ExAct: A Video-Language Benchmark for Expert Action Analysis

Han Yi Yulu Pan Feihong He Xinyu Liu Benjamin Zhang
Oluwatumininu Oguntola Gedas Bertasius
University of North Carolina at Chapel Hill
{hanyi, yulupan, gedas}@cs.unc.edu, {feihonhe, xinyu1, zhangben, oguntola}@unc.edu

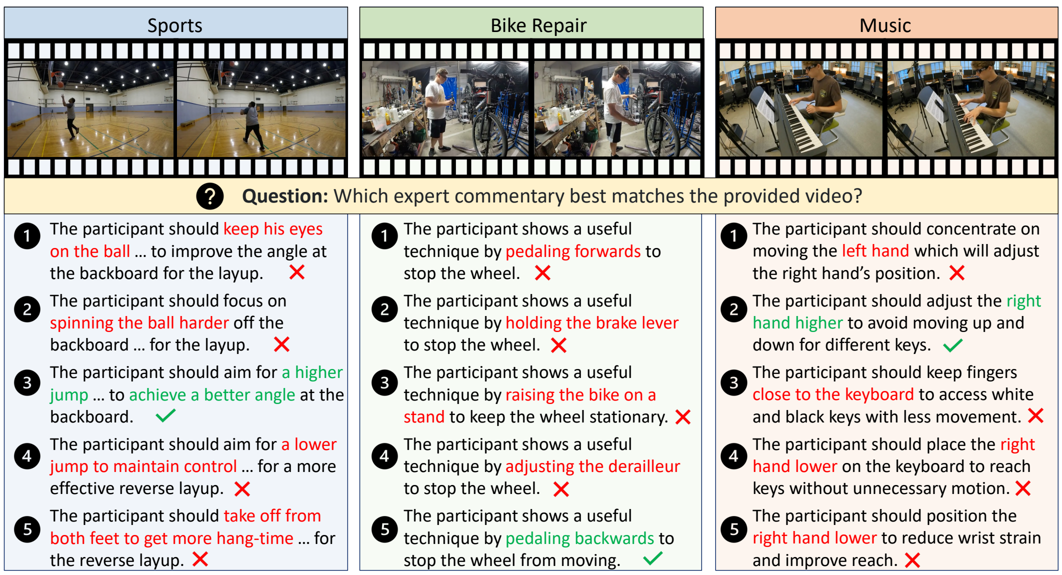

Figure 1: An illustration of several multiple-choice question samples from our expert action analysis benchmark, ExAct. Here, we visualize samples from three domains of skilled activities (i.e., basketball, bike repair, and piano). The correct answers have a green checkmark next to them and were obtained using domain-expert/coach annotations. The phrases in green (correct) and red (incorrect) emphasize the subtle yet critical differences between ground-truth expert descriptions and incorrect candidate answers.

Abstract

We present ExAct, a new video-language benchmark for expert-level understanding of skilled physical human activities. Our new benchmark contains 3,521 expert-curated video question-answer pairs spanning 11 physical activities in 6 domains: Sports, Bike Repair, Cooking, Health, Music, and Dance. ExAct requires the correct answer to be selected from five carefully designed candidate options, thus necessitating a nuanced, fine-grained, expert-level understanding of physical human skills. Evaluating the recent state-of-the-art VLMs on ExAct reveals a substantial performance gap relative to human expert performance. Specifically, the best-performing Gemini 2.5 Pro model achieves only 55.35% accuracy, well below the 82.02% attained by trained human experts. We believe that ExAct will be beneficial for developing and evaluating VLMs capable of precise understanding of human skills in various physical and procedural domains. Dataset and code are available at https://texaser.github.io/exact_project_page/.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

https://doi.org/10.52202/085713-0246

1 Introduction

Today, learning and perfecting a new physical skill requires a significant amount of time, practice, and often guidance from an expert coach/professional in that domain. Unfortunately, personalized coaching remains inaccessible to the majority due to prohibitively expensive costs and a lack of expert availability. Recent advances in AI have sparked interest in developing virtual AI assistants/-coaches, particularly in the text domain [35, 2, 46]. However, text-based large language models (e.g., ChatGPT) are insufficient for learning new *physical skills* as standard LLMs typically lack a nuanced understanding of physical human activities/skills. In contrast, rapidly improving vision-language models (VLMs) could serve as valuable tools for various physical skill learning applications. In particular, by recording a video of a skill demonstration and feeding it to a VLM, people could receive detailed, actionable feedback similar to that of expert human coaches.

Although recent progress has enabled modern VLMs to achieve impressive general image/video recognition capabilities [20, 46, 29, 3, 27, 9], existing studies reveal that such VLMs still struggle to understand fine-grained human activities [6, 36], particularly activities that require expert-level knowledge [5, 19, 26]. This is primarily due to the inadequate underlying visual representations that 1) do not capture expert-level knowledge needed to generate feedback for physical skill learning, and 2) are unable to recognize subtle details in skilled human actions.

In addition to the limitations of existing VLMs, there is a notable lack of evaluation benchmarks tailored for expert-level understanding of skilled human activities. As shown in Table 1, most existing datasets focus on coarse activity recognition [8, 33, 45, 21, 34, 43, 51], which typically only require scene-level recognition rather than fine-grained understanding. While several fine-grained video recognition datasets exist [15, 42, 25, 6], the majority of them are aimed at generic (*e.g.*, putting something into something) rather than expert-level action understanding of skilled human activities (*e.g.*, keeping balance during a dance spin, playing the correct rhythm in a piano piece). Beyond action recognition, a number of video-based skill assessment benchmarks have been recently developed [1, 12, 50, 4, 36, 13]. Most of such skill assessment datasets focus on predicting scalar/categorical performance scores to quantify execution quality. Additionally, these datasets

Dataset	Expert-level Knowledge	Free-form Language Annotations	MCQ Evaluation
Coarse Action Recognition Datasets			
Kinetics-700 [8]	✗	✗	✗
HowTo100M [33]	✗	✗	✗
UCF101 [45]	✗	✗	✗
HMDB [21]	✗	✗	✗
Moments in Time [34]	✗	✗	✗
Hollywood [43]	✗	✓	✗
ActivityNet-QA [51]	✗	✗	✓
Fine-grained Action Recognition Datasets			
Something-SomethingV2 [15]	✗	✗	✗
FineGym [42]	✗	✗	✗
Multisports [25]	✗	✗	✗
TemporalBench [6]	✗	✓	✓
Video-Based Skill Assessment Datasets			
JIGSAWS [1]	✓	✗	✗
Best [12]	✓	✗	✗
FineDiving [50]	✓	✗	✗
FP-Basket [4]	✓	✗	✗
BASKET [36]	✓	✗	✗
Aifit [13]	✓	✓	✗
Skilled Activity Video-Language Datasets			
VidDiffBench [5]	✓	✓	✗
EgoExo-Fitness [26]	✓	✓	✗
EgoExolearn [19]	✓	✓	✗
Ego-Exo4D [16]	✓	✓	✗
EXACT (Ours)	✓	✓	✓

Table 1: Compared to previous action recognition and skill assessment datasets, our proposed EXACT benchmark uniquely combines expert-level, free-form language annotations and a multiple-choice question (MCQ) evaluation format, making it an excellent resource for evaluating modern video-language models at expert-level understanding of skilled human activities.

