VIDEO ACTION DIFFERENCING

Anonymous authors
Paper under double-blind review

ABSTRACT

How do two individuals differ when performing the same action? In this work, we introduce Video Action Differencing, the novel task of identifying subtle differences between videos of the same action, which has numerous applications, such as coaching and skill acquisition. To enable development on this new task, we first create VidDiffBench, a benchmark dataset containing 557 video pairs, with human annotations of 4,719 fine-grained action differences and 2,075 timestamps indicating where these differences occur. Our experiments demonstrate that VidDiffBench poses a significant challenge for state-of-the-art large multimodal models (LMMs), such as GPT-4o, Gemini 1.5 Pro, and Qwen2-VL. By analyzing the failure cases of LMMs on VidDiffBench, we highlight two key challenges for this task: frame-by-frame alignment and fine-grained frame comparison. To overcome these, we propose VidDiff, an agent-based system that breaks the task into three stages: action difference proposal, keyframe localization, and difference verification, each stage utilizing specialized foundation models. The VidDiff method outperforms these baseline LMMs. We release both the dataset[1] and code[2] to encourage and support future research in this domain.

1 INTRODUCTION

The ability to compare two videos of the same action and discern their detailed differences plays a critical role in a wide variety of applications. For instance, in fitness coaching, a novice learning to perform a barbell squat typically watches instructional videos and then compares their actions in a recorded video to identify discrepancies between their movements and those of an expert. In medical training, junior surgeons compare videos of themselves performing surgical procedures with reference videos from experts to identify errors and improve surgical skills.

Despite the significance of comparing actions in videos, effectively analyzing the subtle differences between them remains underexplored. While image comparison has been extensively studied in the field of computer vision (Park et al., 2019; Jhamtani & Berg-Kirkpatrick, 2018; Dunlap et al., 2023; Li et al., 2023; Jiang et al., 2024; Alayrac et al., 2022) and applied to various tasks (Chen et al., 2024; Hu et al., 2023), video action comparison introduces unique challenges. There are two critical obstacles: precise *temporal alignment* and the need for *fine-grained understanding* of action dynamics. Temporal alignment of submovements is essential for meaningful comparisons across video frames, while the identification of subtle differences in action execution requires a nuanced, detailed level of analysis between video pairs.

Current research on video difference understanding largely emphasizes feature visualization (Balakrishnan et al., 2015) or coarse-grained comparisons between different actions or interacting objects (Nagarajan & Torresani, 2024). However, many real-world applications demand fine-grained comparisons between videos of the same action, a challenge that has received comparatively little attention.

We introduce a new task, Video Action Differencing, to advance both academic research and practical applications. Given two videos of the same action, (v_A, v_B), along with a description of the action, the task is to generate two sets of statements: one that is more true for v_A and another for v_B. For example, in a video pair featuring an expert and a novice performing a barbell squat, key differences might include “knees caving in more in video A” and “the squat is deeper in video B”

---

[1] https://huggingface.co/datasets/viddiff/VidDiffBench/

[2] https://anonymous.4open.science/r/VidDiffBench_eval-A0C1/README.md

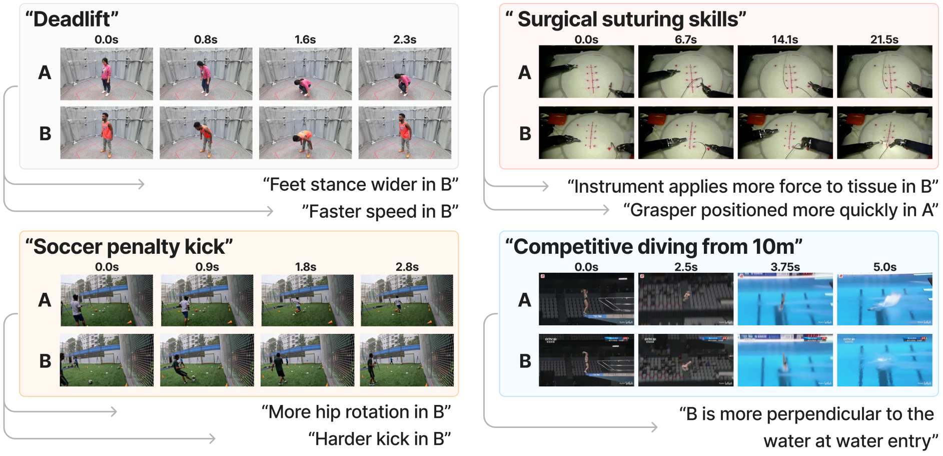
"Deadlift"
0.0s
0.8s
1.6s
2.3s
A
B
"Feet stance wider in B"
"Faster speed in B"
" Surgical suturing skills"
0.0s
6.7s
14.1s
21.5s
A
B
"Instrument applies more force to tissue in B"
"Grasper positioned more quickly in A"
"Soccer penalty kick"
0.0s
0.9s
1.8s
2.8s
A
B
"More hip rotation in B"
"Harder kick in B"
"Competitive diving from 10m"
0.0s
2.5s
3.75s
5.0s
A
B
"B is more perpendicular to the water at water entry"

Figure 1: The Video Action Differencing task and benchmark (VidDiffBench). Given a pair of videos and an action, the task is to generate a list of differences as natural language descriptions. Our VidDiffBench consists of annotated differences across diverse domains, where the differences are relevant to human skill learning. The first row shows the first key challenge: *precise temporal alignment* between segments of the video for comparison. The second row shows the second key challenge: *fine-grained image understanding* of actions in order to perform comparison.

(Figure 1). Since generating the initial difference candidates relies heavily on language capabilities, we also introduce a simpler closed-set setting that focuses on video analysis. In this setting, the target difference strings are provided, and the task is to predict whether each applies more to video A or B.

To facilitate research in this new direction, we present VidDiffBench, a comprehensive benchmark designed for video action differencing. VidDiffBench contains 557 video pairs drawn from domains that require expert feedback, such as fitness, sports, music, and surgery. Each pair is annotated with 4,719 fine-grained differences, along with 2,075 timestamp annotations that identify where these differences occur. Our benchmark is curated with domain expertise, providing a structured taxonomy of differences critical to skill learning. This makes VidDiffBench the first large-scale dataset dedicated to video action differencing, setting a new standard for this emerging task.

