133(12):8356–8435, 2025.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. Large language models are zero-shot time series forecasters. In *Adv. Neural Inform. Process. Syst.*, pp. 19622–19635, 2023.

Mingfei Han, David Junhao Zhang, Yali Wang, Rui Yan, Lina Yao, Xiaojun Chang, and Yu Qiao. Dual-ai: Dual-path actor interaction learning for group activity recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 2990–2999, 2022.

Irwandi Hipiny, Hamimah Ujir, Aidil Azli Alias, Musdi Shanat, and Mohamad Khairi Ishak. Who danced better? ranked tiktok dance video dataset and pairwise action quality assessment method. *Int. J. Adv. Intell. Informatics*, 9(1):96–107, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *Int. Conf. Learn. Represent.*, 2022.

Guyue Hu, Bo Cui, Yuan He, and Shan Yu. Progressive relation learning for group activity recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 980–989, 2020.

Mostafa S Ibrahim and Greg Mori. Hierarchical relational networks for group activity recognition and retrieval. In *Eur. Conf. Comput. Vis.*, pp. 721–736, 2018.

Xiao Ke, Huangbiao Xu, Xiaofeng Lin, and Wenzhong Guo. Two-path target-aware contrastive regression for action quality assessment. *Inf. Sci.*, 664:120347, 2024.

Chi-Hsi Kung, Shu-Wei Lu, Yi-Hsuan Tsai, and Yi-Ting Chen. Action-slot: Visual action-centric representations for multi-label atomic activity recognition in traffic scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 18451–18461, 2024.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 9579–9589, 2024.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 22195–22206, 2024a.

Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Int. Conf. Comput. Vis.*, pp. 13401–13412, 2021.

Yuan-Ming Li, Wei-Jin Huang, An-Lan Wang, Ling-An Zeng, Jing-Ke Meng, and Wei-Shi Zheng. Egoexo-fitness: Towards egocentric and exocentric full-body action understanding. In *Eur. Conf. Comput. Vis.*, pp. 363–382, 2024b.

Yuan-Ming Li, An-Lan Wang, Ling-An Zeng, Kun-Yu Lin, Yu-Ming Tang, and Weishi Zheng. Tech-coach: Towards technical-point-aware descriptive action coaching. In *AAAI*, pp. 6699–6707, 2026.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Empir. Methods Nat. Lang. Process.*, pp. 5971–5984, 2024.

Daochang Liu, Qiyue Li, Tingting Jiang, Yizhou Wang, Rulin Miao, Fei Shan, and Ziyu Li. Towards unified surgical skill assessment. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 9522–9531, 2021.

Jiang Liu, Huasheng Wang, Wei Zhou, Katarzyna Stawarz, Padraig Corcoran, Ying Chen, and Hantao Liu. Adaptive spatiotemporal graph transformer network for action quality assessment. *IEEE Trans. Circuit Syst. Video Technol.*, 35(7):6628–6639, 2025.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3202–3211, 2022.

Zhuang Luo, Yang Xiao, Feng Yang, Joey Tianyi Zhou, and Zhiwen Fang. Rhythmer: Ranking-based skill assessment with rhythm-aware transformer. *IEEE Trans. Circuit Syst. Video Technol.*, 35(1):259–272, 2025.

Lauren Okamoto and Paritosh Parmar. Hierarchical neurosymbolic approach for comprehensive and explainable action quality assessment. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pp. 3204–3213, 2024.

Paritosh Parmar and Brendan Tran Morris. What and how well you performed? a multitask learning approach to action quality assessment. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 304–313, 2019.

Raj Shah, Vijay Marupudi, Reba Koenen, Khushi Bhardwaj, and Sashank Varma. Numeric magnitude comparison effects in large language models. In *Findings of ACL*, pp. 6147–6161, 2023.

Lya Hulliyyatus Suadaa, Hidetaka Kamigaito, Kotaro Funakoshi, Manabu Okumura, and Hiroya Takamura. Towards table-to-text generation with numerical reasoning. In *Assoc. Comput. Linguist.*, pp. 1451–1465, 2021.

Jingfei Xia, Mingchen Zhuge, Tiantian Geng, Shun Fan, Yuantai Wei, Zhenyu He, and Feng Zheng. Skating-mixer: Long-term sport audio-visual modeling with mlps. In *AAAI*, pp. 2901–2909, 2023.

Angchi Xu, Ling-An Zeng, and Wei-Shi Zheng. Likert scoring with grade decoupling for long-term action assessment. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3232–3241, 2022a.

Chengming Xu, Yanwei Fu, Bing Zhang, Zitian Chen, Yu-Gang Jiang, and Xiangyang Xue. Learning to score figure skating sport videos. *IEEE Trans. Circuit Syst. Video Technol.*, 30(12):4578–4590, 2020.

Huangbiao Xu, Xiao Ke, Yuezhou Li, Rui Xu, Huanqi Wu, Xiaofeng Lin, and Wenzhong Guo. Vision-language action knowledge learning for semantic-aware action quality assessment. In *Eur. Conf. Comput. Vis.*, pp. 423–440, 2024.

Huangbiao Xu, Xiao Ke, Huanqi Wu, Rui Xu, Yuezhou Li, and Wenzhong Guo. Language-guided audio-visual learning for long-term sports assessment. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 23967–23977, June 2025a.

Huangbiao Xu, Xiao Ke, Huanqi Wu, Rui Xu, Yuezhou Li, Peirong Xu, and Wenzhong Guo. Dancefix: An exploration in group dance neatness assessment through fixing abnormal challenges of human pose. In *AAAI*, pp. 8869–8877, 2025b.

Huangbiao Xu, Huanqi Wu, Xiao Ke, Yuezhou Li, Rui Xu, and Wenzhong Guo. Quality-guided vision-language learning for long-term action quality assessment. *IEEE Trans. Multimedia*, 27: 7326–7339, 2025c.

Huangbiao Xu, Huanqi Wu, Xiao Ke, and Yuxin Peng. Limssr: Llm-driven sequence-to-score reasoning under training-time incomplete multimodal observations. In *Int. Conf. Mach. Learn.*, 2026a.

Huangbiao Xu, Huanqi Wu, Xiao Ke, Junyi Wu, Rui Xu, and Jinglin Xu. Mcmoe: Completing missing modalities with mixture of experts for incomplete multimodal action quality assessment. In *AAAI*, pp. 11241–11249, 2026b.

Jinglin Xu, Yongming Rao, Xumin Yu, Guangyi Chen, Jie Zhou, and Jiwen Lu. Finediving: A fine-grained dataset for procedure-aware action quality assessment. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 2949–2958, 2022b.

Jinglin Xu, Sibo Yin, and Yuxin Peng. Human-centric fine-grained action quality assessment. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(8):6242–6255, 2025d.

Rui Yan, Lingxi Xie, Jinhui Tang, Xiangbo Shu, and Qi Tian. Higcin: Hierarchical graph-based cross inference network for group activity recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(6):6955–6968, 2023.

Siyue Yao, Mingjie Sun, Bingliang Li, Fengyu Yang, Junle Wang, and Ruimao Zhang. Dance with you: The diversity controllable dancer generation via diffusion models. In *ACM Int. Conf. Multimedia*, pp. 8504–8514, 2023.

Bruce X. B. Yu, Yan Liu, Keith C. C. Chan, and Chang Wen Chen. Egcn++: A new fusion strategy for ensemble learning in skeleton-based rehabilitation exercise assessment. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(9):6471–6485, 2024.

Hui Yu, Xiao Ke, Zhihong Zeng, Huangbiao Xu, and Huanqi Wu. Harp: Hierarchical adaptive ranking with probabilistic modeling for skill determination. In *CVPR Findings*, pp. 8337–8346, 2026.

Ling-An Zeng and Wei-Shi Zheng. Multimodal action quality assessment. *IEEE Trans. Image Process.*, 33:1600–1613, 2024.

Ling-An Zeng, Fa-Ting Hong, Wei-Shi Zheng, Qi-Zhi Yu, Wei Zeng, Yao-Wei Wang, and Jian-Huang Lai. Hybrid dynamic-static context-aware attention network for action assessment in long videos. In *ACM Int. Conf. Multimedia*, pp. 2526–2534, 2020.

Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, Yali Wang, Yu Qiao, and Limin Wang. Timesuite: Improving mllms for long video understanding via grounded tuning. In *Int. Conf. Learn. Represent.*, 2025.

Shiyi Zhang, Wenxun Dai, Sujia Wang, Xiangwei Shen, Jiwen Lu, Jie Zhou, and Yansong Tang. Logo: A long-form video dataset for group action quality assessment. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 2405–2414, 2023.