https://doi.org/10.52202/085713-0246

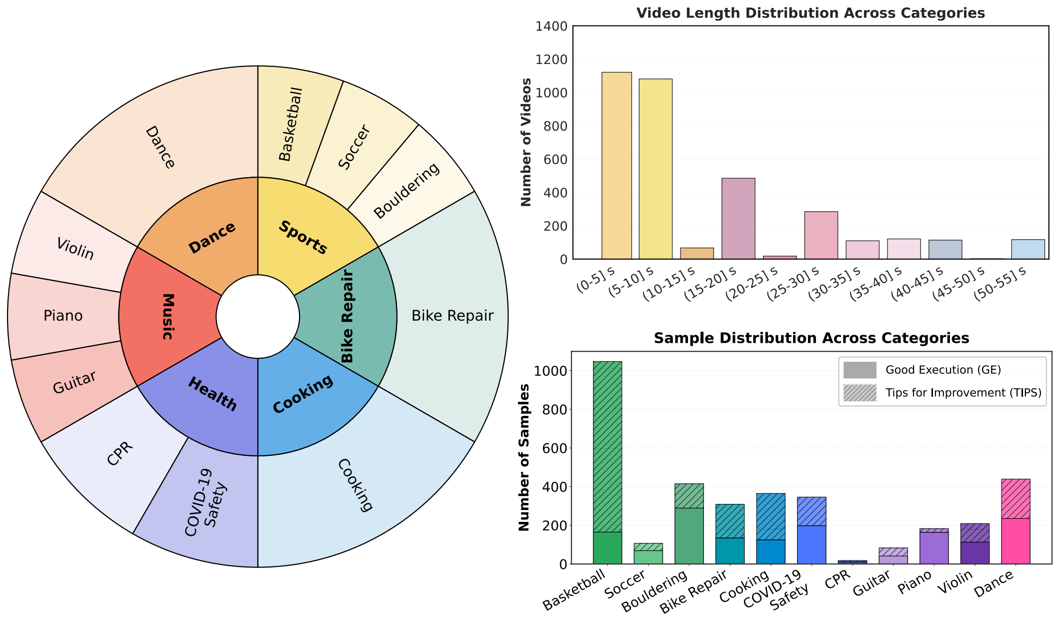

Figure 2: **Left:** Our proposed ExAct benchmark contains 11 skilled activity types spanning 6 broader physical domains: Sports, Music, Dance, Health, Cooking, and Bike Repair. **Top Right:** Distribution of video lengths across the dataset, showing that most clips fall within the 0–10 second range. **Bottom Right:** Sample distribution per activity, categorized by the expert feedback type: Good Execution (GE) and Tips for Improvement (TIPS).

often lack open-ended language annotations, which can provide a rich and intuitive medium to convey skill-specific feedback [5] and can capture subtle errors, temporal nuances, and intent—details that scalar or categorical labels often miss. Finally, a few recent skilled activity video-language datasets [5, 26, 19, 16] use experts to obtain free-form language descriptions akin to verbal coach feedback but do not provide rigorous evaluation benchmarks or tasks, making it difficult to assess how well modern video models can understand nuanced physical human skills.

To address these issues, we introduce ExAct, a video-language benchmark consisting of 3,521 video–question–answer (VQA) pairs designed to evaluate expert-level understanding of skilled physical human actions (sample questions shown in Figure 1). ExAct covers 11 activities across 6 diverse physical domains: Sports (Basketball, Soccer, Bouldering), Bike Repair, Cooking, Health (COVID-19 Safety, CPR), Music (Guitar, Piano, Violin), and Dance, as shown in Figure 2. To construct ExAct, we build on the fine-grained expert commentaries derived from the Ego-Exo4D dataset [16]. The original Ego-Exo4D expert commentaries have several crucial limitations: 1) the original expert commentaries are unstructured and lengthy, 2) they contain automatic speech recognition (ASR) errors, and 3) they include redundant or irrelevant content that is not directly tied to the observed actions. Furthermore, evaluating the quality of open-ended captions/descriptions in the original form of Ego-Exo4D expert commentaries is inherently difficult using standard language metrics such as CIDEr [47], BLEU [37], or ROUGE [28], which do not accurately reflect the correctness or relevance of the instructional feedback in the context of skilled physical activities. To address these challenges, we first apply a structured annotation pipeline that rewrites the original commentaries into concise, self-contained feedback commentaries. Then, our proposed ExAct benchmark evaluates expert action understanding as a multiple-choice question-answering task, which uses a well-defined metric of question-answering accuracy. This multiple-choice question-answering format eliminates the ambiguity and reproducibility issues associated with open-ended evaluation by providing clearly defined answers and controllable question difficulty through the design of distractor options.

We evaluate several state-of-the-art VLMs, including Gemini 2.5 Pro [11], GPT-4.1 [20], GPT-4o [20], Gemini 1.5 Pro [46], LLaVA-Video [55], LLaVA-OneVision [22], Qwen2.5-VL [3], VideoLLaMA [53], InternVL2.5 [9], and PerceptionLM [10] on our new ExAct benchmark. Our results reveal that compared to human experts, most modern VLMs achieve poor results on ExAct. In particular, Gemini 2.5 Pro, the best-performing model in our experiments, achieves 55.35% accuracy. In comparison, non-expert humans achieve 61.86% accuracy, while domain experts achieve 82.02% accuracy. This substantial gap highlights the limitations of current VLMs in expert-level understanding of physical human skills. We hope that ExAct will serve as a rigorous evaluation benchmark for

https://doi.org/10.52202/085713-0246

measuring expert-level understanding of skilled human actions, thus laying the foundation for AI systems that support enhanced human skill learning.

2 Related Work

Vision-Language Models (VLMs). Recent advances in Vision-Language Models (VLMs) have demonstrated impressive capabilities in understanding visual content. Models such as Gemini 2.5 Pro [11], GPT-4.1 [20], GPT-4o [20], Gemini 1.5 [46], LLaVA-OneVision [22], Qwen2.5-VL [3], VideoLLaMA [53], and InternVL2.5 [9] have achieved strong performance in tasks such as action recognition, video captioning, and visual question answering. Although these models show remarkable generalization, their outputs are often limited to high-level descriptions of the image/video content, rather than a detailed, fine-grained understanding of physical human actions and skills. More recent video-centric variants, such as LLaVA-Video [55] and VideoLLaMA [53], attempt to extend static image capabilities to temporal inputs. However, these models still struggle to understand the fine-grained nuances of complex physical human activities, which are essential for human skill understanding. Specifically, most existing VLMs cannot assess how well a task is performed, nor articulate the strengths of the execution and the areas requiring improvement [5, 6, 36]. To address this gap, our proposed EXACT provides a rigorous benchmark to evaluate expert-level understanding of physical human skills.

Multimodal and VQA Benchmarks. Diverse multimodal benchmarks have emerged to evaluate VLM performance, particularly in the video domain. ActivityNet-QA [51] assesses the temporal reasoning of activities within longer videos via question-answering. NExT-QA [49] focuses on compositional temporal reasoning. STAR [48] emphasizes situated reasoning. Benchmarks such as PerceptionTest [41] probe the perceptual abilities of VLMs in various modalities. TemporalBench [6] targets fine-grained temporal understanding. MV-Bench [24] assesses temporal comprehension across 20 challenging tasks. EgoSchema [31] focuses on egocentric human actions. MMWorld [17] aims to evaluate embodied agents in simulated environments. MLVU [57], Video-MMMU [18], MMVU [56], Video-MMLU [44], and Video-MME [14] aim to evaluate modern MLLMs for complex video question answering tasks. Finally, SEED-Bench [23] and MMBench [30] evaluate various multimodal abilities of MLLMs. However, most existing benchmarks focus on factual recall (e.g., "What color is the car?"), event-level understanding (e.g., "What action is being performed?"), or basic temporal reasoning (e.g., "What happened before this?"), and are not designed to capture the subtle nuances of skilled human actions. In contrast, our newly proposed EXACT benchmark focuses explicitly on expert-level analysis of physical human skills.

Video-based Skill Assessment Benchmarks. Several recent works have focused on developing methods to assess human skills from video. Benchmarks such as MITDive [40], UNLV-Dive [39], MTL-AQA [38], and FineDiving [50] provide temporally segmented videos with action labels or action quality scoring for diving. FP-Basket [4] and BASKET [36]) focus on basketball, while LOGO [54]) provides human judgment scores for artistic swimming. Similar efforts also exist in other sports, including figure skating [40] and golf [32]. Furthermore, datasets such as JIGSAWS [1], BEST [12], and EgoExoLearn [19] focus on scenarios beyond sports, such as surgical tasks and daily activities. Most recently, the Ego-Exo4D [16] dataset introduces large-scale egocentric and exocentric video of skilled human activities. Ego-Exo4D includes spoken expert commentaries, offering a unique expert-level supervisory signal for understanding human skills. However, these expert commentaries are typically highly unstructured, noisy due to ASR errors, and often contain irrelevant information. Moreover, Ego-Exo4D does not provide a formal evaluation benchmark/task associated with such expert commentaries. In our work, we leverage such unstructured expert commentaries and construct a rigorous, expert-validated, and easy-to-evaluate EXACT benchmark enabling evaluation of expert-level understanding of physical human actions/skills.