In addition to introducing a new task and benchmark, we propose VidDiff, the first agentic framework that addresses the complexity of video action differencing. VidDiff incorporates large language models (LLMs) to propose differences, aligns relevant frames using contrastive language-image models, and verifies the differences using vision-language models (VLMs). We further benchmark both open-source (Qwen2-VL) and proprietary (GPT-4o, Gemini-1.5 pro) video-language models (VLMs) on VidDiffBench. Our results demonstrate that VidDiff outperforms single-stage models, setting a new benchmark for this task and underscoring the importance of structured approaches in fine-grained video comparison.

2 RELATED WORK

Skilled Action Understanding in Videos Video comparison has many potential applications, and our benchmark focuses on the specific goal of natural language feedback in skill learning. Most of the video action comparison papers from this section's first paragraph are systems for skill feedback, showing that skill feedback is well-motivated. Many works give feedback by classifying coarse motion errors, or by visualizing motions, with applications in yoga (Zhao et al., 2022; Thoutam et al., 2022; Chen et al., 2018; Dittakavi et al., 2022; Chen & Yang, 2020; Xie et al., 2019), physical therapy (Velloso et al., 2013), weightlifting (Parmar et al., 2022; Ogata et al., 2019), and general fitness

(Fieraru et al., 2021; Ashwin et al., 2023). The feedback tends to be coarse-grained. In contrast, our task focuses on open natural language feedback, and identifying fine-grained feedback. Recently, the Ego-Exo4D dataset (Grauman et al., 2023) provides videos with expert commentary on skilled actions, which is promising for developing instructional feedback systems. This, along with existing works that give language feedback (Fieraru et al., 2021; Parmar et al., 2022; Velloso et al., 2013), support our claim that language is a good medium for providing skill feedback to humans. Zooming out from skills feedback, skilled action understanding – which includes foundational capabilities for feedback systems – has attracted enormous interest. For example, in sports, music, dance, and surgery, prior works have tackled action recognition (Verma et al., 2020; Shahroudy et al., 2016; Soomro et al., 2012; Zhang et al., 2013; Wang & Zemel, 2016; Chung et al., 2021); spatial and temporal action localization / segmentation (Shao et al., 2020; Liu et al., 2022; Li et al., 2021b; Zhang et al., 2023b; Ibrahim et al., 2016; Garrow et al., 2021; Li et al., 2021b); human pose and motion estimation / reconstruction (Cai et al., 2022; Tang et al., 2023; Wang et al., 2023; Andriluka et al., 2014; Li et al., 2021a; Fieraru et al., 2021; Zhu et al., 2022; Bera et al., 2023; Liu et al., 2024; Grauman et al., 2023); and hand and tool pose estimation (Doosti, 2019; Johnson et al., 2020; 2016; Gao et al., 2014; Grauman et al., 2023). There are also higher level reasoning tasks like question answering (Li et al., 2024), and action quality assessment (Pirsiavash et al., 2014; Parmar & Tran Morris, 2017).

Visual Difference Understanding Only a few prior works have considered video comparison in actions. They mostly emphasize skill learning in similar categories to our benchmark, but their methods tend to tackle single domains. One approach visualizes the user's motion against a target expert motion in video or in augmented reality (AR) (Trout, 2013; Motokawa & Saito, 2006; Han et al., 2016; Kyan et al., 2015; Kurillo et al., 2008). Since interpreting discrepancies between motions is challenging, especially for novices, other works generate visualizations of differences (Liu et al., 2023; Liao et al., 2023; Balakrishnan et al., 2015). In contrast, we summarize action differences in natural language, which enables direct and interpretable feedback. Also, our benchmark covers many skill categories, encouraging the development of generalizable methods that do not require domain-specific training data and methods. The most related work by Nagarajan & Torresani (2024) focuses on coarse-grained step differences in instructional videos using question-answer pairs. In contrast, our approach targets fine-grained action differences, such as a "deeper squat", which offers more detailed insights for skill learning. Additionally, our VidDiff method is zero-shot for a benchmark spanning multiple skilled domains, while their method requires instruction tuning data and is specialized to cooking and entertainment. Beyond inference-time comparison, a number of important works in skill assessment leverage video pairs in training – the supervision signal is commonly a binary of which video shows more skill Doughty et al. (2018; 2019); Pan et al. (2021); Zhang et al. (2023a). In appendix E, we discuss all related datasets having video pairs, finding that none have labels for fine-grained comparison while having a large scale.

Describing differences between *images* in language is an established task called 'difference captioning' or 'change captioning' (Jhamtani & Berg-Kirkpatrick, 2018; Park et al., 2019; Kim et al., 2021; Yao et al., 2022; Hu et al., 2023). LMM evaluation and instruct-tuning papers address image differencing for pairs or small sets of images (Alayrac et al., 2022; Li et al., 2023; Achiam et al., 2023; Jiang et al., 2024). The task of image set differencing with large sets was introduced in (Dunlap et al., 2023). Our video differencing framework uses image differencing with LMMs as a subroutine, however the task of video action differencing with natural language has not previously been explored.

3 VIDEO ACTION DIFFERENCING

Video Action Differencing is a novel and challenging task, offering significant potential for applications in coaching, skill acquisition, and automated performance feedback. To facilitate the development of models capable of handling such a task, we define two complementary task settings: a *closed* setting, evaluated via multiple-choice format, and a more complex *open* setting, requiring generation of action differences. Both are essential for advancing video understanding, especially in contexts where precise feedback on actions is critical.

3.1 TASK DEFINITION

The goal of video action differencing is to identify key differences between two videos where the same action is performed, in a zero-shot setting. We first introduce the simpler *closed-set* version, followed by the more difficult *open-set* variation.