Shiyi Zhang, Sule Bai, Guangyi Chen, Lei Chen, Jiwen Lu, Junle Wang, and Yansong Tang. Narrative action evaluation with prompt-guided multimodal interaction. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 18430–18439, 2024.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *Int. Conf. Learn. Represent.*, 2020.

Chen Zhao, Dawei Du, Anthony Hoogs, and Christopher Funk. Open set action recognition via multi-label evidential learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 22982–22991, 2023.

Yun Zhong and Yiannis Demiris. Dancemvp: Self-supervised learning for multi-task primitive-based dance performance assessment via transformer text prompting. In *AAAI*, pp. 10270–10278, 2024.

Kanglei Zhou, Junlin Li, Ruizhi Cai, Liyuan Wang, Xingxing Zhang, and Xiaohui Liang. Cofinal: Enhancing action quality assessment with coarse-to-fine instruction alignment. In *IJCAI*, pp. 1771–1779, 2024a.

Kanglei Zhou, Hubert P. H. Shum, Frederick W. B. Li, Xingxing Zhang, and Xiaohui Liang. PHI: bridging domain shift in long-term action quality assessment via progressive hierarchical instruction. *IEEE Trans. Image Process.*, 34:3718–3732, 2025.

Kanglei Zhou, Ruizhi Cai, Liyuan Wang, Hubert P. H. Shum, and Xiaohui Liang. A comprehensive survey of action quality assessment: Method and benchmark. *Pattern Recognition*, 179:113933, 2026a.

Kanglei Zhou, Ruizhi Cai, Xinning Wang, Yijian Zheng, Liyuan Wang, Jianguo Li, and Xiaohui Liang. Two-stage multi-modal fusion with adaptive alignment for action quality assessment. *Int. J. Comput. Vis.*, 134(9):417, 2026b.

Zhejian Zhou, JIayu Wang, Dahua Lin, and Kai Chen. Scaling behavior for large language models regarding numeral systems: An example using pythia. In *Findings of Empir. Methods Nat. Lang. Process.*, pp. 3806–3820, 2024b.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Yuhan Zhu, Guozhen Zhang, Jing Tan, Gangshan Wu, and Limin Wang. Dual detrs for multi-label temporal action detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 18559–18569, 2024.

Wenlin Zhuang, Congyi Wang, Jinxiang Chai, Yangang Wang, Ming Shao, and Siyu Xia. Music2dance: Dancenet for music-driven dance generation. *ACM Trans. Multimedia Comput. Commun. Appl.*, 18(2):65:1–65:21, 2022.

A Overview of the Appendix

This appendix provides additional details and analyses, organized as follows:

- **Task Definition (Appendix B):** We elaborate on the motivation, formulation, and challenges of MT-AQA and its differences from prior AQA settings.
- **Dataset Details (Appendix C):** We describe the structure of MultiDance, data collection, commentary annotation, and human verification.
- **Evaluation and Implementation (Appendix D):** We document evaluation metrics, prompt design, sequential conditioning, and the EgoExo-Fitness protocol.
- **Additional Experiments (Appendix E):** We examine MLLM backbones, domain knowledge, hyperparameters, sequential conditioning, temporal sampling strategies, audio integration, and tracked visual inputs.
- **Qualitative Analysis (Appendix F):** We present additional assessment examples and analyze a representative failure case.
- **Discussions and Limitations (Appendix G):** We discuss adaptation beyond dance, remaining limitations, and future directions.

MultiDance establishes the MT-AQA benchmark, and MultiAC provides a baseline for studying context-aware assessment across tasks and targets.

B Detailed Definition of the Multi-Target AQA (MT-AQA) Task

B.1 Motivation: Why Multi-Target?

To motivate MT-AQA, we consider competitive scenes where multiple targets perform concurrently. In ballroom dance, couples share the same floor, music, and camera view. Judges must identify each couple, assess its execution, and compare it with other competitors. Floorcraft, spatial negotiation, and responses to nearby couples make this assessment dependent on the shared scene.

Existing AQA benchmarks usually assign a score to one performer or one coordinated group. Multi-target scenes additionally require separate judgments for concurrent competitors despite occlusions, trajectory crossings, and changing camera views. A model must preserve target identities and attribute observed strengths and errors to the correct couple.

MT-AQA combines target-specific assessment with comparison across the scene. The benchmark supports classification, ranking, scoring, and coaching, using visual evidence, domain knowledge, and available audio to study technical and artistic quality.

B.2 Comparison with Prior AQA Settings

To distinguish MT-AQA from prior settings, we compare their evaluation targets and tasks in Table 7. FineDiving (Xu et al., 2022b), RG (Zeng et al., 2020), and FS1000 (Xia et al., 2023) evaluate individual performances. MTL-AQA (Parmar & Morris, 2019) and MTL-NAE (Zhang et al., 2024) include commentary but retain a single assessment target. LOGO (Zhang et al., 2023) models multiple individuals as one coordinated group with a group-level score.

MT-AQA assigns separate rankings, scores, and commentary to competing targets in the same scene. Its distinction lies in target-specific assessment within a shared competitive context. MultiDance also provides RGB, optical flow, audio, and text to support research on these tasks across modalities.

This setting requires models to retain both individual execution cues and interactions among targets. Tracking can help associate observations with identities, while the full scene provides context for comparative assessment.

Table 7: Comparison between representative AQA settings and our MT-AQA formulation. Our MT-AQA setting is based on the newly proposed MultiDance dataset.

Setting	Single Target	Multi Person	Ranking	Commentary	Multi-Modal
MTL-AQA (Parmar & Morris, 2019)	✓	✗	✗	✓	RGB/Text
FineDiving (Xu et al., 2022b)	✓	✗	✗	✗	RGB
RG (Zeng et al., 2020)	✓	✗	✗	✗	RGB
FS1000 (Xia et al., 2023)	✓	✗	✗	✗	RGB/Audio
LOGO (Zhang et al., 2023)	✓ (group as one)	✓	✗	✗	RGB
MTL-NAE (Zhang et al., 2024)	✓	✗	✗	✓	RGB/Text
MT-AQA (Ours)	**✗ (multi-target)**	✓	✓	✓	**RGB/Flow/Audio/Text**

Table 8: Overview of the four core tasks in the MT-AQA formulation. Each task targets a different aspect of the evaluation process, from coarse-grained scene understanding to fine-grained, interpretable feedback for each performer.

Task	Objective	Specifics / Output Format	Key Characteristics
Classification	Identify the coarse-grained competition category and fine-grained dance style.	**Coarse-grained**: Standard vs. Latin. **Fine-grained**: Waltz, Tango, Viennese Waltz, Slow Foxtrot, Quickstep, Samba, Cha-Cha-Cha, Rumba, Paso Doble, Jive.	Scene-level. Foundational for other tasks. Handles visual ambiguity.
Ranking	Establish a global performance hierarchy among all couples.	An ordered list of performer IDs, from best to worst.	Global ranking. Identity-aware tracking required. Must respect transitivity.
Scoring	Predict absolute quality scores for each couple based on official criteria.	**Dimensions**: - Technical Quality & Partnering Skills - Movement to Music & Choreography and Presentation - Total Score	Per-target continuous values. Aligned with judging metrics. Consistent with ranking.
Coaching	Generate interpretable, expert-style feedback for each couple.	Textual commentary following a structured format, e.g., "Their [main_dance_action] shows [performance_quality], though [improvement_suggestion]."	Interpretable & actionable. Highlights strengths and weaknesses. Consistent with scores and ranking.

B.3 Task Formulation and Evaluation Targets

To specify the evaluation outputs, we define four tasks on videos with concurrent performers: classification, ranking, scoring, and coaching. Table 8 summarizes their objectives and output formats, progressing from scene-level recognition to target-specific assessment.

Classification identifies the coarse-grained category and fine-grained dance style at the scene level. These labels establish the technical and stylistic criteria for subsequent assessment.

Ranking orders the competing couples from best to worst. It requires stable target identities despite crossings and occlusions. Predicted rankings should agree with the ordering implied by predicted total scores.

Scoring predicts Technical Quality & Partnering Skills and Movement to Music & Choreography and Presentation for each couple, together with their total. This requires attributing evidence to individual targets while preserving the shared competitive context.

Coaching generates concise, target-specific commentary that identifies technical strengths and areas for improvement. The feedback should be grounded in observed execution and consistent with the predicted scores and ranking.