3 EXACT Benchmark Construction

We construct EXACT using a four-stage pipeline. In stage I, we pre-process raw expert commentaries using GPT-4o, correcting ASR errors and segmenting them into concise, self-contained feedback commentaries. In Stage II, we construct multiple-choice QA pairs, each consisting of one correct commentary and four distractors. In Stage III, we remove low-quality or biased samples through length filtering and blind-LLMs. Finally, in Stage IV, domain experts review each QA pair to ensure

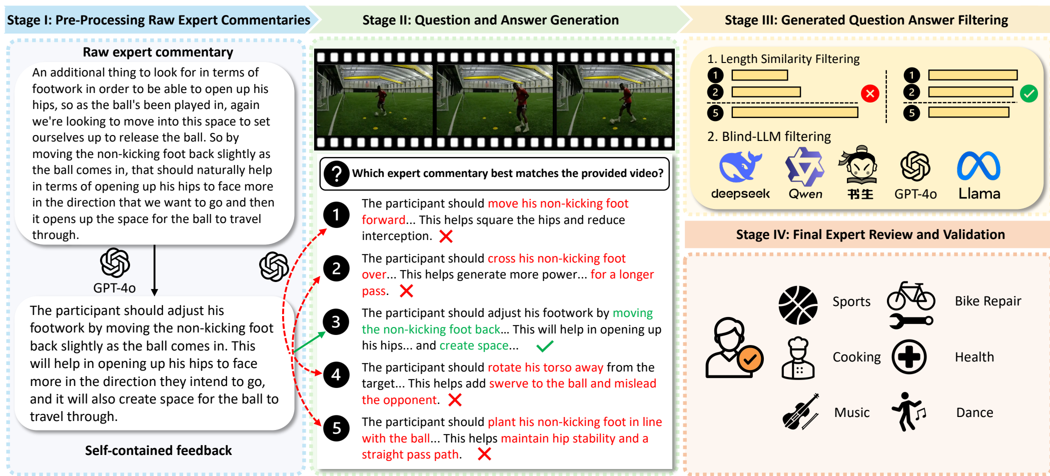

Figure 3: Overview of our benchmark construction pipeline. In stage I, we pre-process raw expert commentaries using GPT-4o, correcting errors and segmenting them into concise, self-contained feedback commentaries. In Stage II, we construct multiple-choice QA pairs, each consisting of one correct expert commentary and four carefully generated distractors. The four red arrows indicate the LLM-generated distractors, while the green arrow represents the correct expert commentary. In Stage III, we filter out low-quality or biased samples using length-based heuristics and blind-LLMs. Finally, in Stage IV, domain experts review all QA pairs to ensure visual grounding and linguistic accuracy.

visual grounding and linguistic accuracy. Our benchmark construction pipeline is illustrated in Figure 3.

3.1 Stage I: Pre-Processing Raw Expert Commentaries

The transcribed commentaries from Ego-Exo4D are often lengthy, noisy, and unstructured. They may also include automatic speech recognition (ASR) errors, redundant phrases, and off-topic remarks, as illustrated in the upper-left example of Figure 3. In the first stage of our benchmark construction pipeline, we use an LLM to refine raw expert commentaries from Ego-Exo4D into a more compact and less noisy format. Specifically, to do this, we prompt GPT-4o to 1) correct transcription mistakes, 2) remove irrelevant or repetitive content, and 3) segment the cleaned text into concise, self-contained feedback commentaries. Each commentary is then also assigned to one of the two categories: *good execution* or *tips for improvement*, reflecting the two main ways in which experts deliver their feedback, i.e., by affirming what was done well and/or suggesting what can be improved. We then use such compact and structured commentaries to construct multiple-choice question-answer pairs as described next. The complete prompt templates are provided in the supplementary material.

3.2 Stage II: Question and Answer Generation

After pre-processing raw expert commentaries in Stage I, we proceed with multiple-choice QA pair construction. Specifically, we treat the structured commentary from Stage I as a positive commentary and ask the LLM (e.g., GPT-4o) to generate four distractor commentaries that require fine-grained expert-level understanding to distinguish the correct answer from the incorrect ones. This leads to a 5-way multiple-choice QA setup. To create negative commentaries, we use the following strategies:

For expert commentaries assigned to the *good execution* category (see Subsection 3.1), we apply two strategies to generate negative commentaries:

- **Action replacement:** A key action in the original commentary is substituted with a plausible but incorrect alternative (*e.g.*, replacing "*... performs a three-point shot*" with "*... performs a layup*").
- **Absent-action insertion:** A new event or action is inserted that was never mentioned or shown (*e.g.*, adding "*... tightens the brake lever ...*" to "*... keeps the bike steady on a repair stand ...*").

For commentaries in the *tips for improvement* category, we apply four alternative strategies:

https://doi.org/10.52202/085713-0246

- **Action misinterpretation:** Misinterpreting the mistake in an execution. *Example:* Replacing "*... elbow is too low ...*" with "*... grip is too tight ...*".
- **Incorrect technical reasoning:** Correctly identifying a flaw but providing an implausible or technically inaccurate explanation. *Example:* "*... doesn't bend knees ... reduces jump height*" vs. "*... doesn't bend knees ... prevents ball spin*" (i.e., knee movement affects jumping, not ball spin).
- **False cause–effect relationship:** Introducing a misleading causal link between the error and an unrelated factor. *Example:* "*... feet are misaligned, leading to an off-balance shot*" vs. "*... doesn't tuck in jersey, leading to an off-balance shot*".
- **Ineffective suggestion:** Proposing a correction to the execution that does not address the problem. *Example:* "*... doesn't keep eyes on the ball ...*" vs. "*... should stand closer to the baseline*".

Albeit simple in format, our questions effectively probe multi-faceted reasoning abilities—spanning temporal, causal, spatial, and domain-specific understanding—beyond mere visual recognition (see supplementary material S4 for examples).

3.3 Stage III: Generated Question Answer Filtering

After constructing initial QA samples (Subsection 3.2), we apply several additional filtering steps to ensure high-quality and unbiased samples. Prior studies [7, 52] have shown that large language models (LLMs) can often exploit subtle statistical or stylistic patterns in multiple-choice answers to correctly identify the ground-truth option without relying on visual input. Such language-driven shortcuts, often referred to as language-related bias, pose a serious threat to the integrity of robust evaluation. To mitigate such biases, we adopt two filtering strategies:

1) **Length similarity:** Similar to human test-takers, LLMs may exploit surface-level cues such as answer length to eliminate implausible distractors. To mitigate this bias, we enforce a length similarity constraint: each distractor must be between 80% and 120% the length of the positive (i.e., correct) expert commentary, and the absolute word count difference must not exceed 8 words. QA samples that violate this constraint are excluded from the benchmark. Note that the 80–120% length range and 8-word absolute difference threshold were determined empirically through iterative pilot studies with human annotators.

2) **Blind-LLM filtering:** Inspired by findings from TemporalBench [6], we observe that LLMs can sometimes identify the correct answer by detecting shared linguistic patterns, particularly when distractors are lightly edited variants of the ground truth. To avoid such language-driven biases, we present each QA sample consisting solely of the five textual options without any video to 5 state-of-the-art LLMs (GPT-4o and DeepSeek-VL-R1 (proprietary), Qwen2.5-72B, LLaMA3.3-70B, and InternLM2.5-20B (open-source, all ranked highly on multiple LLM leaderboards)). Each model is prompted with the question: *"Which expert commentary best matches the provided video?"* If more than 20% of the models (i.e., exceeding the random chance of selecting the correct answer in a 5-way setup) select the positive expert commentary, we consider the sample susceptible to language-only bias and remove it from the dataset.

3.4 Stage IV: Final Expert Review and Validation

In the final stage, we conduct a two-phase manual review to ensure that each QA question is clearly formulated and can be reliably answered through expert-level video analysis. Unlike the previous stages, which rely solely on textual inputs, this phase gives domain experts full access to the video alongside the five answer options. This setup allows for a comprehensive multimodal evaluation of answer correctness and distractor plausibility.

First, for each QA sample, domain experts meticulously review the corresponding video clip and all five answer options. Their task is to select the option most consistent with the visual information presented in the video. After submitting their answer, the ground-truth label is revealed. The expert is then asked to verify three criteria: (1) whether the visual content of the video clip supports the correctness of the positive (i.e., correct) expert commentary; (2) whether any of the distractor commentaries also describe actions that are visible or valid within the video; and (3) whether all five candidate answer options are free from grammatical, logical, or instructional flaws. The sample will be removed if any of these criteria are not met. The second criterion is especially important because if a distractor candidate also describes something that appears in the video, the question becomes

https://doi.org/10.52202/085713-0246

ill-posed due to multiple correct answers. For example, suppose that the positive expert commentary states, "*The participant plants their non-kicking foot beside the ball*," while a distractor commentary states, "*The participant plants their foot behind the ball*." In this case, both candidate commentaries are visually observable, which means that the expert cannot answer it using a single answer. We remove all such samples to avoid ambiguity.