Closed-Set Video Action Differencing: In the closed-set task, the input consists of an action description string s, a video pair (v_A, v_B), and a list of k candidate difference statements $\mathbf{D} = \{d_0, d_1, \ldots, d_{k-1}\}$, such as "the jump is higher." For each k, the model makes k predictions $\mathbf{P} = \{p_0, p_1, \ldots, p_{k-1}\}$, where each prediction is either 'A' (if the statement applies more to v_A) or 'B' (if it applies more to v_B). This setup simulates real-world scenarios, such as coaching, where specific differences of interest are already known. For benchmark purposes, the dataset only includes instances where there is a clear ground-truth label ('A' or 'B') for each difference, which makes evaluation both reliable and automatic.

Open-Set Video Action Differencing: In the open-set task, the input includes the action description string s, a video pair (v_A, v_B), and an integer N_{diff}. The model must generate at most N_{diff} difference statements $\mathbf{D}$ and their associated predictions $\mathbf{P}$, which label the differences as 'a' (for video v_A) or 'b' (for video v_B). This setting is more challenging, as the model must not only identify relevant differences but also generate those differences without any pre-defined options, closely mimicking real-world conditions.

3.2 EVALUATION METRIC

Closed-Set Evaluation: In the closed-set task, the evaluation is straightforward: prediction accuracy is measured as the percentage of correct predictions, where 50% corresponds to random guessing and 100% represents perfect performance. This automatic, unbiased metric provides a reliable baseline for performance comparison.

Open-Set Evaluation: The open-set task introduces additional complexity due to the potential for *ambiguity*—different annotators may disagree on which differences are most important. To address this, we use the recall@N_{diff} metric. Here, we match each ground-truth difference with a predicted difference using a large language model (LLM), specifically GPT-4o. Only 'positive differences'—where the ground-truth label is either 'a' or 'b'—are considered. The recall is calculated as the number of correctly matched and predicted positive differences, divided by the total number of positive differences. We set N_{diff} to be 1.5 times the number of ground-truth differences in the taxonomy, a reasonable limit given that the taxonomy was carefully designed by experts to cover the most important skill-relevant differences. Further details on prompts and matching procedures are provided in appendix F.2.

4 BENCHMARK DATASET AND ANNOTATIONS

The Video Action Differencing task presents a novel challenge in video understanding, requiring precise and systematic comparison of subtle action differences. As no comprehensive benchmark to evaluate this task exists, we introduce VidDiffBench – a comprehensive benchmark specifically designed to test and advance the ability of models to detect fine-grained differences in complex actions. Our benchmark consists of publicly available videos and our human-generated annotations are freely available on HuggingFace Hub[3]. VidDiffBench covers a wide range of actions relevant to skill learning and performance feedback, and is constructed to challenge models across varying levels of difficulty, ensuring its relevance for long-term model development. Table 4 summarizes the key dataset statistics.

4.1 VIDEO DATASETS

The video collection for VidDiffBench was designed to capture a diverse range of actions where performance feedback is essential, ranging from simple exercises to complex professional tasks. This diversity ensures that models are challenged not only on temporal alignment but also on the subtlety and complexity of visual differences. Actions in VidDiffBench span multiple levels of

---

[3] https://huggingface.co/datasets/viddiff/VidDiffBench

Category	Source Dataset	Activity	Video Pair	Difference	Timestamp
Fitness	HuMMan (Cai et al., 2022)	8	193	1,466	310
Ballsports	Ego-Exo4d (Grauman et al., 2023)	4	100	1,013	595
Surgery	JIGSAWS (Gao et al., 2014)	3	168	1,568	672
Music	Ego-Exo4d (Grauman et al., 2023)	2	29	203	320
Diving	FineDiving (Xu et al., 2022)	1	67	469	140
Total		**18**	**557**	**4,719**	**2,075**

Table 1: Summary of VidDiffBench statistics across categories and datasets. We show the number of unique activities, the number of video pairs, annotations for differences, and timestamps.

difficulty—from the basic "hip rotations" in fitness exercises to the intricate "surgical knot tying." This wide coverage tests models across varying degrees of granularity and action complexity.

VidDiffBench features five categories: *Fitness*, *Ballsports*, *Diving*, *Music*, and *Surgery*.

- *Fitness* videos are simple, single-human exercises sourced from HuMMan (Cai et al., 2022), characterized by clean consistent backgrounds, consistent camera viewing angles, and consistent movement patterns.
- *Ballsports* includes basketball and soccer actions from Ego-Exo4D (Grauman et al., 2023), recorded across various environments with diversity in background and action detail.
- *Diving* features high-level Olympic performances from the FineDiving dataset (Xu et al., 2022), capturing subtle and complex movements in professional diving.
- *Music* contains guitar and piano exercises, sourced from Ego-Exo4D (Grauman et al., 2023), focusing on detailed finger and hand movements.
- *Surgery* includes long, intricate procedures such as "knot tying" and "needle passing" from the JIGSAWS dataset (Gao et al., 2014), testing the models on complex medical tasks.

Within each action, video pairs are randomly sampled to ensure a wide range of comparison difficulty, from simple actions to more advanced tasks requiring fine-grained understanding.

4.2 VIDEO ACTION DIFFERENCE ANNOTATIONS

A critical innovation of VidDiffBench is its detailed human-annotated dataset, designed to address two major challenges in action differencing: *ambiguity* in identifying relevant differences and *calibration* consistency among annotators. To tackle ambiguity, we introduce a structured difference taxonomy for each action, ensuring clarity on what aspects are being compared. Then we assign annotators to label video pairs with differences – to handle the calibration challenge we ensure labeling consistency by maintaining a consistent annotator identity within each action. Additionally, we provide frame-level localization annotations of differences, enabling more detailed analysis. In the following section, we describe these components in greater detail.

4.2.1 ANNOTATION TAXONOMY

For each action, we define a structured *difference taxonomy* – a list of key visual differences relevant to the task. For instance, in the basketball jump shot, one difference might be "the ball is more in front of the body." Annotators assign labels to video pairs as follows: 'A' if the difference is more pronounced in video A, 'B' if it's more pronounced in video B, and 'C' if the difference is negligible. By fixing this taxonomy, we address the *ambiguity challenge* – that different annotators may not focus on the same differences. This allows for more objective and consistent comparisons.