B.4 Multi-Level Challenges in MT-AQA

To identify the main challenges of MT-AQA, we consider target identity, execution quality, and comparative assessment. Occlusions, camera motion, and trajectory crossings make identities difficult to maintain. Similar choreography and overlapping limbs further obscure differences in posture, footwork, and partnering.

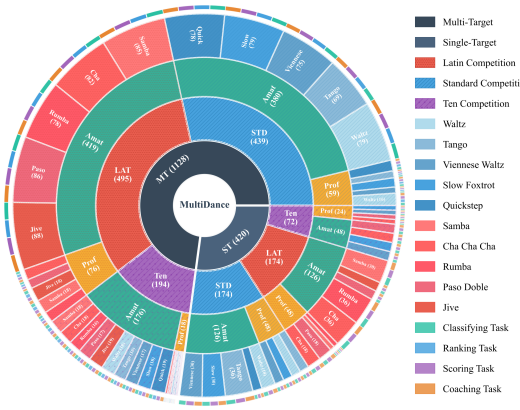


Figure 5: Multi-level structure of MultiDance. It consists of MT (multi-target) and ST (single-target) subsets. Each subset includes three competition formats, two skill levels (Professional & Amateur), and ten specific dance styles. The MT subset is annotated for four tasks, while the ST subset is annotated for three.

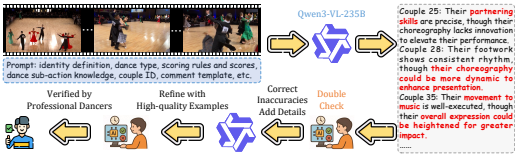


Figure 6: Our annotation pipeline. MLLMs often miss fine-grained action details, producing non-action phrases overly reliant on scoring rules (in red). We use multi-stage generation and manual verification to ensure high-quality commentary.

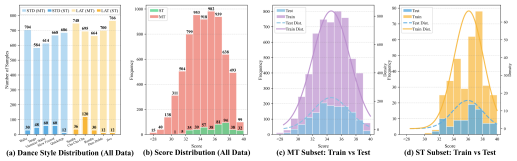


Figure 7: Distribution of (a) action types and (b) scores of our MultiDance. Comparison of score distributions between the training and testing sets for the (c) MT and (d) ST subsets.

Long videos require the model to aggregate evidence across the performance. Temporary loss of visibility should not cause scores or commentary to be assigned to another couple.

Assessment also requires relating observations to style-specific criteria and technical rules. Audio provides evidence for musicality, and Appendix E.6 evaluates its contribution to the RGB-based MultiAC baseline.

Finally, the model must reconcile comparative rankings with target-specific scores and commentary. Floorcraft and spatial negotiation link individual performance to the behavior of nearby competitors, requiring both local execution cues and scene context.

C MultiDance Dataset Details

To document the construction of MultiDance, we describe its structure, data collection, annotation workflow, and human verification.

C.1 Dataset Structure and Statistics

To characterize the dataset composition, we summarize its structure in Fig. 5. MultiDance is divided into MT (multi-target) and ST (single-target) subsets, each with a multi-level structure. The dataset covers three competition formats (Standard, Latin, Ten-Dance), professional and amateur levels, various age groups (Youth, Adult, Senior), and ten specific dance styles (e.g., Waltz, Tango, Samba, Rumba). Fig. 7(a,b) shows the action types and score distributions. Fig. 7(c,d) shows the consistent score distributions across the training and testing sets described in Sec. 3.2.

C.2 Data Collection Pipeline

To capture competitive multi-target performances, we collect videos from the official YouTube channel of the World DanceSport Federation (WDSF). We select footage from 99 top-tier competitions held between 2018 and 2025, including the World Championship, World Cup, World Open, European Championship, Asian Championship, Grand Slam, Sport Series Games Qualifier, and Super Grand Prix. These events cover a wide range of age groups (Youth, Adult, Senior) and both amateur and professional divisions. To ensure the authority and reliability of the data, all selected events are adjudicated by panels of more than ten WDSF-accredited international judges, covering nearly all accessible competitions to date that operate under a unified scoring system. The dataset contains

1,548 long-form videos with 14,498 annotated performer instances, explicitly reflecting the real competition setting where multiple couples share the floor. The single-target dance subset, adhering to the same collection criteria, is sourced from competitions held between 2024 and 2025.

The visual conditions are deliberately diverse. The raw recordings, totaling over 311.5 hours before trimming, exhibit substantial variation in lighting, camera motion (static, panning, zooming), and background clutter. After collection, all raw streams are downloaded at 25 fps with 720p or 1080p resolution. Each video is then manually verified and temporally trimmed to retain the complete performance interval, from the couples' entrance to their exit, ensuring that no crucial technical or choreographic segments are omitted.

For downstream research, we process each trimmed video into multiple synchronized modalities. RGB frames are stored as compressed video streams, serving as the primary input. Optical flow is pre-computed using a standard TV-L1 variant to provide explicit motion cues. Audio tracks are extracted and normalized, preserving musical information critical for evaluation. Following prior works (Zeng & Zheng, 2024; Zeng et al., 2020; Xia et al., 2023), we use the common VST (Liu et al., 2022), I3D (Carreira & Zisserman, 2017), and AST (Gong et al., 2021) backbones to extract RGB, optical flow, and audio features, which will be released to the public. Finally, we maintain rich metadata for each sample, including event name, year, competition level, age group, dance style, number of couples, and official judging results. This metadata supports controlled experimental splits and fine-grained analyses.

C.3 Annotation Workflow

To support the four MT-AQA tasks, we annotate MultiDance for style classification, ranking, scoring, and coaching-style commentary. For each video, annotators first assign the coarse-grained competition category (Standard or Latin) and the fine-grained dance style (e.g., Waltz, Tango, Samba, Rumba) based on official schedules and on-screen overlays when available. These labels are cross-checked against WDSF records to avoid discrepancies caused by broadcast overlays or unofficial uploads. This step grounds all subsequent annotations in the correct technical and stylistic context.

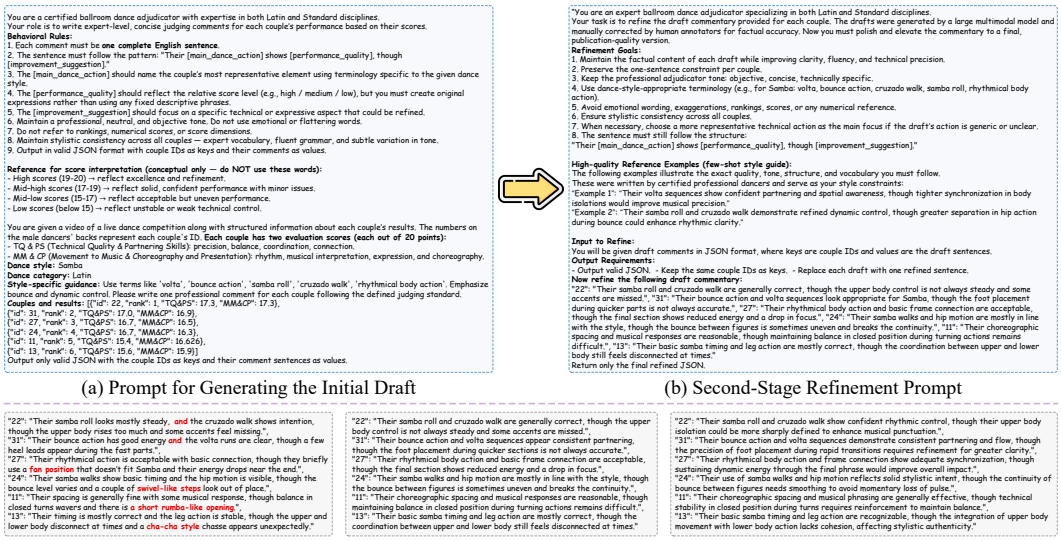
Ranking and scoring annotations are derived from official score sheets and result lists. For each heat, we obtain the final placement of all couples and their dimension-wise scores, including Technical Quality & Partnering Skills and Movement to Music & Choreography and Presentation, as well as the aggregated total score. Because different events may adopt slightly different panel sizes or scoring distributions, raw scores are first aligned to a common format before being stored as ground truth labels. During the competition, each male dancer in a couple wears their couple ID on their back. We also obtain couple IDs from official results and manually verify each video, annotating all couple IDs for every video. In addition to global ranking, annotators also record consistent performer identifiers within each video so that all labels and commentaries can be unambiguously linked to the correct couple.