This verification process is conducted by 16 experts, each assigned to one of the 11 activity categories based on their area of expertise. Each sample is reviewed by at least one qualified expert to ensure consistency and domain relevance. For domains reviewed by two experts, we observed that experts reached agreement in over 90% of cases—demonstrating strong consistency in their judgments. A screenshot of the annotation interface, along with additional implementation details, is provided in the supplementary material.

4 Experimental Setup

Evaluation Metrics. We use standard question-answering accuracy as the primary evaluation metric. Our benchmark includes a total of 3,521 QA samples spanning 11 fine-grained physical skilled activities. For each activity, we compute the accuracy as the percentage of questions for which the model selects the correct answer. To summarize performance across the dataset, we report the average accuracy across all 11 activities. We additionally report per-domain accuracy (Sports, Bike Repair, Cooking, Health, Music, and Dance) to highlight domain-specific generalization.

Baseline Models. To thoroughly assess the challenges posed by EXACT, we evaluate various state-of-the-art Video-Language Models (VLMs), including proprietary models: Gemini 2.5 Pro [11], GPT-4.1 [20], GPT-4o [20], Gemini 1.5 Pro [46], and open-source models: LLaVA-Video [55], LLaVA-OneVision [22], Qwen2.5-VL [3], VideoLLaMA [53], InternVL2.5 [9], and PerceptionLM [10]. These models vary significantly in architecture, training corpus, and modality integration strategies, offering a broad and representative basis to evaluate expert-level feedback capabilities.

Implementation Details. All model inferences are conducted using 4 NVIDIA Tesla H100 GPUs, each with 96 GB of memory. For fairness, we adopt uniformly sampling strategy and extract 32 frames per video clip for all models. Each frame is extracted at a resolution of 796×448 and then resized internally according to the input resolution requirements of each model. Unless otherwise specified, we use the same prompt template and follow the official inference code provided by each model. We also ablate on several key hyperparameters in Subsection 5.4. Additional implementation details, including full prompt formulations, are provided in the supplementary material.

5 Experimental Results

5.1 Main Results

In Table 2, we report the performance of several state-of-the-art video-language models (VLMs) on our EXACT benchmark. Overall, none of the methods achieve accuracy exceeding 56%. Among the evaluated models, Gemini 2.5 Pro achieves the highest overall accuracy at 55.35%. This model outperforms all others across all six domains. Notably, it achieves over 60% accuracy in the domains of Bike Repair and Health. These domains tend to rely more heavily on procedural knowledge, such as understanding sequences of actions and planning the next steps. In contrast, domains like Sports, Cooking, Music, and Dance involve highly specialized, fine-grained physical movements that require detailed visual perception and precise temporal understanding.

We also observe that proprietary models (e.g., GPT and Gemini) consistently outperform the best-performing open-source alternatives. This performance gap highlights the limitations of current public models in capturing the fine-grained, domain-specific reasoning required for expert-level skill understanding. Additionally, model size appears to be a significant factor in fine-grained video comprehension. Smaller models such as PerceptionLM-8B [10] and VideoLLaMA3-7B [53] achieve only 24.65% and 26.38% accuracy respectively—barely above random chance. In comparison, larger open-source models reach at least 33% accuracy, suggesting that scale contributes to more effective representation and complex activity analysis.

 https://doi.org/10.52202/085713-0246

Model	Overall (%)	Results by Domain (%)					
		Sports	Bike Repair	Cooking	Health	Music	Dance
Random Choice	20.00	20.00	20.00	20.00	20.00	20.00	20.00
Human Non-Expert	61.86	62.97	55.02	66.58	71.43	54.11	59.22
Human Expert	82.02	82.09	81.23	80.27	87.09	80.21	81.55
Open-source VLMs							
PerceptionLM-8B [10]	24.65	24.22	28.16	25.75	22.53	22.95	26.42
VideoLLaMA3-7B [53]	26.38	26.64	23.30	29.32	26.65	23.79	27.79
InternVL2.5-78B [9]	33.48	31.93	36.57	33.70	37.91	32.00	34.62
LLaVA-OneVision-72B [22]	35.44	33.65	43.04	33.42	35.44	30.53	43.51
Qwen2.5-VL-72B-Instruct [3]	35.67	35.62	37.86	33.97	36.26	32.63	38.50
LLaVA-Video-72B [55]	41.58	41.81	42.72	44.11	32.42	38.74	48.52
Proprietary VLMs							
Gemini 1.5 Pro [46]	43.91	42.83	52.10	51.78	41.21	41.89	39.86
GPT-4o [20]	44.70	43.47	52.75	46.30	53.30	33.89	46.70
GPT-4.1 [20]	50.89	51.37	58.90	54.25	51.10	40.84	51.48
Gemini 2.5 Pro [11]	**55.35**	**52.58**	**65.05**	**58.36**	**60.71**	**53.05**	**53.98**

Table 2: ExAct evaluation results (QA accuracy) across six diverse physical domains: Sports (Basketball, Soccer, Bouldering), Bike Repair, Cooking, Health (COVID-19 safety, CPR), Music (Guitar, Piano, Violin), and Dance. The results show a significant gap between the performance of modern Vision-Language Models (VLMs) and human experts, indicating that there is a significant room for improvement for future video-language models.

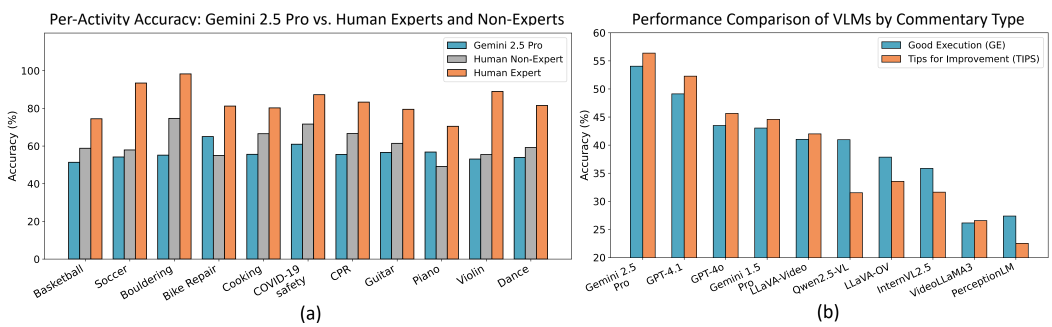

Figure 4: (a) Accuracy of the best-performing model (Gemini 2.5 Pro) compared to human experts and non-experts across domains. While Gemini 2.5 Pro achieves the highest accuracy among VLMs, it still falls short of human performance: experts consistently achieve over 80% accuracy in most domains, with soccer and bouldering reaching up to 95%. (b) Performance of VLMs on Good Execution (GE) and Tips for Improvement (TIPS) commentary categories. Models such as PerceptionLM, InternVL2.5, LLaVA-OneVision, and Qwen2.5-VL perform better on GE, while all proprietary models show stronger performance on TIPS.

5.2 Human Performance Analysis

To quantify the gap between human and model performance, we conduct a human evaluation involving two groups: **experts**, comprising trained coaches and professionals with significant domain expertise (i.e., $> 10\ years$), and **non-experts**, individuals without professional expertise in a given domain. As shown in Figure 4 (a), experts achieve consistently high accuracy—often exceeding 80% across most domains, with soccer and bouldering reaching around 95%. Additionally, we observe that while non-experts perform considerably worse (60-70% accuracy), they still surpass the best-performing VLM model (i.e., Gemini 2.5 Pro). These results highlight a substantial gap between human and VLM capabilities, emphasizing the challenges of ExAct and the limitations of current VLMs.

https://doi.org/10.52202/085713-0246

Num. Frames	Acc. (%)
8	39.22
16	39.24
32	**41.58**
64	39.73

(a) **Number of Frames:** Using 32 input video frames leads to the best accuracy.

Model Size	Acc. (%)
7B	27.07
72B	**41.58**

(b) **Impact of LLM size:** Using larger LLMs leads to significantly better performance.

Prompting	Acc. (%)
w/o Time Info	40.04
w/ Time Info	**41.58**

(c) **Prompting strategies:** Incorporating time information into the prompt improves accuracy.

Table 3: Ablation study on different design choices: number of input frames, impact of LLM size, different prompt strategies.