We consulted domain experts to create the taxonomies for each action category. For Fitness and Surgery, we worked with a personal trainer and an attending surgeon, respectively, to identify visually salient differences between novice and expert performers. For Ballsports and Music, we extracted relevant differences from expert commentary in the Ego-Exo4D dataset using a large language model (LLM). For Diving, we leveraged the FINA diving manual, processed by an LLM, to identify key distinctions. Differences that were difficult to visually assess, such as "wrist snap" in basketball, were excluded to maintain focus on visually discernible differences.

This method resulted in 147 distinct difference descriptions, which are detailed in Appendix G.2. This fixed taxonomy allows for precise evaluation of model performance across video pairs and helps identify failure cases where models struggle with particular types of differences.

4.2.2 ANNOTATING ACTION DIFFERENCES

For each action a_j and its corresponding differences, annotators reviewed video pairs (v_A, v_B) side-by-side, with the ability to step through frames. Each difference was labeled as 'A' if it applied more to video v_A, 'B' if it applied more to v_B, or 'C' if the difference was insignificant. Consistent annotation was achieved by assigning a single annotator to each action, ensuring that models are evaluated uniformly across all samples. This avoids the *calibration* challenge, that different annotators may have different thresholds for significance.

To verify annotation quality, a second annotator reviewed 25% of the samples. We assessed disagreements where one annotator marked 'A' and the other marked 'B', which occurred in only 2% of cases, indicating low error rates. Annotators were provided with clear visual guidelines to ensure accurate and impartial labeling. On average, annotators spent three minutes per video pair to evaluate five differences, balancing thoroughness and efficiency.

4.2.3 ANNOTATING DIFFERENCE LOCALIZATIONS

In addition to action differences, VidDiffBench provides localization annotations, pinpointing the exact frames in each video where key differences occur. Since identifying localizing frames and aligning them across videos is a key step in performing video action differencing, these annotations enable analysis of model weaknesses, for example through ablation tests in our results section.

We define specific *key points* for each action, representing critical frames where important movements occur. For example, in a squat, key points might include "knees start to bend" and "reaches lowest position." Differences are then linked to these key points, allowing for precise localization annotations. Further details are provided in Appendix C.2.

4.3 DATASET SPLITS AND STATISTICS

Dataset Splits To account for varying levels of difficulty in VidDiffBench, we categorize actions into *easy*, *medium*, and *hard* splits. GPT-4o was used to assign actions to these splits based on descriptions, difference lists, and video lengths. The easy split includes simple movements like Fitness exercises, while medium and hard splits contain more complex actions like Ballsports, Diving, Music, and Surgery. This ensures that models are challenged across a range of difficulties, from basic movements to subtle, fine-grained comparisons.

Dataset Statistics VidDiffBench includes 557 video pairs, 5,580 annotated differences, and 2,075 key point annotations across Fitness, Weightlifting, Ballsports, Surgery, Music, and Diving domains. Video lengths range from a few seconds to several minutes, providing comprehensive coverage of different action complexities. This diversity ensures that VidDiffBench is a robust benchmark for testing and advancing models in fine-grained action comparison. Under the closed setting, the A/B ratio is 0.493/0.507, and in the open setting, the A/B/C ration is 0.259/0.264/0.476.

5 VIDDIFF METHOD

We propose a three-stage framework, **VidDiff**, that effectively addresses the video action differencing task in a zero-shot setting. The method follows a structured pipeline consisting of three key components: Difference Proposer, Frame Localizer, and Action Differencer. Each stage builds on the previous one to progressively refine and validate the identified differences, as in Figure 2. The method described is for the open setting. The method for the closed setting is the same, except the LLM query for candidate differences in stage 1 is replaced with the ground truth differences.

1. Difference Proposer: The Difference Proposer module generates candidate differences for a given action description s. It leverages the extensive knowledge embedded in large language models (LLMs) to predict likely differences between the two videos. For example, given the description "A

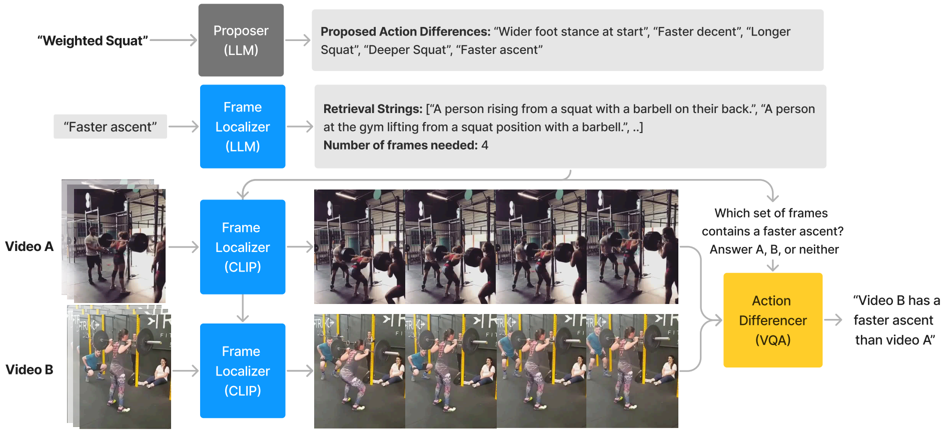
"Weighted Squat"
Proposer (LLM)
Proposed Action Differences: "Wider foot stance at start", "Faster decent", "Longer Squat", "Deeper Squat", "Faster ascent"
"Faster ascent"
Frame Localizer (LLM)
Retrieval Strings: ["A person rising from a squat with a barbell on their back.", "A person at the gym lifting from a squat position with a barbell.", ..]
Number of frames needed: 4
Video A
Frame Localizer (CLIP)
Video B
Frame Localizer (CLIP)
Which set of frames contains a faster ascent? Answer A, B, or neither
Action Differencer (VQA)
"Video B has a faster ascent than video A"

Figure 2: VidDiff Framework. The process begins with a user-supplied action description (e.g., "weighted squat"). The Difference Proposer generates potential differences using a large language model (LLM). The Frame Localizer assigns frames where these differences are observable. Finally, the Action Differencer validates each difference using a vision-language model, determining whether it applies more to video A or video B.

practice basketball jump shot", the module might generate difference candidates such as "the athlete jumps higher". These difference statements, which are visually assessable, form the basis for further analysis. The goal of this stage is to create a diverse set of meaningful and relevant comparisons.