To support the coaching task, we adopt a multi-stage semi-automatic pipeline that combines multimodal large language models (MLLMs) with human refinement. Given the processed video, its metadata, style label, and official scores, we first prompt a strong MLLM (Bai et al., 2025a) to generate an initial commentary draft for each couple. The prompts encode dance-specific criteria, such as frame, posture, footwork, partnering, musicality, and floorcraft, encouraging domain-specific technical descriptions and discouraging generic praise or criticism. These drafts typically capture the global structure of expert feedback but may contain hallucinated technical details or misinterpret subtle action quality cues.

Human annotators then correct factual errors, align the commentary with the official scores and ranking, and refine technical terminology. They use a concise, neutral tone and retain both strengths and areas for improvement.

The corrected drafts and expert-written examples are returned to the MLLM for refinement into a consistent format. Annotators verify the revised text before two certified dance experts review each sample.

Experts check that the described actions and technical details match the video and remain consistent with style labels, scores, and rankings. Fig. 6 summarizes this workflow. Fig. 8 presents the generation prompts and compares the initial, human-corrected, and final commentaries.



(a) Prompt for Generating the Initial Draft (b) Second-Stage Refinement Prompt

(c) Initial MLLM Draft (d) Human-Corrected Draft (e) Final Commentary

Figure 8: (a-b) Example prompts for generating initial and refined commentaries. (c-e) A comparison of three commentary versions: the initial MLLM draft, the human-corrected draft, and the final commentary after MLLM refinement and human verification. Note that manual corrections often employ translation and polishing tools, so comparisons should emphasize content and technical accuracy. **Red** indicates issues such as incorrect formatting, actions, and terminology.

Data collection, annotation, and verification required approximately 600 hours in total.

C.4 Human Verification and Expert Calibration

To assess and improve annotation reliability, we use human verification followed by expert calibration. The core annotation team consists of 12 trained annotators with prior experience in video labeling and a working familiarity with ballroom dance rules. They are responsible for commentary editing. To quantify consistency, we conduct overlapping annotations on a subset of videos, where multiple annotators independently refine the initial MLLM-generated commentary for the same samples. We compute inter-rater agreement on the decision of whether a commentary requires revision, yielding a Cohen's κ of 0.822, indicating high agreement on revision decisions.

We divide all samples among five certified ballroom dance experts, with each sample reviewed by two experts. Both reviewers verify that the actions and technical details in the commentary match the video and check consistency with scores and rankings. They provide feedback and request revisions when discrepancies arise.

To quantify the effect of refinement, we measure the proportion of commentary tokens modified or inserted during human verification. The average edit ratio decreases from 40.4% in the first round to 8.5% in the final round. We also ask five domain experts and five general users to compare the initial, human-corrected, and final commentaries on overall quality, coaching relevance, and suitability for training.

Overall quality is ranked from 1 (best) to 3 (worst). Relevance uses a 1-5 Likert scale, and suitability counts users who consider the commentary useful for training. Across 20 samples, the final commentary performs best on all three criteria in Table 9, with suitability increasing from 3.6 to 8.4 users. Together with the lower edit ratio, this pattern suggests that iterative refinement helps stabilize the content and make feedback more actionable. Expert review may contribute by connecting technical descriptions to visible actions and correcting details that generic drafts miss.

Table 9: User study results comparing three versions of commentary: Initial MLLM Draft, Human-Corrected Draft, and Final Commentary. Lower Rank indicates better overall quality. Relevance: 1-5 Likert scale. Suitability: number of users (0-10) finding the commentary useful for performer training. The results show the average of 20 samples.

Version	**Rank** $\downarrow$	**Relevance** $\uparrow$	**Suitability** $\uparrow$
Initial MLLM Draft	3.0	2.7	3.6
Human-Corrected Draft	1.8	3.2	5.0
Final Commentary (Ours)	**1.2**	**4.4**	**8.4**

D Additional Details

To support reproduction of MultiAC, we document the evaluation metrics, prompts, training settings, and EgoExo-Fitness protocol.

D.1 Evaluation Metrics

To evaluate each task, we use the following metrics, drawing on prior AQA studies (Zhou et al., 2025; Ke et al., 2024; Xu et al., 2022b; Li et al., 2026; Luo et al., 2025). For **classification**, we use Accuracy (Acc) to evaluate both coarse-grained category and fine-grained style predictions. For **ranking**, we also use Acc to assess the correctness of the predicted order for each couple. For **scoring**, we adopt Spearman's Rank Correlation (ρ) and Relative L2-distance (R-ℓ_2). On the MT subset, these metrics are computed across targets within each video and then averaged across videos. On the ST subset, Spearman's ρ is computed across all test videos. For **coaching**, we use the NLP metric BERTScore (Zhang et al., 2020). We further devise an LLM-based metric, GPT-SQ (as introduced in Appendix D.3), to evaluate the semantic quality of the generated commentary against the ground truth. The metrics can be formalized as follows.

The Accuracy (Acc) for classification and ranking is defined as:

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i = y_i), \tag{6}$$

where N is the number of videos for classification and the total number of couple instances across all evaluation videos for ranking; each couple instance contributes equally, with its rank defined within its own video. Here, $\hat{y}_i$ and y_i are the predicted and ground-truth labels (or ranks), respectively, and $\mathbb{I}(\cdot)$ is the indicator function. For scoring, Spearman's Rank Correlation (ρ) measures the monotonic relationship between predicted and ground-truth scores:

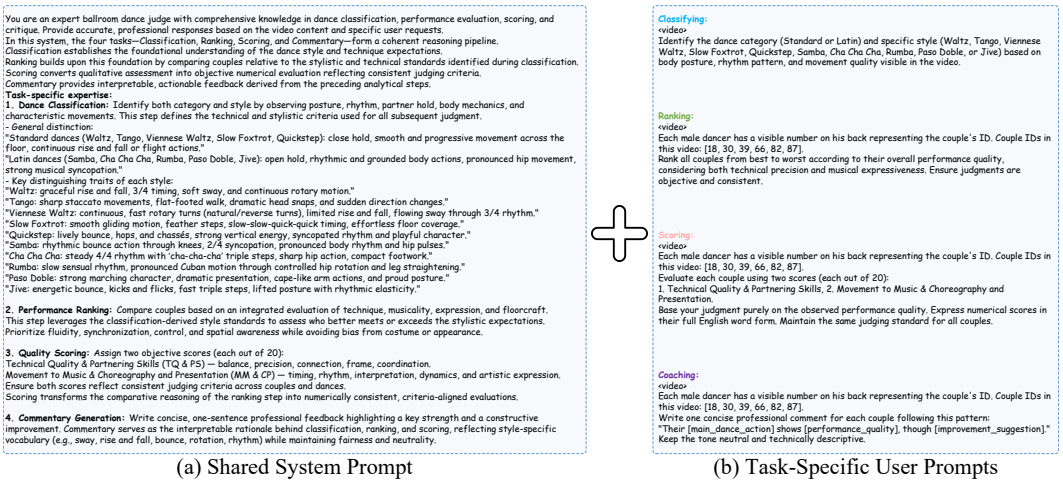
$$\rho = \frac{\sum_i (r_i - \bar{r})(\hat{r}_i - \bar{\hat{r}})}{\sqrt{\sum_i (r_i - \bar{r})^2}\sqrt{\sum_i (\hat{r}_i - \bar{\hat{r}})^2}}, \tag{7}$$

where $r_i = \text{rank}(s_i)$ and $\hat{r}_i = \text{rank}(\hat{s}_i)$ are the ranks of the ground-truth and predicted scores, respectively, and $\bar{r}$ and $\bar{\hat{r}}$ denote their means. Tied scores are assigned average ranks. The Relative L2-distance (R-ℓ_2) quantifies the normalized regression error:

$$\text{R-}\ell_2 = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{s_n - \hat{s}_n}{s_{\max} - s_{\min}}\right)^2, \tag{8}$$

where $s_{\max}$ and $s_{\min}$ denote the maximum and minimum scores in the current dataset, respectively. For coaching evaluation, BERTScore computes the similarity between the candidate text $\hat{x}$ and reference text x using contextual embeddings. It is composed of precision (P_{BERT}), recall (R_{BERT}), and the F1 measure (F_{BERT}):

$$R_{\text{BERT}} = \frac{1}{|x|}\sum_{x_i \in x}\max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j, P_{\text{BERT}} = \frac{1}{|\hat{x}|}\sum_{\hat{x}_j \in \hat{x}}\max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j, \tag{9}$$



(a) Shared System Prompt (b) Task-Specific User Prompts

Figure 9: Examples of (a) shared system prompts and (b) specific user prompts.

$$F_{\text{BERT}} = 2 \cdot \frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}, \tag{10}$$

where $\mathbf{x}_i$ and $\hat{\mathbf{x}}_j$ are the pre-normalized vectors for tokens in the reference and candidate sentences, respectively. In our experiments, we report the F1 score (F_{BERT}) as the primary metric.