5.3 Performance by Expert Commentary Type

Figure 4 (b) shows the performance of various VLMs across two expert commentary categories: Good Execution (GE) and Tips for Improvement (TIPS). Models such as PerceptionLM [10], InternVL2.5 [9], LLaVA-OneVision [22], and Qwen2.5-VL [3] perform better on GE samples. However, they struggle with questions from the TIPS category, which require identifying subtle mistakes in skilled activity executions. In contrast, stronger models such as Gemini 2.5 Pro [11] and GPT-4.1 [20] perform better on questions from the TIPS category, indicating a greater capacity for expert-level understanding of errors/mistakes in physical human activities.

5.4 Ablation Studies

In Table 3, we conduct ablation studies on four key design choices: (a) number of input frames, (b) impact of LLM size, (c) different prompt strategies, and (d) spatial video resolution. For (a)–(c), we use LLaVA-Video [55], the strongest-performing open-source baseline. We ablate the effect of spatial video resolution using Qwen2.5-VL, which supports native-resolution inputs.

Number of Input Frames. Table 3a shows the effect of varying the number of input frames using a uniform sampling strategy. We observe a consistent improvement in performance as the number of frames increases, with the model achieving the highest accuracy at 32 frames. This suggests that incorporating more temporal context enhances the model's ability to understand skilled human actions. We also observe that increasing the number of frames beyond 32 does not yield further gains, which may be because most VLMs are not optimized to process longer video sequences effectively.

Impact of LLM Size. In Table 3b, we use the same model architecture while varying the size of LLM in LLaVA-Video: 7B and 72B. As the size of LLM parameters increases, we observe a consistent improvement in accuracy. This trend highlights the critical role of the language model for capturing fine-grained video-language cues necessary for human skill analysis.

Prompting Strategies. In Table 3c, we analyze the impact of different prompting strategies on model performance. We first evaluate the role of including time information (*e.g.* *"The video is {video_time}s long, and {len(frames)} uniformly sampled frames occur at {frame_time}."* in the prompt. Removing the time information leads to a performance degradation of 1.54%, indicating that the understanding of time plays a meaningful role in the model's analysis of human skill.

Spatial Video Resolution. We investigate how different spatial input resolutions affect the performance of Qwen2.5-VL by evaluating three settings: the original resolution of 796×448 (35.67% accuracy), a $1.5\times$ downsampled version at 531×299 (37.40% accuracy), and a $2\times$ downsampled version at 398×224 (35.03% accuracy). Interestingly, the model achieves the highest accuracy at the mid-level resolution of 531×299. We hypothesize that although higher resolutions provide more visual detail, they may lead to performance degradation due to increased token length and a mismatch with the lower-resolution inputs commonly seen during pretraining.

6 Conclusion

We introduce ExAct, a new video-language benchmark designed to evaluate expert-level understanding of skilled human activities across a diverse set of physical and procedural domains. Our new benchmark uses fine-grained, expert-level, language annotations and a multiple-choice evaluation

https://doi.org/10.52202/085713-0246

format to enable a rigorous evaluation of expert-level understanding of physical human skills. Our experiments reveal a significant gap between state-of-the-art VLMs and human experts' performance, indicating a significant room for future improvement in video-language model design. We believe that ExAct will be pivotal in the development and evaluation of video language models capable of skilled human activity understanding.

Limitations. Although our benchmark spans multiple domains, it captures only a fraction of real-world activities. Additional tasks from underrepresented or specialized fields such as surgery or mechanical engineering may elicit different behaviors from current models and offer further insights. Moreover, certain domains (e.g., COVID-related tasks) may be time-sensitive or outdated, potentially affecting the relevance of some samples. Finally, while all participants consent to data usage, the inclusion of real-world videos containing identifiable human faces raises potential privacy concerns. The ethical implications surrounding the reuse and dissemination of such visual data warrant careful consideration and responsible handling.

References

[1] Narges Ahmidi, Lingling Tao, Shahin Sefati, Yixin Gao, Colin Lea, Benjamin Bejar Haro, Luca Zappella, Sanjeev Khudanpur, René Vidal, and Gregory D Hager. A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. *IEEE Transactions on Biomedical Engineering*, 64(9):2025–2041, 2017.

[2] Anthropic. Claude. `https://www.anthropic.com`, 2023. Large language model.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[4] Gedas Bertasius, Hyun Soo Park, Stella X Yu, and Jianbo Shi. Am i a baller? basketball performance assessment from first-person videos. In *Proceedings of the IEEE international conference on computer vision*, pages 2177–2185, 2017.

[5] James Burgess, Xiaohan Wang, Yuhui Zhang, Anita Rau, Alejandro Lozano, Lisa Dunlap, Trevor Darrell, and Serena Yeung-Levy. Video action differencing. *arXiv preprint arXiv:2503.07860*, 2025.

[6] Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. Temporalbench: Towards fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*, 2024.

[7] Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. Matryoshka multimodal models. In *Workshop on Video-Language Models@ NeurIPS 2024*, 2024.

[8] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

[9] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.

[10] Jang Hyun Cho, Andrea Madotto, Effrosyni Mavroudi, Triantafyllos Afouras, Tushar Nagarajan, Muhammad Maaz, Yale Song, Tengyu Ma, Shuming Hu, Suyog Jain, et al. Perceptionlm: Open-access data and models for detailed visual understanding. *arXiv preprint arXiv:2504.13180*, 2025.

[11] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

https://doi.org/10.52202/085713-0246

[12] Hazel Doughty, Walterio Mayol-Cuevas, and Dima Damen. The pros and cons: Rank-aware temporal attention for skill determination in long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[13] Mihai Fieraru, Mihai Zanfir, Silviu Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu. Aifit: Automatic 3d human-interpretable feedback models for fitness training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9919–9928, 2021.

[14] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

[15] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.

[16] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024.

[17] Xuehai He, Weixi Feng, Kaizhi Zheng, Yujie Lu, Wanrong Zhu, Jiachen Li, Yue Fan, Jianfeng Wang, Linjie Li, Zhengyuan Yang, et al. Mmworld: Towards multi-discipline multi-faceted world model evaluation in videos. *arXiv preprint arXiv:2406.08407*, 2024.

[18] Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025.

[19] Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, et al. Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22072–22086, 2024.

[20] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[21] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.

[22] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[23] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

[24] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024.

[25] Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. Multisports: A multi-person video dataset of spatio-temporally localized sports actions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13536–13545, 2021.

 https://doi.org/10.52202/085713-0246

[26] 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 *European Conference on Computer Vision*, pages 363–382. Springer, 2024.

[27] 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. *arXiv preprint arXiv:2311.10122*, 2023.

[28] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[30] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.

[31] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

[32] William McNally, Kanav Vats, Tyler Pinto, Chris Dulhanty, John McPhee, and Alexander Wong. Golfdb: A video database for golf swing sequencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[33] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019.

[34] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019.

[35] OpenAI. Chatgpt (may 2024 version). Large language model developed by OpenAI, 2024. `https://chat.openai.com`.

[36] Yulu Pan, Ce Zhang, and Gedas Bertasius. Basket: A large-scale video dataset for fine-grained skill estimation. *arXiv preprint arXiv:2503.20781*, 2025.

[37] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[38] Paritosh Parmar and Brendan Tran Morris. What and how well you performed? a multitask learning approach to action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 304–313, 2019.

[39] Paritosh Parmar and Brendan Tran Morris. Learning to score olympic events. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 20–28, 2017.

[40] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Assessing the quality of actions. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 556–571. Springer, 2014.

[41] Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=HYEGXFnPoq`.

[42] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2616–2625, 2020.

[43] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer, 2016.

[44] Enxin Song, Wenhao Chai, Weili Xu, Jianwen Xie, Yuxuan Liu, and Gaoang Wang. Video-mmlu: A massive multi-discipline lecture understanding benchmark. *arXiv preprint arXiv:2504.14693*, 2025.

[45] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[46] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[47] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[48] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*, 2024.

[49] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021.

[50] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2949–2958, 2022.

[51] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019.

[52] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024.

[53] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

[54] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2405–2414, 2023.

[55] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

[56] Yilun Zhao, Lujing Xie, Haowei Zhang, Guo Gan, Yitao Long, Zhiyuan Hu, Tongyan Hu, Weiyuan Chen, Chuhan Li, Junyang Song, et al. Mmvu: Measuring expert-level multi-discipline video understanding. *arXiv preprint arXiv:2501.12380*, 2025.