2. Frame Localizer: The Frame Localizer module focuses on identifying the most relevant frames in the video where the proposed differences can be observed. By retrieving the most salient segments from both frames, we solve the key challenge of aligning *precise temporal alignment*, which makes the next stage more effective. Using a large language model, we generate visual cue text strings to guide the localization process. A pretrained CLIP model (Radford et al., 2021) is used to compute frame similarity based on these retrieval strings. To improve temporal alignment, we employ a likelihood model that ensures consistency with the sequence of sub-actions in the videos, solved efficiently using the Viterbi algorithm (Kukleva et al., 2019).

3. Action Differencer: In the final stage, the Action Differencer module validates the proposed differences using vision-language models (VLMs). Given the localized frames from both videos, this module poses multiple-choice questions (derived from the generated difference candidates) to a VLM, which determines whether each difference is more pronounced in v_A, v_B, or if it is indistinguishable. This stage transforms the problem into a structured multiple-choice task, ensuring that each identified difference is rigorously evaluated based on visual evidence.

6 RESULTS

In this section, we present the results of evaluating large multimodal models (LMMs) and VidDiff and on the challenging task of video action differencing, using both closed-set and open-set benchmarks. Our experiments showcase the complexity of this task, particularly in capturing subtle, fine-grained action differences across diverse video categories. We demonstrate that existing state-of-the-art LMMs, such as GPT-4o and Gemini, struggle with these challenges, while our proposed VidDiff method outperforms the baselines, especially in the close-set evaluation. Through detailed error analysis and ablation studies, we uncover key factors that influence model performance, shedding light on future directions for improving video-based model capabilities.

6.1 MAIN RESULTS

As described in Section 3.2, we evaluate our approach on both the *closed-set* and *open-set* tasks. In the closed-set task, models are provided with predefined difference descriptions and must predict whether the difference applies to video A or B. In the open-set task, models are tasked with both generating the difference description and making a prediction. These tasks are fundamental to assessing models' capabilities in fine-grained action comparison.

For our experiments, we benchmark large multimodal models (LMMs) that have demonstrated strong performance in video tasks. Specifically, we use top models from the Video-MME benchmark (Fu et al., 2024): GPT-4o (Achiam et al., 2023), Gemini-1.5-Pro (Reid et al., 2024), and the leading open-source models, Qwen2-VL-7B (Wang et al., 2024; Bai et al., 2023) and LLaVA-Video (Zhang et al., 2024). Following model guidelines, we provide Gemini, Qwen, and VideoLLaVA with raw video inputs, while for GPT-4o we feed frame samples, with text prompts explaining which frames belong to which video. For categories with shorter, fine-grained actions (e.g., Fitness, Ballsports, and Diving), we sample frames at 4-6 fps, while for longer actions (e.g., Music and Surgery), we sample at 2 fps. Our method, VidDiff, is evaluated alongside these baselines, were the proposer LLM is `gpt-4o-2024-08-06`, the localizer embedding model is `CLIP ViT-bigG-14` and frame, and frame differencer VLM is `laion2b_s39b_b160k`.

The results are results shown in Table 2 and Table 3.

Closed-Set Benchmark Performance The closed-set results, presented in Table 2, reveal that video action differencing is a highly challenging task. While some models surpass the random-guessing baseline of 50%, their improvements are modest, especially in the harder splits where no model performs significantly better than chance. VidDiff achieves the best performance on the medium split and comes in a close second on the easy split. Notably, Gemini outperforms GPT-4o on the easy split, but struggles more on the medium split, while the open-source Qwen model consistently lags behind.

Table 2: Results for closed setting (accuracy). Best scores in **bold**, second best <u>underlined</u>. Scores are better than random, with statistical significance highlighted in gray. Significance is p-value$<$ 0.05 on a binomial test.

	Easy	Medium	Hard	Avg
GPT-4o	58.8	53.0	<u>50.1</u>	54.0
Gemini-1.5-Pro	**65.8**	51.9	49.8	<u>55.8</u>
Claude-3.5-Sonnet	56.6	<u>53.5</u>	48.3	52.8
LLaVA-Video-7B	56.6	52.0	48.3	52.3
Qwen2VL-7B	49.0	52.6	49.6	50.4
VidDiff (ours)	<u>65.3</u>	**55.4**	**50.4**	**57.0**

Open-Set Benchmark Performance In the open-set task (Table 3), our method outperforms all other models across most splits, except on the medium difficulty. Among the LMMs, GPT-4o performs much better than Gemini. We analyze this gap by breaking down errors into two categories: *difference recall error*, where the model fails to generate the ground-truth difference, and *flipped prediction error*, where the generated difference is correct but the prediction ('A' or 'B') is incorrect. Closed-set results show minimal flipped prediction error, suggesting that Gemini's main weakness is in difference recall. Specifically, on the easy split, Gemini's recall error is 66% compared to GPT-4o's 30%. Despite generating a similar number of differences as GPT-4o, Gemini struggles to identify the most important ones in our taxonomy, which hampers its performance. Success in the open setting requires strong language capabilities, and this limitation is the bottleneck for handling subtle differences. This explains why, when using the same language proposer, our model performs similarly to GPT-4o.

Table 3: Results for open setting (recall@N_{diff}). Best scores in **bold**, second best <u>underlined</u>.

