VIDEO ACTION DIFFERENCING

James Burgess[1], **Xiaohan Wang**[1], **Yuhui Zhang**[1], **Anita Rau**[1], **Alejandro Lozano**[1], **Lisa Dunlap**[2], **Trevor Darrell**[2], **Serena Yeung-Levy**[1]
[1]Stanford, [2] UC Berkeley

ABSTRACT

How do two individuals differ when performing the same action? In this work, we introduce Video Action Differencing (VidDiff), the novel task of identifying subtle differences between videos of the same action, which has numerous applications, such as coaching and skill learning. To enable development on this new task, we first create VidDiffBench, a benchmark dataset containing 549 video pairs, with human annotations of 4,469 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 and Qwen2-VL. By analyzing the failure cases of LMMs on VidDiffBench, we highlight two key challenges for this task: localizing relevant sub-actions over two videos and fine-grained frame comparison. To overcome these, we propose the VidDiff method, an agentic workflow that breaks the task into three stages: action difference proposal, keyframe localization, and frame differencing, each stage utilizing specialized foundation models. To encourage future research in this new task, we release the benchmark[1] and code[2].

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.

There are two critical obstacles. First is precise *localization of sub-actions* where differences might occur: finding differences requires aligning sub-action frames where differences might occur. Second is *fine-grained understanding*: the ability to perceive subtle visual differences in motions.

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 (VidDiff). 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" (Figure 1). Since generating the initial difference candidates relies heavily on language capabilities, we also introduce a simpler 'closed' 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 549 video pairs drawn from domains

[1]Benchmark: https://huggingface.co/datasets/jmhb/VidDiffBench
[2]Project page: http://jmhb0.github.io/viddiff

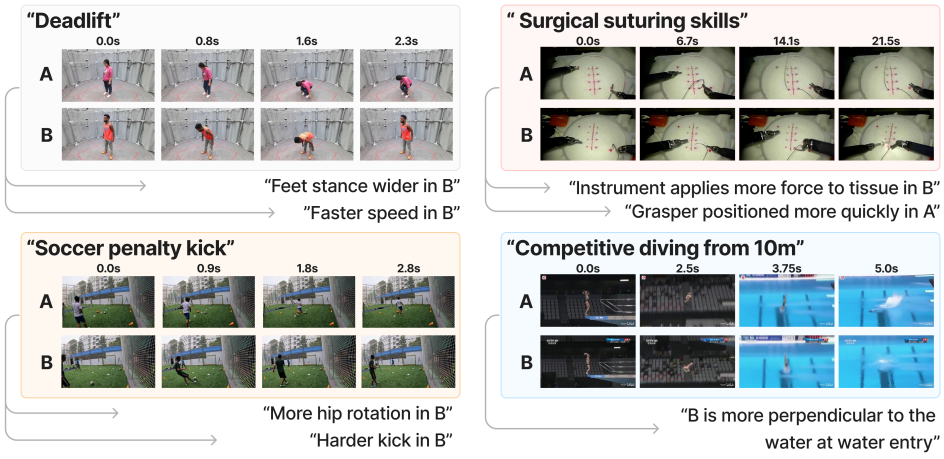

Figure 1: The Video Action Differencing task and benchmark (VidDiffBench). Given a pair of videos and an action, the task is to generate a list of differences as natural language descriptions. Our VidDiffBench consists of annotated differences across diverse domains, where the differences are relevant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

that require expert feedback, such as fitness, sports, music, and surgery. The videos are annotated with 4,469 fine-grained differences ($\sim$8 per video pair), along with 2,075 timestamp annotations that identify where these differences occur. To ensure the annotated differences are relevant to skill learning, we create a taxonomy of action differences that leverages domain expertise. This makes VidDiffBench the first large-scale dataset dedicated to video action differencing.

In addition to introducing a new task and benchmark, we propose the VidDiff method, an agentic workflow (Anthropic, 2025) that addresses the complexity of video action differencing. The method incorporates large language models (LLMs) to propose differences, localizes relevant frames using contrastive language-image models (CLIP), and compares frames for differences using vision-language models (VLMs). We further benchmark both open-source (Qwen2-VL, LLaVA-Video) and proprietary (GPT-4o, Gemini-1.5 pro, Claude 3.5 Sonnet)) large multimodal models (LMMs) on VidDiffBench. Our results demonstrate that VidDiff performs strongly over open and closed settings, 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; Aklilu et al., 2024); 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). Skilled action domains also tackle 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. Beyond inference-time comparison, a number of important works in skill assessment leverage video pairs in training – here the supervision signal is typically a binary variable indicating which video demonstrates greater 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 being large scale, unlike our VidDiffBench

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 (VidDiff) 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 skill-based differences between two videos of the same action, 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 is 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 a 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 like 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 is 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 differences, but also generate those differences without any pre-defined options, closely mimicking real-world conditions.

3.2 Evaluation Metric

Our choice of benchmark evaluation metrics is driven by two major challenges for designing annotations: *ambiguity* and *calibration*. First, there is ambiguity around what differences are important for performing an action skillfully. Second, annotators are calibrated differently – they have different thresholds for whether a difference like "wider feet stance" is different enough to be annotated.