[57] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024.

 https://doi.org/10.52202/085713-0246

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the key contributions of the paper, including "EXACT will be beneficial for developing and evaluating VLMs capable of precise understanding of human skills in various physical and procedural domains." These claims are consistent with the technical content and experimental results presented in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We provide limitations in Section 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

https://doi.org/10.52202/085713-0246

Answer: [NA]

Justification: The paper does not include any theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All necessary details to reproduce the main experimental results are provided in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

https://doi.org/10.52202/085713-0246

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide open access to both the dataset and code used in our experiments. The dataset is available at: `https://huggingface.co/datasets/Alexhimself/ExAct`, and the evaluation code is available at: `https://github.com/Texaser/Exact`. Detailed instructions for reproducing the main experimental results are included in the repository and supplemental material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide the details in Section 4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We do not report error bars. All results come from a single deterministic inference pass with pre-trained models—no model training or other stochastic procedures are involved—so repeated runs would yield identical numbers, while retraining the large models to obtain multiple seeds would be prohibitively expensive.

Guidelines:

- The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the details in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research adheres fully to the NeurIPS Code of Ethics. We ensured responsible use of data, preserved participant anonymity where applicable, and followed ethical practices in the collection, use, and reporting of all results. No known risks, harms, or ethical violations were introduced by this work.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

https://doi.org/10.52202/085713-0246

Justification: The paper discusses both potential positive and negative societal impacts. It highlights the benefits of the proposed benchmark for evaluating expert-level understanding and supporting applications in education and skill training. In the limitations section, we also acknowledge potential negative societal impacts. Although all participants consented to data usage, the inclusion of real-world videos containing identifiable human faces may still pose privacy concerns.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The benchmark does not involve the release of models or datasets with a high risk of misuse. All data sources are curated from publicly available, non-sensitive content with appropriate usage rights, and the benchmark is intended for academic research under a permissive license.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All external assets used in this work—including datasets, models, and code—are properly cited in the paper. Their respective licenses and terms of use (e.g., MIT, CC-BY, or custom academic licenses) have been reviewed and respected. We ensure that all assets are used in accordance with their intended academic and research use permissions.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces a new benchmark, and we provide detailed documentation alongside the released assets. This includes task definitions, annotation guidelines, dataset statistics, data format specifications, and instructions for evaluation and usage. The documentation is provided alongside the released code and dataset at the uploaded URL.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: As the paper involves research with human annotators, we provide the full text of the instructions given to participants, example screenshots of the annotation interface in the supplemental material.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

https://doi.org/10.52202/085713-0246

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The study involves human annotators evaluating non-anonymized video clips. While the videos may contain identifiable individuals, the annotation task itself poses no known risks to participants. All annotators were informed of the task content and compensation before participation. Based on our institutional guidelines, the nature of the task did not require formal IRB approval.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Large language models (LLMs) are a core component of our evaluation methodology. Their usage is clearly documented in the main paper, including the specific models employed (e.g., GPT-4o, Gemini 1.5 Pro) and their roles within the evaluation pipeline.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

https://doi.org/10.52202/085713-0246

ExAct: A Video-Language Benchmark for Expert Action Analysis

Supplementary Material

Our supplementary materials include the following sections: Section S1: Full Prompts for Constructing ExAct, Section S2: Annotation Interface Details, Section S3: Qualitative Results, Section S4: Evaluated Reasoning Dimensions in ExAct, Section S5: Sample statistics after each filtering stage.

S1 Full Prompts for Constructing ExAct

We provide the detailed prompts used for (1) pre-processing raw expert commentaries, (2) generating negative/distractor commentaries for ExAct, and (3) formatting inputs for VLMs.

Prompt for Pre-Processing Raw Expert Commentaries. We present the prompts used for preprocessing raw expert commentaries into Good Execution and Tips for Improvement in Figure 5 and Figure 7, respectively. Corresponding examples of Good Execution and Tips for Improvement are shown in Figure 6 and Figure 8.

You are given a raw, transcribed expert commentary from a dataset describing a participant performing an {activity}. These transcriptions may contain automatic speech recognition (ASR) errors, redundant phrases, off-topic remarks, or unstructured language.
Your primary goal is to identify the good executions mentioned in the commentary and express them in a clean, concise, and coherent manner. Specifically:

1. Correct any transcription or grammatical errors.
2. Remove irrelevant, repetitive, or filler content.
3. Your goal is to determine the good executions mentioned in the commentary and write them in a coherent, concise manner.
4. If there are no good executions, respond with: "The expert mentions no good executions."

Figure 5: A prompt for pre-processing raw expert commentaries from the good execution (GE) category.

Sample 1

Input
A grand great rhythm and flow as he crosses over. This is a great picture of going off the correct foot, which is his right foot, raising that left knee. He's got his eye and head up, total great body extension. He's got in great spatial awareness of where he is in relation to the basket and the rim. He's not too far under, he's not too far out.

Output
The participant has a great rhythm and flow as he crosses over. He goes off of the correct foot, raises the correct knee, and has his head and eyes up. Overall, great body extension and spatial awareness under the rim.

Sample 2

Input
So here the dancer is jumping side to side, almost like a little short gallop. Really nice leg space. I definitely would recommend bending the knees a little bit more. And then just really like, yeah, I think if we bent the knees that would just kind of bring a better level and ease for him to jump side to side. Good job with the arms swaying back and forth on that part and then good eye connection to the audience facing front.

Output
The participant demonstrates nice leg space during the jumps. Additionally, the participant nicely sways their arms and makes good eye contact with the audience.

Figure 6: Samples of converting raw expert commentaries into concise good execution (GE) commentaries.

https://doi.org/10.52202/085713-0246

You are given a raw, transcribed expert commentary from a dataset describing a participant performing an {activity}. These transcriptions may contain automatic speech recognition (ASR) errors, redundant phrases, off-topic remarks, or unstructured language.
Your primary goal is to identify the tips for improvement mentioned in the commentary and express them in a clean, concise, and coherent manner. Specifically:

1. Correct any transcription or grammatical errors.
2. Remove irrelevant, repetitive, or filler content.
3. Your goal is to determine the tips for improvement mentioned in the commentary and write them in a coherent, concise manner.
4. If there are no tips for improvement, respond with: "The expert mentions no tips for improvement."

Figure 7: A prompt for pre-processing raw expert commentaries from the tips for improvement (TIPS) category.

Sample 1

Input
This is another example again of the person taking too heavy of a touch and then overextending his leg to be able to wrap around the ball. So actually you see now how he's contacting the ball. Instead of kind of catching it, he's actually having to almost like, his foot placement is much higher on the center of the ball here. He's actually coming like downward with the force in this direction when he's actually contacting it, which could actually potentially make the ball jump up when he hits it.

Output
The participant should try to take a lighter touch and focus on not overextending their leg to wrap around the ball. The participant should avoid coming downward when contacting the ball because it could make the ball jump up when they hit it.

Sample 2

Input
This is a really excellent shift that the player had. If they wanted to be as expressive, they could change the weight as they're going up to come into the fingerboard a little bit more so it comes across as a more expressive slide.

Output
The participant should change the weight as they go up to come into the fingerboard so it comes across as a more expressive slide.

Figure 8: Samples of converting raw expert commentaries into concise tips for improvement (TIPS) commentaries.

https://doi.org/10.52202/085713-0246

Prompt for Negative Commentary Generation. Here we present the prompts used for negative/distractor commentary generation. For Good Execution (GE) commentaries, we adopt two strategies, as shown in Figure 9. For Tips for Improvement commentaries, we use four strategies, as shown in Figure 10.

You are a bouldering expert tasked with creating **wrong but plausible commentary** to train a video understanding model. You will be given high-expertise bouldering commentary, and your task is to generate **four wrong comments** based on that expert commentary. These wrong comments should be grounded in visible actions from the video and appear reasonable but must either provide an incorrect justification or directly misinterpret the actions.

Requirements:

- All comments must be grounded in observable actions from the video and avoid references to non-visual elements.
- Match the length, detail, and complexity of the expert's original comments without obvious stylistic differences.
- The difference between correct and incorrect comments should lie in the reasoning or specific actions mentioned.
- Do not generate comments that might sometimes be true; ensure the actions are *definitely* incorrect based on expert feedback.
- Avoid using negative adjective words such as "improper," "bad," "not good," or "not perfect."
- Ensure the incorrect comments appear plausible, limiting each to 1–2 subtle errors.
- Vary the type of error across the four generated comments. Keep all comments logical and coherent.

Some techniques for creating wrong comments:

- **Action replacement:** A key action in the original commentary is substituted with a plausible but incorrect alternative.
- **Absent-action insertion:** A new event or action is inserted that was never mentioned or shown.