D.2 Prompt Design

To connect task-specific outputs with shared assessment criteria, we use a system prompt and task-specific user prompts, as shown in Fig. 9. The system prompt defines the scene, role, and domain knowledge. User prompts specify each task and its output format. Preceding predictions are appended as textual context according to Eq. 3.

Task-Specific Text Tokens. For classification, the model outputs "[coarse-grained category], [fine-grained style]". For ranking, it generates a best-to-worst sequence of couple IDs. For scoring, it provides each couple's ID with all their scores. For coaching, it provides feedback on each couple's highlights and areas for improvement, following the template: "Their [main_dance_action] shows [performance_quality], though [improvement_suggestion]." This requires the model to not only identify key actions but also to assess their quality and pinpoint areas for refinement. We employ a fixed template to standardize quantitative evaluation while allowing performance-specific descriptions.

Domain Knowledge. We use general dance rules and technical criteria as shared prompt context, following knowledge-guided assessment (Xu et al., 2025a; Li et al., 2026; Xu et al., 2024). This context is common to all samples and does not contain clip-specific labels or reference commentary.

D.3 LLM-Based Metric Design

To assess commentary beyond lexical similarity, we use GPT-SQ, an LLM-based metric following prior work (Li et al., 2026; Chen et al., 2024). We use GPT-5.5-Pro to compare generated and reference commentary for accuracy, relevance, fluency, and informativeness, assigning a score from 0 to 10 that is normalized to $[0, 1]$. Fig. 10 provides the evaluation prompt.

D.4 Implementation Details

To implement coarse-to-fine adaptation, we fine-tune Qwen3-VL-8B (Bai et al., 2025a) with two-stage LoRA (Hu et al., 2022). In the 1st stage, we use a 512-frame budget and a nominal sampling rate of 5 fps, with a learning rate of 2e-5 for 1 epoch. In the 2nd stage, the frame budget is 128 and the nominal rate is 15 fps, with a learning rate of 1e-4 for 3 epochs. The effective sampling density depends on the video duration and frame limit, as described below. All experiments are conducted

on a single A800 GPU with a batch size of 16, and other parameters follow the defaults from (Bai et al., 2025a).

Sequential Conditioning. For MT videos, inference follows classification, ranking, scoring, and coaching. Ranking receives the predicted category and style. Scoring receives these predictions and the predicted ranking. Coaching receives all preceding predictions, including the target-specific scores. The video remains available at each step, and the same model generates every output. For ST and EgoExo-Fitness videos, the ranking step is omitted. Predicted outputs are serialized as textual history with their task names and performer IDs.

During training, preceding task outputs are selected from ground-truth annotations or model predictions with a fixed probability of 0.5 for each source. The selected outputs form $H_{train}^{(k)}$, which is supplied as context for predicting the ground-truth answer of task k. This exposes the model to both accurate and imperfect task context. At inference, $H_{pred}^{(k)}$ consists entirely of model-generated outputs. No ground-truth task output is supplied at test time. Sequential conditioning creates explicit dependencies. Appendix E.4 examines cross-task agreement together with task performance.

D.5 EgoExo-Fitness Evaluation Protocol

Dataset and Task Adaptation. To evaluate multi-task adaptation beyond dance, we use EgoExo-Fitness (Li et al., 2024b), which provides synchronized egocentric and exocentric fitness videos from 40 participants across 12 action categories. Its action labels, quality scores on a 1-5 scale, and human commentary support classification, scoring, and coaching. This three-task protocol differs from the original guidance-based execution verification benchmark.

Input and Data Partition. We use 1,525 third-person videos with human coaching commentary, action labels, and quality scores. Single action clips are obtained from the provided boundaries and released 30-fps frames. We randomly split them into 1,145 training and 380 test videos, approximately following a 3:1 ratio, while ensuring that the two sets come from disjoint source clips. All methods use the same test clips and frame budget.

Training and Comparison. Table 3 compares zero-shot MLLMs with single-stage multi-task fine-tuning and MultiAC. The fine-tuned models use the same Qwen3-VL-8B initialization, training clips, task supervision, and total training updates. Ranking is omitted because each clip contains one performer. All methods receive the same fitness domain knowledge without using the ground-truth category to select guidance. At test time, task history contains only model predictions.

Evaluation Metrics. Classification uses top-1 Acc over the 12 action categories. For scoring, we follow the dataset's score aggregation and average the available human scores for each action clip. Predictions remain on the same scale from 1 to 5. We evaluate AQA performance using Spearman's ρ and R-ℓ_2, defined in Eqs. 7 and 8, respectively. For coaching, we retain all available human comments as references. We evaluate coaching performance using BERTScore (Eq. 10) and GPT-SQ. GPT-SQ uses the semantic quality criteria in Appendix D.3, with fitness terminology and all available reference comments supplied together. The evaluator model and prompt are fixed across methods. The resulting scores are reported on the same $[0, 1]$ scale used for MultiDance.

Scope. This experiment evaluates adaptation to fitness actions with existing human annotations. Each clip contains one performer, complementing the competitive multi-target evaluation on MultiDance.

E Expanded Experimental Results

To examine the components of MultiAC, we compare backbones and training settings and evaluate cross-task conditioning, temporal sampling, audio integration, and tracked visual inputs.

E.1 Impact of Different MLLM Backbones

To examine whether MultiAC generalizes across MLLM backbones, we evaluate Video-LLaVA-7B (Lin et al., 2024), VideoChat2-7B (Li et al., 2024a), Qwen2.5-VL-7B (Bai et al., 2025b),

Table 10: Impact of different MLLM backbones on MultiDance. Acc_{coarse} and Acc_{fine} denote the accuracy for coarse-grained category and fine-grained style classification. All MLLMs are fine-tuned and tested with the same prompts.

Methods	Classification		Ranking	Scoring - TQ&PS		Scoring - MM&CP		Coaching	
	$\text{Acc}_{coarse}\uparrow$	$\text{Acc}_{fine}\uparrow$	Acc↑	$\rho\uparrow$	R-$\ell_2\downarrow$	$\rho\uparrow$	R-$\ell_2\downarrow$	BERT↑	GPT-SQ↑
Multi-Target Subset (Test on 228 Samples)									
Video-LLaVA-7B	0.912	0.408	0.428	0.435	0.289	0.452	0.265	0.712	0.345
VideoChat2-7B	0.930	0.434	0.456	0.462	0.264	0.478	0.241	0.734	0.368
Qwen2.5-VL-7B	0.956	0.461	0.475	0.480	0.248	0.495	0.223	0.748	0.382
InternVL3-8B	0.965	**0.478**	0.483	0.475	0.251	0.502	0.215	0.751	0.389
Qwen3-VL-8B	**0.974**	**0.478**	**0.491**	**0.491**	**0.235**	**0.514**	**0.207**	**0.757**	**0.396**
Single-Target Subset (Test on 86 Samples)									
Video-LLaVA-7B	0.884	0.360	-	0.568	0.275	0.582	0.254	0.695	0.372
VideoChat2-7B	0.907	0.384	-	0.592	0.252	0.605	0.231	0.708	0.395
Qwen2.5-VL-7B	0.930	0.395	-	0.615	0.238	0.628	0.215	0.719	0.412
InternVL3-8B	**0.953**	0.407	-	0.621	0.225	0.632	0.208	0.725	0.418
Qwen3-VL-8B	0.942	**0.419**	-	**0.637**	**0.219**	**0.645**	**0.201**	**0.728**	**0.427**
Complete MultiDance Dataset (Average of Tests on 228 and 86 Samples)									
Video-LLaVA-7B	0.933	0.430	0.457	0.573	0.242	0.601	0.228	0.777	0.368
VideoChat2-7B	0.949	0.487	0.488	0.602	0.217	0.598	0.234	**0.801**	0.423
Qwen2.5-VL-7B	0.965	0.484	0.495	0.612	0.220	0.635	0.211	0.781	0.438
InternVL3-8B	0.981	0.538	0.514	0.618	**0.204**	0.643	0.200	0.795	0.464
Qwen3-VL-8B	**1.000**	**0.554**	**0.518**	**0.638**	0.209	**0.651**	**0.198**	0.796	**0.475**

InternVL3-8B (Zhu et al., 2025), and Qwen3-VL-8B (Bai et al., 2025a) under the same adaptation framework.