	Easy	Medium	Hard
GPT-4o	<u>39.5</u>	**35.8**	<u>32.3</u>
Gemini-1.5-Pro	22.7	12.9	21.2
Claude-3.5-Sonnet	31.1	32.5	31.0
LLaVA-Video-7B	7.8	9.0	8.5
Qwen2VL-7B	11.2	8.8	1.6
VidDiff (ours)	**40.1**	<u>34.7</u>	**32.5**

6.2 ABLATION STUDIES

We conducted ablation studies to better understand the individual contributions of different components within VidDiff. These studies focus on the Closed setting, isolating the effects of the frame differencing and frame localization stages.

Frame Differencer Image Comparison In the final stage of VidDiff, the model performs visual question answering (VQA) on frames retrieved from the two videos. To evaluate the effectiveness of this process, we conducted a test using the ground-truth timestamp annotations from VidDiffBench. The results (Table 4) show that even with perfect frame alignment, zero-shot VLMs struggle to consistently detect subtle differences in images. Performance decreases significantly on the medium and hard splits, which suggests room for improvement in zero-shot VLMs' image understanding capabilities.

Split	**Easy**	**Medium**	**Hard**
Acc	78.6	61.2	51.0

Table 4: Ablation study results for frame differencing VQA with ground truth frames. Questions are 3-way multiple-choice.

Frame Localization Design We also analyzed the performance of the Frame Localizer in the closed-set case for the easy split, using ground-truth difference proposals to measure VQA accuracy. Table 5 shows that random frame retrieval leads to significant performance drops, while the addition of Viterbi-based decoding (which enforces a fixed action transcript) substantially improves accuracy. The improvement suggests that temporal alignment plays a critical role in achieving robust video differencing.

In summary, these ablation studies confirm that both accurate frame localization and careful VQA processing are essential to achieving strong performance in video action differencing.

Method	**Accuracy**
Oracle (GT timestamps)	78.6
Random	50.1
Ours w/o Viterbi Decoding	57.4
Ours	65.8

Table 5: Ablation on frame localization using different retrieval techniques on easy.

6.3 DIFFERENCE-LEVEL ERROR ANALYSIS

VidDiffBench's predefined taxonomy allows us to analyze model performance on 147 specific types of action differences, highlighting where models succeed and fail. The results for each difference are detailed in Appendix Table 14, and we perform a statistical significance test to compare models against the random-guessing baseline.

We find that model performance is highly dependent on the visual complexity of the action and the difficulty of localization. Successful examples (Figure 3, left column) show high accuracy for simple, easily localized actions, such as "wider foot stance" in hip rotations (83% accuracy) or "guiding the ball" in a basketball layup (90% accuracy). These cases feature coarse differences that are apparent in most frames, or require only approximate localization.

Conversely, failure cases (Figure 3, right column) often involve precise localization or fine-grained differences. For instance, identifying the angle of a diver's entry into the water in a "10m dive" requires frame-perfect alignment, and recognizing subtle changes in speed in "piano scales" is difficult when reasoning over multi-frames. These challenges highlight the limitations of current models in handling fine-grained video analysis.

7 CONCLUSION

In this paper, we introduce the novel task of Video Action Differencing, aimed at comparing actions in videos. We define this task, compile a meticulously annotated benchmark, and propose a zero-shot agent-based framework, VidDiff. Our findings demonstrate that this task is feasible with current foundation models, although more challenging splits in the benchmark reveal significant opportunities for further methodological improvements. We believe that Video Action Differencing

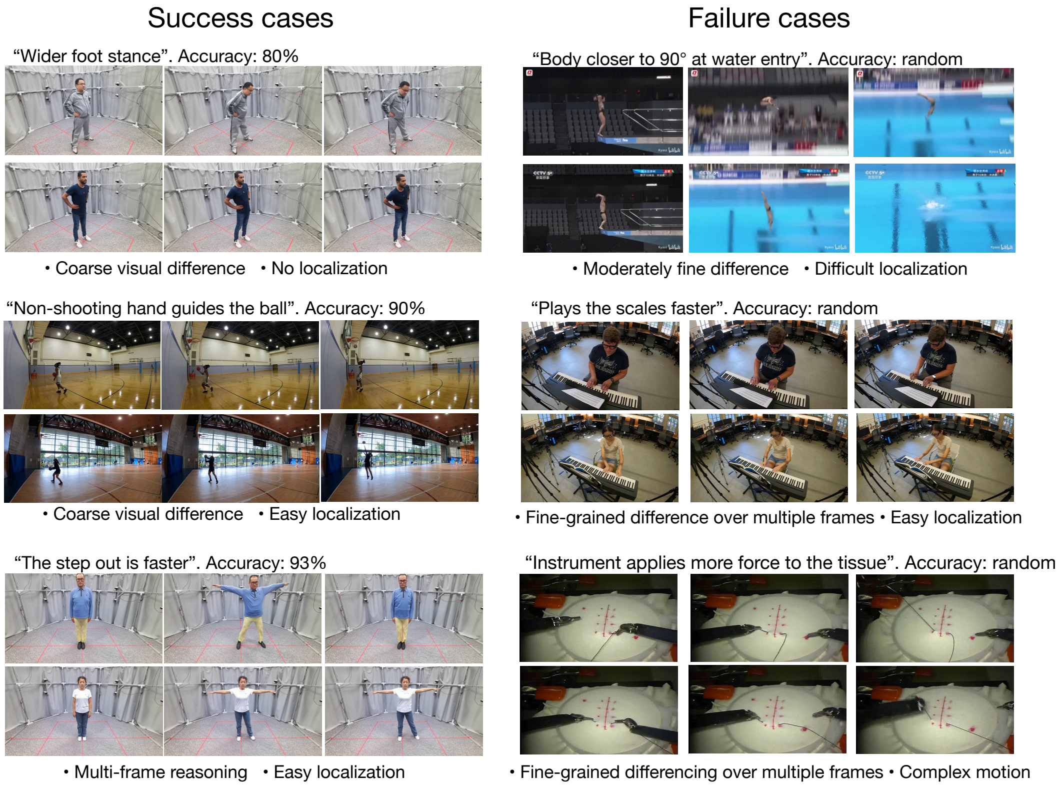
Success cases
Failure cases
“Wider foot stance”. Accuracy: 80%
• Coarse visual difference • No localization
“Body closer to 90° at water entry”. Accuracy: random
• Moderately fine difference • Difficult localization
“Non-shooting hand guides the ball”. Accuracy: 90%
• Coarse visual difference • Easy localization
“Plays the scales faster”. Accuracy: random
• Fine-grained difference over multiple frames • Easy localization
“The step out is faster”. Accuracy: 93%
• Multi-frame reasoning • Easy localization
“Instrument applies more force to the tissue”. Accuracy: random
• Fine-grained differencing over multiple frames • Complex motion

Figure 3: Examples of success cases (left) with high accuracy, and failure cases (right). Successful cases typically involve coarse differences, easy localization, or simple actions, while failure cases often require precise localization or complex motion analysis.

represents a promising research direction with broad applications in fields such as skill acquisition, sports analytics, and scientific research.