Output Format:

Good execution: "The participant demonstrates a good initiation of upward movement, properly preparing their legs to generate momentum upwards."

Wrong Comments:

- **Action replacement:** "The participant demonstrates a good initiation of downward movement, correctly preparing their arms to generate momentum downwards."
- **Action replacement:** "The participant properly prepares their arms to generate momentum sideways, effectively aiding their lateral movement along the boulder."
- **Absent-action insertion:** "The participant demonstrates a clever strategy by swinging their body to the side before leaping to the next hold, avoiding direct upward movement."
- **Absent-action insertion:** "The participant initiates a powerful dyno by planting both feet on the foothold and lunging directly for the top of the wall, using agility over controlled movement."

The high-expertise comment is as follows:

Figure 9: Prompt for negative/distractor commentary generation for good execution commentaries. We use a bouldering example for illustration.

https://doi.org/10.52202/085713-0246

You are a cooking expert tasked with creating **wrong but plausible commentary** to train a video understanding model. You will be given high-expertise cooking commentary, and your task is to generate **four wrong comments** based on that expert commentary. These wrong comments should be grounded in visible actions from the video and appear reasonable but must either provide an incorrect justification or directly misinterpret the actions.

Requirements:

- All comments must be grounded in observable actions from the video and avoid references to non-visual elements.
- Match the length, detail, and complexity of the expert's original comments without obvious stylistic differences.
- The difference between correct and incorrect comments should lie in the reasoning or specific actions mentioned.
- Do not generate comments that might sometimes be true; ensure the actions are *definitely* incorrect based on expert feedback.
- Avoid using negative adjective words such as "improper," "bad," "not good," or "not perfect."
- Ensure the incorrect comments appear plausible, limiting each to 1–2 subtle errors.
- Vary the type of error across the four generated comments. Keep all comments logical and coherent.

Some techniques for creating wrong comments:

- **Action misinterpretation:** Misinterpreting the mistake in an execution.
- **Incorrect technical reasoning:** Correctly identifying a flaw but providing an implausible or technically inaccurate explanation.
- **False cause–effect relationship:** Introducing a misleading causal link between the error and an unrelated factor.
- **Ineffective suggestion:** Proposing a correction to the execution that does not address the problem.

Output Format:

Tips for improvement: "The participant should add herbs, spices, and tea while waiting for the mixture to come to a simmer to improve efficiency and flavor infusion."

Wrong Comments:

- **Action misinterpretation:** "The participant should wait until the mixture has finished simmering before adding herbs, spices, and tea, as this prevents any flavors from being cooked out of the ingredients."
- **Incorrect technical reasoning:** "The participant should add herbs, spices, and tea while the mixture is boiling vigorously, as the intense heat heightens the flavor of these ingredients."
- **False cause–effect relationship:** "The participant should add herbs, spices, and tea just after the mixture stops simmering, as this allows the flavors to cool simultaneously with the dish for balanced taste."
- **Ineffective suggestion:** "The participant should blend the herbs, spices, and tea into a smooth paste before adding them to the mixture after it comes to a simmer, ensuring a more uniform flavor throughout."

The high-expertise comment is as follows:

Figure 10: Prompt for negative/distractor commentary generation from tips for improvement commentaries. We use a cooking example for illustration.

VLM Prompt for Processing ExAct. Here we provide the template for the input prompt to LLaVA-Video. The prompts used for other models are very similar, only with a different video separation token. The default image token is the video separation input for LLaVA-Video. The scenario prompt briefly introduces the activity being performed by the participant (*e.g.*, "The participant is practicing basketball."). We also include time-related instructions, such as the total video duration and the timestamps of uniformly sampled frames in the prompt to guide the VLMs. The prompt presents five candidate answer options labeled **Option 1** to **Option 5**, including one correct answer and four distractors.

<DEFAULT_IMAGE_TOKEN>
The video is {video_time} seconds long, and {len(frames)} uniformly sampled frames occur at {frame_time}.
{scenario_prompt}
Below are different feedback statements about the person's performance in this video:
Option 1. Option 2. Option 3. Option 4. Option 5.
Based on what you observe in the video, which expert commentary best matches the provided video?
Just respond with the option number (1–5) and nothing else.

Figure 11: An input prompt to Vision-Language models (VLMs) for processing ExAct.

S2 Annotation Interface Details

In this section, we provide more details related to the annotation interface used for annotating ExAct. We develop a user-friendly web-based platform tailored for human annotators. A total of 16 experts participated in the annotation process, each with at least ten years of experience in their respective domains. The website is built using `github.io`, with `Formspree` used to collect submission data from annotators. All annotators are compensated at a rate of $50 per hour. The interface begins with an introduction to the project, followed by detailed annotation guidelines. It supports saving progress, allowing annotators to complete their assigned tasks over multiple sessions rather than in a single sitting. Each expert is only assigned samples within their domain of expertise. Upon completing their assigned tasks, annotators submit their responses via the form. Figure 12 and Figure 13 shows our data annotation interface.

Guidelines for Annotators. Please follow the instructions below when annotating:

1. Carefully watch the video clip.
2. Read all five options and select the one you believe is correct.
3. Click the "Confirm Selection" button to submit your answer.
4. After submitting your selection, the ground-truth answer will be revealed. Please then evaluate the sample based on the following criteria:
   (a) Does the video clearly support the action described in the ground-truth commentary?
   (b) Do any of the other options also appear valid based on the video?
   (c) Are there any language issues, such as grammatical errors or illogical phrasing, in any of the options?
5. Click "Continue" to move on to the next sample.

 https://doi.org/10.52202/085713-0246

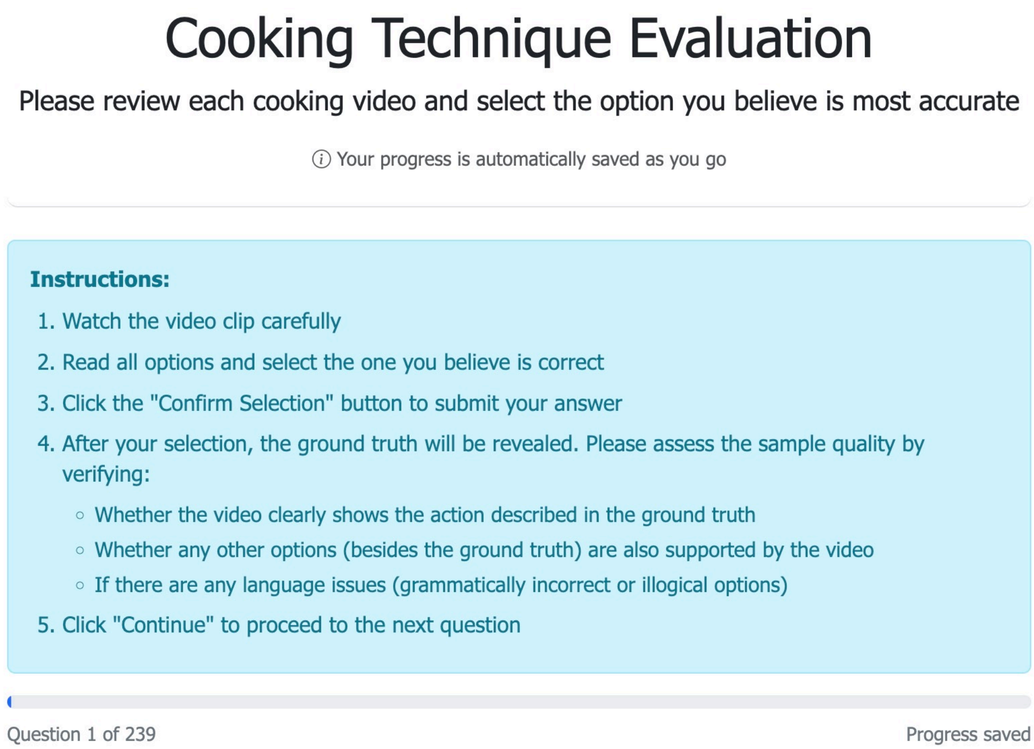

Figure 12: User instructions for the annotation platform interface.