Table 10 shows improvements over the corresponding zero-shot baselines in Table 2 across all backbones, suggesting that task-aligned adaptation benefits multiple model families. Shared domain criteria and task supervision may help translate general video understanding into target-specific judgments. Gains on quantitative tasks also suggest that MultiAC learns assessment cues beyond the annotation model's linguistic style. Qwen3-VL-8B achieves the highest scoring ρ for both dimensions across the MT, ST, and complete-dataset settings, supporting its use as our default backbone.

The complete-dataset results use different training protocols. For the zero-shot general MLLMs in Table 2, we aggregate the MT and ST metrics by test sample count:

$$m_{\text{all}} = \frac{228\, m_{\text{MT}} + 86\, m_{\text{ST}}}{228 + 86}, \tag{11}$$

where m_{MT} and m_{ST} are the metrics on 228 MT and 86 ST test videos. Classification, scoring, and coaching results are rounded to three decimal places. Ranking uses only the MT subset. For ρ, this is an average of subset correlations, not a correlation recomputed from pooled predictions.

All backbones in Table 10 are fine-tuned. Their complete-dataset rows report evaluation of models jointly trained on MT and ST data. MultiAC follows this protocol in both tables, so its results are not obtained by averaging separately trained models.

E.2 Ablation on Domain Knowledge Components

To isolate the contribution of each domain knowledge component, we remove style-level knowledge, technique-level knowledge, reasoning-chain knowledge, or the task scene definition, and also evaluate complete removal of domain knowledge.

Style-level knowledge specifies dance categories and style characteristics. Technique-level knowledge describes judging criteria such as fluidity, synchronization, and body mechanics. Reasoning-chain knowledge describes task relations in the system prompt. Removing it preserves prediction conditioning, which is separately ablated in Appendix E.4. The task scene definition specifies the expert role and competitive setting.

Table 11 shows a broad decline when domain knowledge is removed, with different components affecting different tasks. Removing style knowledge most affects classification and coaching, while removing technique knowledge most affects ranking and scoring, reducing ρ from 0.503 to 0.445.

Table 11: Ablation on Domain Knowledge Components. Scoring metrics are averaged over the two score types. Acc_c and Acc_f are accuracy for coarse- and fine-grained classification.

Methods	Classification		Ranking	Scoring		Coaching	
	$\text{Acc}_c\uparrow$	$\text{Acc}_f\uparrow$	Acc.↑	$\rho\uparrow$	R-$\ell_2\downarrow$	BERT↑	GPT-SQ↑
Complete Prompt (Ours)	**0.974**	**0.478**	**0.491**	**0.503**	**0.221**	**0.757**	**0.396**
w/o Style-Level Knowledge	0.925	0.404	0.462	0.465	0.268	0.740	0.368
w/o Technique-Level Knowledge	0.969	0.469	0.442	0.445	0.288	0.746	0.378
w/o Reasoning-Chain Knowledge	0.961	0.456	0.468	0.472	0.262	0.750	0.382
w/o Task Scene Definition	0.965	0.447	0.474	0.479	0.256	0.751	0.384
w/o Domain knowledge	0.917	0.386	0.427	0.428	0.303	0.735	0.360

Table 12: Ablation on Hyperparameters of MultiAC. Scoring metrics are averaged over the two score types. Acc_c and Acc_f are accuracy for coarse- and fine-grained classification.

Hyperparameters	Classification		Ranking	Scoring		Coaching	
	$\text{Acc}_c\uparrow$	$\text{Acc}_f\uparrow$	Acc.↑	$\rho\uparrow$	R-$\ell_2\downarrow$	BERT↑	GPT-SQ↑
Temporal Sampling (Stage 1 & Stage 2)							
3 fps & 10 fps	0.952	0.452	0.472	0.485	0.238	0.742	0.380
5 fps & 15 fps (Default)	**0.974**	**0.478**	**0.491**	**0.503**	**0.221**	**0.757**	**0.396**
8 fps & 20 fps	0.969	0.465	0.485	0.492	0.228	0.751	0.388
LoRA Rank							
Rank = 8	0.961	0.461	0.478	0.488	0.232	0.748	0.385
Rank = 16 (Default)	**0.974**	**0.478**	**0.491**	**0.503**	**0.221**	**0.757**	**0.396**
Rank = 32	0.969	0.469	0.489	0.498	0.225	0.754	0.392
Learning Rate (Stage 1 & Stage 2)							
1e-5 & 5e-5	0.956	0.456	0.475	0.482	0.240	0.745	0.382
2e-5 & 1e-4 (Default)	**0.974**	**0.478**	**0.491**	**0.503**	**0.221**	**0.757**	**0.396**
5e-5 & 2e-4	0.965	0.465	0.482	0.490	0.230	0.750	0.389

Style descriptions may help establish what to look for and how to describe it, while technical criteria provide a basis for judging execution quality. The larger decline without any domain knowledge suggests that these components offer complementary guidance for recognition and assessment.

E.3 Ablation on Hyperparameters of MultiAC

To further understand the sensitivity of our two-stage adaptation strategy, we additionally conduct a set of lightweight hyperparameter ablations that stay strictly within the MultiAC framework. We vary the temporal sampling rates, LoRA ranks, and learning rates used in the two stages while keeping all other settings unchanged. Specifically, we compare the default 5 fps/15 fps configuration with variants using 3 fps and 8 fps in the long-context stage, as well as 10 fps and 20 fps in the fine-grained stage, to examine how the temporal density affects multi-target reasoning. We also experiment with reducing the LoRA rank from 16 to 8 and increasing it to 32 to evaluate how parameter capacity influences the model's ability to learn subtle action cues. Finally, we test learning rates of 1e-5, 2e-5, 5e-5 in Stage 1 and 5e-5, 1e-4, 2e-4 in Stage 2 to assess training stability under different adaptation strengths. These ablations jointly provide a clearer picture of how temporal sampling density, adaptation capacity, and update magnitude impact MultiAC, while keeping the analysis tightly aligned with our original two-stage design without introducing additional architectural components. The detailed results are presented in 12.

E.4 Sequential Conditioning and Cross-Task Consistency

To assess the contribution of preceding predictions, we compare independent querying with sequential conditioning on the 228 MT test videos. Both settings use the same trained checkpoint, sampled video frames, system prompt, task questions, and decoding configuration. Independent querying omits preceding task outputs. Sequential conditioning supplies all preceding predictions as described in Appendix D.4. This comparison isolates the inference-time contribution of prediction

Table 13: Comparison of inference protocols on the MT subset. Scoring metrics are averaged over both dimensions. τ_b measures ranking-score agreement and C is the pairwise contradiction rate. Classification uses the same first-task output in both settings.

Protocol	**Classification**		**Ranking**	**Scoring**		**Coaching**		**Consistency**	
	$\mathrm{Acc}_c \uparrow$	$\mathrm{Acc}_f \uparrow$	Acc↑	$\rho \uparrow$	R-$\ell_2 \downarrow$	BERT↑	GPT-SQ↑	$\tau_b \uparrow$	$C \downarrow$
Independent querying	0.974	0.478	0.465	0.469	0.254	0.750	0.371	0.547	0.235
Sequential conditioning (Ours)	**0.974**	**0.478**	**0.491**	**0.503**	**0.221**	**0.757**	**0.396**	**0.643**	**0.180**

Table 14: Stage 2 sampling comparison under a 128-frame budget. The bold row is the reported MultiAC setting. Scoring metrics are averaged over both dimensions.

Sampling strategy	**Classification**		**Ranking**	**Scoring**		**Coaching**	
	$\mathrm{Acc}_c \uparrow$	$\mathrm{Acc}_f \uparrow$	Acc↑	$\rho \uparrow$	R-$\ell_2 \downarrow$	BERT↑	GPT-SQ↑
Uniform full-video (Ours)	**0.974**	**0.478**	**0.491**	**0.503**	**0.221**	**0.757**	**0.396**
Center window	0.952	0.452	0.470	0.476	0.248	0.742	0.372
Random window	0.965	0.469	0.483	0.489	0.230	0.750	0.388

history. It complements the existing task-relation prompt ablation, which retains prediction history while removing relational descriptions from the system prompt.

In addition to classification, ranking, scoring, and coaching metrics, we measure agreement between predicted rankings and predicted total scores. For each video v, let r_i be the predicted rank of performer i, with smaller values indicating better performance, and let s_i be the sum of the predicted TQ&PS and MM&CP scores. We compute Kendall's τ_b between r_i and $-s_i$, using tie correction. We also measure the pairwise contradiction rate:

$$C_v = \frac{\sum_{i<j} \mathbb{I}[(r_i - r_j)(s_i - s_j) > 0]}{\sum_{i<j} \mathbb{I}[r_i \neq r_j,\ s_i \neq s_j]}. \tag{12}$$

Metrics are computed within each video and then averaged. We assess agreement alongside task accuracy because mutually consistent predictions can still be incorrect.