8 FUTURE WORK AND LIMITATIONS

While our work demonstrates the potential of Video Action Differencing, there are several areas for future improvement. Enhancing frame retrieval techniques could improve performance on more complex video splits. Additionally, training Vision-Language Models (VLMs) on comparison-specific data may result in better identification of nuanced differences. Further, developing methods tailored to specialized domains such as healthcare or education could unlock more targeted applications. Limitations in our current approach include reliance on general foundation models, which may struggle with domain-specific tasks or fine-grained comparisons. We hope this work encourages further exploration into broader video comparison methods and inspires advancements in these areas.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pp. 3686–3693, 2014.

TS Ashwin, Vijay Prakash, and Ramkumar Rajendran. A systematic review of intelligent tutoring systems based on gross body movement detected using computer vision. *Computers and Education: Artificial Intelligence*, 4:100125, 2023.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Guha Balakrishnan, Frédo Durand, and John Guttag. Video diff: Highlighting differences between similar actions in videos. *ACM Transactions on Graphics (TOG)*, 34(6):1–10, 2015.

Asish Bera, Mita Nasipuri, Ondrej Krejcar, and Debotosh Bhattacharjee. Fine-grained sports, yoga, and dance postures recognition: A benchmark analysis. *IEEE Transactions on Instrumentation and Measurement*, 2023.

Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, et al. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In *European Conference on Computer Vision*, pp. 557–577. Springer, 2022.

Hua-Tsung Chen, Yu-Zhen He, and Chun-Chieh Hsu. Computer-assisted yoga training system. *Multimedia Tools and Applications*, 77:23969–23991, 2018.

Steven Chen and Richard R Yang. Pose trainer: correcting exercise posture using pose estimation. *arXiv preprint arXiv:2006.11718*, 2020.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024.

Jihoon Chung, Cheng-hsin Wuu, Hsuan-ru Yang, Yu-Wing Tai, and Chi-Keung Tang. Haa500: Human-centric atomic action dataset with curated videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13465–13474, 2021.

Bhat Dittakavi, Divyagna Bavikadi, Sai Vikas Desai, Soumi Chakraborty, Nishant Reddy, Vineeth N Balasubramanian, Bharathi Callepalli, and Ayon Sharma. Pose tutor: an explainable system for pose correction in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3540–3549, 2022.

Bardia Doosti. Hand pose estimation: A survey. *arXiv preprint arXiv:1903.01013*, 2019.

Hazel Doughty, Dima Damen, and Walterio Mayol-Cuevas. Who's better? who's best? pairwise deep ranking for skill determination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6057–6066, 2018.

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*, pp. 7862–7871, 2019.

Lisa Dunlap, Yuhui Zhang, Xiaohan Wang, Ruiqi Zhong, Trevor Darrell, Jacob Steinhardt, Joseph E Gonzalez, and Serena Yeung-Levy. Describing differences in image sets with natural language. *arXiv preprint arXiv:2312.02974*, 2023.

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*, pp. 9919–9928, 2021.

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.

Yixin Gao, S Swaroop Vedula, Carol E Reiley, Narges Ahmidi, Balakrishnan Varadarajan, Henry C Lin, Lingling Tao, Luca Zappella, Benjamın Béjar, David D Yuh, et al. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *MICCAI workshop: M2cai*, volume 3, pp. 3, 2014.

Carly R Garrow, Karl-Friedrich Kowalewski, Linhong Li, Martin Wagner, Mona W Schmidt, Sandy Engelhardt, Daniel A Hashimoto, Hannes G Kenngott, Sebastian Bodenstedt, Stefanie Speidel, et al. Machine learning for surgical phase recognition: a systematic review. *Annals of surgery*, 273(4):684–693, 2021.

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. *arXiv preprint arXiv:2311.18259*, 2023.

Ping-Hsuan Han, Kuan-Wen Chen, Chen-Hsin Hsieh, Yu-Jie Huang, and Yi-Ping Hung. Ar-arm: Augmented visualization for guiding arm movement in the first-person perspective. In *Proceedings of the 7th Augmented Human International Conference 2016*, pp. 1–4, 2016.

Xinyue Hu, Lin Gu, Qi A. An, Mengliang Zhang, Liangchen Liu, Kazuma Kobayashi, Tatsuya Harada, Ronald M. Summers, and Yingying Zhu. Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering. *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023. URL `https://api.semanticscholar.org/CorpusID:260125237`.

Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1971–1980, 2016.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. *arXiv preprint arXiv:1808.10584*, 2018.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning, April 2024. URL `https://tiger-ai-lab.github.io/Blog/mantis`.

David Johnson, Isabelle Dufour, Daniela Damian, and George Tzanetakis. Detecting pianist hand posture mistakes for virtual piano tutoring. In *Proceedings of the international computer music conference*, pp. 168–171, 2016.

David Johnson, Daniela Damian, and George Tzanetakis. Detecting hand posture in piano playing using depth data. *Computer Music Journal*, 43(1):59–78, 2020.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. Dspy: Compiling declarative language model calls into self-improving pipelines. 2024.

Hoeseong Kim, Jongseok Kim, Hyungseok Lee, Hyun a Park, and Gunhee Kim. Viewpoint-agnostic change captioning with cycle consistency. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2075–2084, 2021.