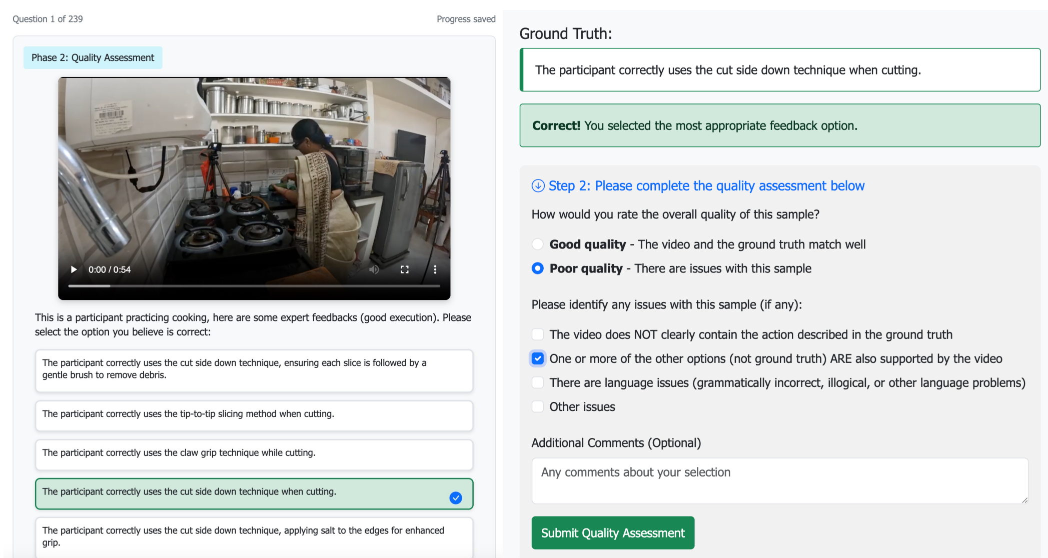

Figure 13: Two-phase manual review interface on the annotation website.

S3 Qualitative Results

In this section, we present four QA examples (Figures 14–17) that range from easy to difficult.

Sample 1 (Figure 14) represents a relatively easy example. All models, as well as both human experts and non-experts, correctly identify the correct answer.

Sample 2 (Figure 15) presents a moderately challenging sample. Only human experts and some of the best models identify the correct answer.

Question: Which expert commentary best matches the provided video?

1. The participant needs to improve on providing enough arc accuracy and rotation on their jump shot to ensure the ball reaches the midpoint between the rims. ✓
2. The player should work on keeping a stiffer wrist during the release to maintain stability, which will ensure the ball travels precisely to the center of the hoop.
3. The player should aim to add more spin to the ball to create a backspin effect, which will assist the ball in reaching the center point of the hoop.
4. The player should concentrate on jumping higher to increase the shot's velocity, which will make the ball accurately land in the midpoint between the rims.
5. The player needs to focus on reducing the arc of their jump shot to increase momentum, which will help the ball reach the midpoint between the rims.

Model response	Human
GPT-4o: Option 1 ✓	Human Expert: Option 1 ✓
Gemini 1.5 Pro: Option 1 ✓	Human Non-Expert: Option 1 ✓
LLaVA-Video: Option 1 ✓	

Figure 14: Sample 1 (Basketball): All models, as well as both human experts and non-experts, select the correct answer.

Sample 3 (Figure 16) presents a more difficult case. Only human experts and GPT-4o select the correct answer.

Sample 4 (Figure 17) illustrates a particularly difficult case. None of the models are able to select the correct answer. Only human experts succeed. This highlights a significant gap between current VLM capabilities and expert-level understanding, especially for tasks that require nuanced, domain-specific reasoning.

These examples collectively highlight the limitations of current VLMs in expert-level understanding of physical human skills.

https://doi.org/10.52202/085713-0246

Question: Which expert commentary best matches the provided video?

1. The participant lowers their shoulder to focus on playing at the tip, improving their ability to execute fast passages.
2. The musician prominently uses their wrist to achieve playing close to the frog, allowing better transition between dynamics.
3. The violinist skillfully extends their upper arm away from the body to maintain control while playing near the frog, helping in balancing the instrument.
4. The participant effectively uses their upper arm, bringing it closer to their body to play near the frog. ✓
5. The violinist utilizes finger strength to move towards the tip of the bow, ensuring a brighter sound while keeping the bow under control.

Model response		Human	
GPT-4o: Option 4	✓	Human Expert: Option 4	✓
Gemini 1.5 Pro: Option 4	✓	Human Non-Expert: Option 3	✗
LLaVA-Video: Option 3	✗		

Figure 15: Sample 2 (Violin): Some models and human experts select the correct answer.

Question: Which expert commentary best matches the provided video?

1. The participant does an excellent job at drying the swab with a cloth to ensure any excess liquid is inside the tube rather than collected on the swab, which is crucial for accurate test results.
2. The participant does an excellent job at shaking the tube to ensure any excess liquid is inside the tube rather than collected on the swab, which is crucial for accurate test results.
3. The participant does an excellent job at pressing the swab to ensure any excess liquid is inside the tube rather than collected on the swab, which is crucial for accurate test results.
4. The participant does an excellent job at swirling the swab in the air to ensure any excess liquid is inside the tube rather than collected on the swab, which is crucial for accurate test results.
5. The participant does an excellent job at squeezing the tube to ensure any excess liquid is inside the tube rather than collected on the swab, which is crucial for accurate test results. ✓

Model response		Human	
GPT-4o: Option 5	✓	Human Expert: Option 5	✓
Gemini 1.5 Pro: Option 1	✗	Human Non-Expert: Option 1	✗
LLaVA-Video: Option 1	✗		

Figure 16: Sample 3 (COVID-19 Safety): Only GPT-4o and human experts select the correct answer.

https://doi.org/10.52202/085713-0246

Question: Which expert commentary best matches the provided video?

1. The participant executes a nice jump to the side with well-bent knees while clapping her hands above her head to maintain rhythm and movement.
2. The participant executes a nice jump to the side with well-bent knees and performs a nice roll with her upper body to maintain rhythm and movement. ✓
3. The participant executes a nice jump to the side with well-bent knees and performs a smooth cartwheel to maintain rhythm and movement.
4. The participant executes a nice jump upwards with locked knees and performs a rigid turn with her upper body to maintain rhythm and movement.
5. The participant executes a nice leap to the front with well-straightened legs and performs a graceful arm sweep to maintain rhythm and movement.

Model response	Human
GPT-4o: Option 1 ✗	Human Expert: Option 2 ✓
Gemini 1.5 Pro: Option 1 ✗	Human Non-Expert: Option 3 ✗
LLaVA-Video: Option 1 ✗	

Figure 17: Sample 4 (Dance): None of the models select the correct answer. Only human experts identify the correct response.

https://doi.org/10.52202/085713-0246

S4 Evaluated Capabilities in EXACT

This section details the reasoning capabilities implicitly evaluated by our QA framework. Although each question follows a simple multiple-choice format, correctly answering them requires diverse capabilities that extend beyond visual recognition. The evaluated reasoning types include:

- **Fine-grained recognition:** Detecting subtle differences in body positioning, movement patterns, and technique execution (e.g., "plays with a very tight wrist" vs. "uses a hybrid picking technique").
- **Temporal reasoning:** Understanding timing precision, action phases, and temporal dependencies (e.g., "releases the ball too early" vs. "too late after the peak jump").
- **Causal understanding:** Linking technique flaws to performance outcomes (e.g., "doesn't bend knees...reduces jump height" vs. "doesn't bend knees...prevents ball spin").
- **Procedural reasoning:** Evaluating adherence to task sequences (e.g., "checks pulse before calling for help" vs. "delays calling for help and requesting an AED").
- **Spatial & geometric reasoning:** Understanding angles, trajectories, and spatial relationships (e.g., "flattens the arc on a left-handed layup" vs. "lowers the entry angle to avoid rotation").
- **Domain-specific expertise:** Applying professional knowledge across sports, music, healthcare, cooking, bike repair, and dance (e.g., "maintains CPR rate at 100–120 bpm" vs. "120–140 bpm").
- **Physical understanding:** Reasoning about biomechanics, force generation, and energy transfer (e.g., "angles knees inward for power" vs. "opens stance toward the rim").

S5 Sample statistics after each filtering stage

Starting from over 400k expert commentaries in Ego-Exo4D, Stage I reduced the number of samples to 108k ($\approx$ –73%). Stage III's length similarity filtering further reduced it from 108k to 60k ($\approx$ –44%). The blind-LLM filtering further shrank the sample size from 60k to 4.5k ($\approx$ –92.5%). Finally, Stage IV expert verification led to 3.5k samples ($\approx$ –22% of the 4.5k), leaving about 0.9% of the original samples. These statistics demonstrate that Stage III is critical as it filters low-quality or easily "cheatable" samples before sending them to the experts for final verification.

https://doi.org/10.52202/085713-0246