Table 13 shows that sequential conditioning improves downstream task performance and ranking-score agreement. For example, scoring ρ rises from 0.469 to 0.503 and the contradiction rate falls from 0.235 to 0.180. Earlier predictions may serve as shared reference points for later judgments, reducing incompatible interpretations of the same performance. Classification remains unchanged because it receives no preceding predictions. These results suggest that explicit history helps connect the tasks beyond their shared model parameters, while the remaining contradictions show that this connection is still imperfect.

E.5 Ablation on Temporal Sampling Strategies

To examine how temporal coverage affects Stage 2 adaptation, we compare uniform full-video sampling with center and random windows in Table 14.

We initialize all variants from the same Stage 1 checkpoint and vary only Stage 2 temporal selection. Uniform sampling retains 128 frames across the full video. Center-window sampling selects 128 frames at 15 fps from a centered interval, while random-window sampling chooses a valid start position uniformly. No variant uses annotations for temporal selection. All variants share the training settings and frame budget.

At test time, every checkpoint uses the same fixed full-video sampling policy to isolate training-time selection. The uniform full-video row repeats the MultiAC MT baseline.

Uniform full-video sampling performs best across all metrics, followed by random and center windows. Scoring ρ, for example, reaches 0.503 with full-video coverage and 0.489 with random windows. Random windows may broaden temporal exposure during training, while full-video sampling preserves observations across performance phases and changing target visibility. Under the shared

Table 15: Audio feature ablation on the MT subset. The bold row denotes the main RGB-only MultiAC baseline. All settings use textual prompts.

Input setting	Scoring-TQ&PS		Scoring-MM&CP		Coaching	
	$\rho \uparrow$	R-$\ell_2 \downarrow$	$\rho \uparrow$	R-$\ell_2 \downarrow$	BERT↑	GPT-SQ↑
RGB only (MultiAC)	0.491	0.235	0.514	0.207	0.757	0.396
RGB+"null audio tokens"	0.494	0.232	0.519	0.205	0.757	0.398
RGB+"shifted audio"	0.497	0.230	0.524	0.201	0.758	0.400
RGB+"aligned audio"	0.498	0.226	0.548	0.190	0.763	0.409

full-video test policy, the results suggest that global coverage and train-test sampling consistency deserve attention alongside local sampling density. This supports interpreting coarse-to-fine adaptation through task objectives and supervision.

E.6 Audio Feature Integration

Setting. To assess the contribution of audio within our compute budget, we extend Qwen3-VL-8B with pre-extracted AST features (Gong et al., 2021). Our main VL setting supports broad comparison with mainstream vision-language MLLMs under shared RGB and text inputs.

Audio Integration. Audio segmentation and sampling settings follow prior work (Zeng & Zheng, 2024). A trainable two-layer projection maps timestamped features to the language backbone's embedding dimension and inserts them into the prompt as continuous embeddings:

$$F_a = \mathcal{P}_\phi(\mathcal{F}_{AST}(A)) \in \mathbb{R}^{L_a \times d_{llm}}, \qquad \hat{Y}^{(k)} = \mathcal{F}_{llm}\left(F_v, F_a, F_t^{(k)}\right), \tag{13}$$

where A is the audio track, $L_a = 128$ is the audio token count, and d_{llm} is the embedding dimension. We initialize from MultiAC, freeze AST and the base backbone, and train the projection and LoRA parameters with the existing CE objective and history mixing.

Comparison. The null-audio control uses zero-valued features with the same architecture and training budget as the aligned-audio model. The shifted-audio setting evaluates the aligned checkpoint with temporally shifted features and unchanged timestamp positions. Table 15 reports both score dimensions and coaching metrics, with the original RGB-only MultiAC as a reference.

Table 15 shows that aligned audio improves scoring and coaching, with a larger gain for the music-related MM&CP dimension. Its ρ rises from 0.514 to 0.548, suggesting that audio supplies complementary cues that are difficult to recover from movement alone. Aligned audio also outperforms null and shifted audio, consistent with a benefit from temporal correspondence beyond the additional input tokens. These trends suggest that retaining synchronized evidence may be useful for music-related assessment and feedback, while direct movement-to-music coordination warrants dedicated evaluation.

E.7 Tracked Visual Inputs

To examine how target-specific inputs complement shared competitive context, we compare three MultiAC input settings on the 228 MT test videos: full video, tracked crops, and full video with target-level features.

We use SAM-Track (Cheng et al., 2023) to obtain couple-specific tracks. The crop-only setting concatenates the cropped videos along time and replaces the full-video input, retaining the same Qwen3-VL-8B backbone and aggregate frame and visual-token budgets. Concatenation increases the sequence duration, so the shared sampling budget provides fewer observations per target. The feature-augmented setting retains the full video and inserts projected target-level visual features into the prompt, following the embedding integration in Appendix E.6. Tracking and feature extraction add preprocessing cost, and the feature-augmented setting adds prompt tokens.

Table 16 shows that crop-only inputs reduce all metrics. Concatenating tracks spreads the fixed frame budget across a longer sequence, while cropping limits access to shared spatial cues. Some

Table 16: Tracked-input comparison on the MT subset. The bold row repeats the reported MultiAC baseline. Scoring metrics are averaged over both dimensions.

Input setting	Classification		Ranking	Scoring		Coaching	
	$\mathrm{Acc}_c \uparrow$	$\mathrm{Acc}_f \uparrow$	Acc↑	$\rho \uparrow$	R-$\ell_2 \downarrow$	BERT↑	GPT-SQ↑
Full video (MultiAC)	**0.974**	**0.478**	**0.491**	**0.503**	**0.221**	**0.757**	**0.396**
Tracked crops only	0.921	0.434	0.457	0.466	0.239	0.741	0.383
Full video + target features	0.978	0.487	0.502	0.515	0.224	0.763	0.407

You are an expert evaluator specializing in assessing the semantic quality of generated dance commentary. Your task is to compare a generated comment with a reference human comment and assign an overall semantic quality score (GPT-SQ) on a scale of 0 to 10.

Evaluation Dimensions
You should judge the generated comment based on the following four dimensions, weighted equally unless otherwise implied:
1. Accuracy
- Whether the comment describes technically correct dance actions.
- Whether movements, terminology, or stylistic elements are correctly referenced.
- Penalize hallucinations, fabricated movements, or incorrect dance terminology.
2. Relevance
- Whether the comment focuses on the same movement aspects and evaluation intentions as the reference.
- Whether the generated comment aligns with the core feedback theme of the reference comment.
- Irrelevant details or topic drift should reduce the score.
3. Fluency
- Grammatical correctness and natural language quality.
- Smooth readability befitting professional dance evaluation.
4. Information Quality
- Whether the generated comment captures sufficient detail compared to the reference.
- Does it convey stylistically meaningful and technically interpretable feedback?
- Penalize vague, generic, or under-specified descriptions.

Scoring Guidelines (0-10)
- 9-10: Semantically highly aligned with the reference; accurate, relevant, fluent, and informative with only minor omissions.
- 7-8: Mostly aligned; minor inaccuracies or missing details but still professional and meaningful.
- 5-6: Partial alignment; noticeable inaccuracies, vague phrasing, or missing key aspects.
- 3-4: Low alignment; contains major errors, irrelevant content, or significant lack of detail.
- 0-2: Completely misaligned or includes serious hallucinations or stylistic mistakes.

Your Task
Given the reference human comment and the generated comment, provide:
1. A brief justification (2-4 sentences) describing strengths and weaknesses of the generated comment relative to the reference.
2. A single numerical GPT-SQ score (0-10) based on the criteria above.

Input Format
Reference Comment:
‹REF_COMMENT›
Generated Comment:
‹GEN_COMMENT›

Output Format (must follow exactly):
Justification: ‹your analysis›
GPT-SQ Score: ‹0-10›

Prompt for GPT-SQ Metric

Figure 10: Prompt design for the GPT-SQ metric, used to evaluate the semantic quality of generated commentaries.



Classifying
GT: Standard, Viennese Waltz.
Ours: Standard, Waltz.

Ranking
GT: 66, 28, 59, 41, 25, 24.
Ours: 28, 25, 66, 59, 41, 24.