Anna Kukleva, Hilde Kuehne, Fadime Sener, and Jurgen Gall. Unsupervised learning of action classes with continuous temporal embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12066–12074, 2019.

Gregorij Kurillo, Ruzena Bajcsy, Klara Nahrsted, and Oliver Kreylos. Immersive 3d environment for remote collaboration and training of physical activities. In *2008 IEEE Virtual Reality Conference*, pp. 269–270. IEEE, 2008.

Matthew Kyan, Guoyu Sun, Haiyan Li, Ling Zhong, Paisarn Muneesawang, Nan Dong, Bruce Elder, and Ling Guan. An approach to ballet dance training through ms kinect and visualization in a cave virtual reality environment. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6 (2):1–37, 2015.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023.

Haopeng Li, Andong Deng, Qiuhong Ke, Jun Liu, Hossein Rahmani, Yulan Guo, Bernt Schiele, and Chen Chen. Sports-qa: A large-scale video question answering benchmark for complex and professional sports. *arXiv preprint arXiv:2401.01505*, 2024.

Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13401–13412, 2021a.

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*, pp. 13536–13545, 2021b.

Chen-Chieh Liao, Dong-Hyun Hwang, Erwin Wu, and Hideki Koike. Ai coach: A motor skill training system using motion discrepancy detection. In *Proceedings of the Augmented Humans International Conference 2023*, pp. 179–189, 2023.

Ruofan Liu, Erwin Wu, Chen-Chieh Liao, Hayato Nishioka, Shinichi Furuya, and Hideki Koike. Pianosyncar: Enhancing piano learning through visualizing synchronized hand pose discrepancies in augmented reality. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 859–868. IEEE, 2023.

Yanchao Liu, Xina Cheng, and Takeshi Ikenaga. Motion-aware and data-independent model based multi-view 3d pose refinement for volleyball spike analysis. *Multimedia Tools and Applications*, 83(8):22995–23018, 2024.

Yi Liu, Limin Wang, Yali Wang, Xiao Ma, and Yu Qiao. Fineaction: A fine-grained video dataset for temporal action localization. *IEEE transactions on image processing*, 31:6937–6950, 2022.

Yoichi Motokawa and Hideo Saito. Support system for guitar playing using augmented reality display. In *2006 IEEE/ACM International Symposium on Mixed and Augmented Reality*, pp. 243–244. IEEE, 2006.

Tushar Nagarajan and Lorenzo Torresani. Step differences in instructional video. *arXiv preprint arXiv:2404.16222*, 2024.

Ryoji Ogata, Edgar Simo-Serra, Satoshi Iizuka, and Hiroshi Ishikawa. Temporal distance matrices for squat classification. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops*, pp. 0–0, 2019.

Jia-Hui Pan, Jibin Gao, and Wei-Shi Zheng. Adaptive action assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8779–8795, 2021.

Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4624–4633, 2019.

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

Paritosh Parmar, Amol Gharat, and Helge Rhodin. Domain knowledge-informed self-supervised representations for workout form assessment. In *European Conference on Computer Vision*, pp. 105–123. Springer, 2022.

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*, pp. 556–571. Springer, 2014.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1010–1019, 2016.

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*, pp. 2616–2625, 2020.

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.

Yansong Tang, Jinpeng Liu, Aoyang Liu, Bin Yang, Wenxun Dai, Yongming Rao, Jiwen Lu, Jie Zhou, and Xiu Li. Flag3d: A 3d fitness activity dataset with language instruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22106–22117, 2023.

Vivek Anand Thoutam, Anugrah Srivastava, Tapas Badal, Vipul Kumar Mishra, GR Sinha, Aditi Sakalle, Harshit Bhardwaj, and Manish Raj. Yoga pose estimation and feedback generation using deep learning. *Computational Intelligence and Neuroscience*, 2022, 2022.

Josh Trout. Digital movement analysis in physical education. *Journal of Physical Education, Recreation & Dance*, 84(7):47–50, 2013.

Eduardo Velloso, Andreas Bulling, and Hans Gellersen. Motionma: Motion modelling and analysis by demonstration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1309–1318, 2013.

Manisha Verma, Sudhakar Kumawat, Yuta Nakashima, and Shanmuganathan Raman. Yoga-82: a new dataset for fine-grained classification of human poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 1038–1039, 2020.

Kuan-Chieh Wang and Richard Zemel. Classifying nba offensive plays using neural networks. In *Proceedings of MIT Sloan sports analytics conference*, volume 4, 2016.

Kuan-Chieh Wang, Zhenzhen Weng, Maria Xenochristou, João Pedro Araújo, Jeffrey Gu, Karen Liu, and Serena Yeung. Nemo: Learning 3d neural motion fields from multiple video instances of the same action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22129–22138, 2023.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Haoran Xie, Atsushi Watatani, and Kazunori Miyata. Visual feedback for core training with 3d human shape and pose. In *2019 Nicograph International (NicoInt)*, pp. 49–56. IEEE, 2019.

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*, pp. 2949–2958, 2022.

Linli Yao, Weiying Wang, and Qin Jin. Image difference captioning with pre-training and contrastive learning. In *AAAI Conference on Artificial Intelligence*, 2022.

Shao-Jie Zhang, Jia-Hui Pan, Jibin Gao, and Wei-Shi Zheng. Adaptive stage-aware assessment skill transfer for skill determination. *IEEE Transactions on Multimedia*, 2023a.

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*, pp. 2405–2414, 2023b.

Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *Proceedings of the IEEE international conference on computer vision*, pp. 2248–2255, 2013.

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.

Ziyi Zhao, Sena Kiciroglu, Hugues Vinzant, Yuan Cheng, Isinsu Katircioglu, Mathieu Salzmann, and Pascal Fua. 3d pose based feedback for physical exercises. In *Proceedings of the Asian Conference on Computer Vision*, pp. 1316–1332, 2022.

Kevin Zhu, Alexander Wong, and John McPhee. Fencenet: Fine-grained footwork recognition in fencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3589–3598, 2022.