Scoring

	ID	24	25	28	41	59	66
GT:	TQ&PS	17.000, 17.000, 18.084, 17.334, 17.250, 19.250					
	MM&CP	17.500, 17.916, 18.416, 17.666, 17.916, 19.334					
Ours:	TQ&PS	17.167, 19.166, 19.000, 17.250, 18.000, 18.416					
	MM&CP	17.750, 19.250, 19.334, 18.084, 18.416, 18.916					

Coaching
GT: 24: Their footwork precision in quick-change sequences shows adequate control, though improved weight transfer timing would elevate the smoothness of their overall flow.
25: Their momentum through progressive figures shows dynamic energy, though a steadier upper body carriage during fast rotations would support better frame stability.
28: Their continuous movement across the floor shows confident timing and partnership, though sharper edge definition in change steps would improve their directional clarity.
41: Their coordination in synchronized turns shows reliable connection, though greater uniformity in body alignment during transitions would refine their spatial presentation.
59: Their musical phrasing during rotational sequences shows expressive intent, though maintaining consistent rise and fall throughout the phrase would strengthen technical integrity.
66: Their rotation through natural and reverse turns shows exceptional fluidity and balance, though a more consistent sway control would enhance the vertical line of their frame.
Ours: 24: Their waltz shows exceptional technical precision and seamless partnering, though their musical interpretation could benefit from more dynamic variation.
25: Their waltz shows solid technical execution and balanced partnering, though their musical expression could be more emotionally engaging.
28: Their waltz shows consistent technical quality and good partnering, though their choreography lacks the complexity to elevate their presentation.
41: Their waltz shows strong technical control and expressive partnering, though their choreography could be more innovative to enhance visual impact.
59: Their waltz shows competent technical execution and clear partnering, though maintaining consistent rise and fall throughout dynamic sections would strengthen continuity.
66: Their waltz shows adequate technical foundation and coordinated partnering, though their movement to music lacks the fluidity needed for higher-level expression.

Figure 11: Failure case. When the video contains several performer groups with highly similar visual appearances, the model's fine-grained identity discrimination degrades markedly. Correct predictions are color-coded.

targets may therefore be insufficiently observed, limiting comparative assessment. Adding target features to the full video improves most metrics, with ρ reaching 0.515. This suggests that local detail can complement shared context, motivating target-aware representations that retain the full scene. The accompanying increase in R-ℓ_2 to 0.224 also indicates that better relative ordering need not yield more accurate score magnitudes.

F More Qualitative Results

To complement the quantitative evaluation, we examine additional assessment examples and a representative failure case.

F.1 Failure Case Analysis

To identify limitations in target-specific assessment, we examine a crowded scene in Fig. 11. Three of the six couples wear similar outfits, with black suits for male dancers and red dresses for female dancers. Under trajectory crossings and camera motion, subtle cues such as back numbers become difficult to distinguish. The mismatched target-level outputs suggest that identity ambiguity can propagate into rankings, scores, and commentary. Recognizing technical execution may therefore

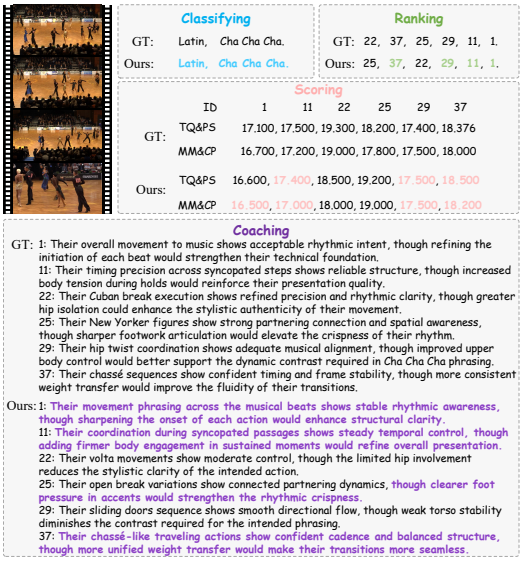


Figure 12: Qualitative results of the generated multi-task outputs. Our MultiAC understands action attributes from coarse to fine, forming a reasoning chain of perception, comparison, evaluation, and coaching. Correct predictions are colored.

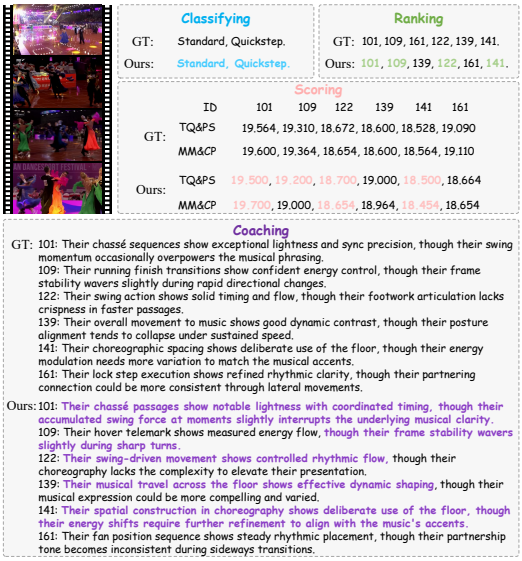


Figure 13: Qualitative results of the generated multi-task outputs. Our MultiAC understands action attributes from coarse to fine, forming a reasoning chain of perception, comparison, evaluation, and coaching. Correct predictions are colored.

be insufficient when its evidence is attributed to the wrong couple. This case points to persistent identity cues as a useful complement to action understanding.

F.2 Additional Case Studies

To illustrate how preceding assessments inform coaching, we present additional outputs in Figs. 12 and 13. The examples connect predicted categories, rankings, and scores with commentary on technical strengths and areas for improvement. Viewing these outputs together may help readers inspect whether the feedback supports a couple's predicted assessment. This suggests a practical benefit of multi-task outputs for interpreting judgments, with their usefulness depending on accurate target attribution and descriptions grounded in visible actions.

G Discussions and Limitations

G.1 Discussions

To clarify the scope of our evaluation, we consider how assessment knowledge may transfer across target configurations and domains. Joint training on the ST and MT subsets of MultiDance improves performance across tasks. One possible explanation is that ST videos provide less ambiguous views of individual execution, while MT videos supply competitive comparisons and shared-scene interactions. Their combination may help connect target-specific quality cues with comparative judgment. The qualitative surfing case further suggests potential beyond dance and motivates quantitative evaluation across competitive domains.

We also evaluate classification, scoring, and coaching on the public EgoExo-Fitness dataset (Li et al., 2024b). MultiAC improves over single-stage multi-task fine-tuning in Table 3 under the protocol in Appendix D.5. These gains suggest that organizing supervision from scene recognition to detailed assessment may remain useful when competitive ranking is absent. The transferable aspect of the framework may thus lie partly in connecting tasks at different granularities. EgoExo-Fitness broadens the assessment domains with existing human annotations, while MultiDance evaluates competitive interactions among concurrent targets.

Other single-target dance datasets offer complementary research opportunities. Imperial-Dance (Zhong & Demiris, 2024) provides motion-capture skeleton sequences, music, and expertise-level labels, which differ from our RGB-based competitive scoring protocol. EgoExo4D (Dance) (Grauman et al., 2025) investigates learning across egocentric and exocentric views in casual dance settings, with different objectives from criteria-driven competitive assessment. Extending our framework to these representations and evaluation protocols remains a promising direction for future work.

G.2 Limitations and Future Work

To clarify the limits of the current evaluation, we distinguish official score prediction from direct movement-to-music assessment. Our main MultiAC baseline uses RGB video and textual context. The audio extension in Appendix E.6 improves MM&CP prediction, while this combined score also reflects choreography and presentation. Dedicated synchronization annotations and evaluation are needed to determine how precisely the model assesses movement-to-music coordination. Musicality-related commentary from the RGB-only baseline remains limited by the absence of audio.

Sequential conditioning can propagate errors from earlier tasks into later scores and commentary. Mixing ground-truth and predicted history during training exposes the model to imperfect context. Appendix E.4 shows improved agreement alongside assessment quality, but some contradictions remain. This suggests that prediction history provides a useful yet incomplete constraint on later judgments. Future work could examine confidence-aware conditioning or explicit consistency checks to retain informative context while limiting the influence of unreliable predictions.

Visually similar targets remain difficult to distinguish under occlusion and overlapping trajectories, as shown in Appendix F. The feature integration results in Appendix E.7 suggest potential benefits from combining target-level detail with full-scene context. The remaining errors motivate more persistent identity cues and careful allocation of the visual budget across targets. Studying these factors separately could clarify when tracking supports assessment and when missed observations or identity switches limit its value.

We also plan to extend MultiDance to training sessions, a wider range of skill levels, and additional dance styles to study assessment beyond the current competition setting.