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 (assuming a balanced evaluation set). There is no *ambiguity* because we provide the possible differences. There is no *calibration* issue because the answer must be 'A' or 'B' (and not 'C' for "not different"). Overall, it's an automatic metric that focuses on video understanding.

Open-Set Evaluation: In the open-set task, we use an LLM query (GPT-4o) to match the ground truth difference strings to predicted difference strings in a 'partial matching'. Then we only consider "positive differences" – where the ground-truth label is 'A' or 'B' and not 'C'. Then the recall@N_{diff} is calculated as the number of correctly matched and predicted positive differences, divided by the total number of positive differences. To handle the *ambiguity* of what differences are relevant, we set N_{diff} to be 1.5 times the number of labeled differences, so models can predict more differences without penalty. This is a reasonable number because the annotation taxonomy is designed to cover skill-relevant differences. Moreover, we handle the *calibration* challenge of whether a difference is 'above a threshold' by only considering the positive differences where ground truth is 'A' or 'B'.

4 Benchmark Dataset and Annotations

The Video Action Differencing task presents a novel challenge in video understanding, requiring precise 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-created 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 localization but also on the subtlety and complexity of visual differences. Actions in VidDiffBench span multiple levels of

[3] https://huggingface.co/datasets/jmhb/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	98	996	595
Surgery	JIGSAWS (Gao et al., 2014)	3	166	1,386	672
Music	Ego-Exo4d (Grauman et al., 2023)	2	29	180	320
Diving	FineDiving (Xu et al., 2022)	1	63	441	140
Total		**18**	**549**	**4,469**	**2,075**

Table 1: Summary of VidDiffBench statistics across categories and datasets: number of unique activities, 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. The are five categories:

- *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 some diversity in background and camera angle, as well as action detail.
- *Diving* features high-level Olympic performances from the FineDiving dataset (Xu et al., 2022), capturing subtle and complex movements in professional diving. The backgrounds may different, but the camera angles are consistent.
- *Music* contains guitar and piano exercises from Ego-Exo4D (Grauman et al., 2023), focusing on detailed finger and hand movements. Background and camera angles can vary.
- *Surgery* includes long, intricate procedures such as "knot tying" and "needle passing" from the JIGSAWS dataset (Gao et al., 2014). The background and camera angles are consistent.

Within each action, video pairs are randomly sampled to ensure a wide range of comparison difficulty. The range of tasks is broad in terms of action complexity and background variation.

4.2 Video Action Difference Annotations

A critical innovation of VidDiffBench is its detailed human-annotated dataset, designed to address two major challenges in annotating the video differencing task: *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, which can enable analysis for future model development. 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, a skill-relevant difference might be "the ball is more in front of the body"; on the other hand, we do not include differences not directly relevant to skill performance like "the athlete is taller". 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 differences. We filtered differences that were difficulty to visually asses, such as "more wrist snap" in basketball jump shot (because video resolution was not high enough).

This method resulted in 148 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 about eight differences.

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: for example the difference "faster squat descent" is defined as the frame spanning "knees start to bend" and "reaches lowest position". 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 549 video pairs, 4,469 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 ratio is 0.259/0.264/0.476.

5 VidDiff Method

We propose a three-stage framework, the VidDiff method, 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.

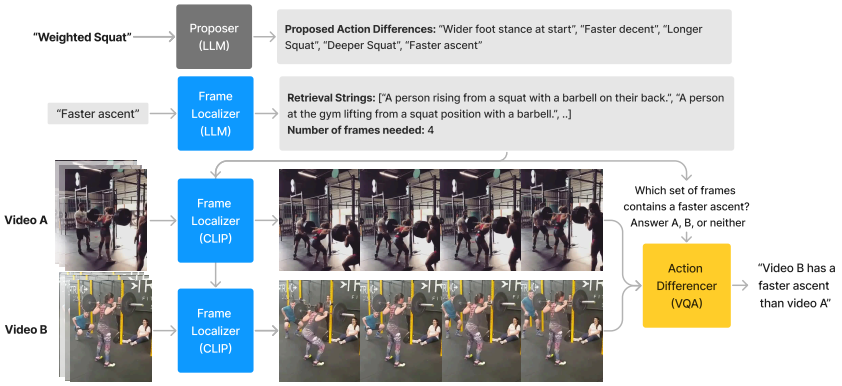

Figure 2: VidDiff Method. One input is an action description (e.g. "weighted squat"). The Difference Proposer generates potential differences using a large language model (LLM). The Frame Localizer assigns frames where these differences are observable. Finally, the Action Differencer checks each difference using a vision-language model, determining whether it applies more to video A or video B, or neither.

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 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 *temporal localization of sub-actions*, which makes the next stage more effective. Our approach is to do temporal sub-action segmentation. The LLM takes uses the action description string to produce a list of sub-actions, along with retrieval strings to guide localization. A pretrained CLIP model (Radford et al., 2021) is used to compute frame similarity based on these retrieval strings, and then we assign each frame to one of the sub-actions. Here, we use a Viterbi-based algorithm (Kukleva et al., 2019), which assigns each frame to a sub-action based on its similarity score, while enforcing that the frames follow the fixed sequence of sub-actions. Finally, the LLM predicts a mapping between the sub-actions and their corresponding differences, yielding a set of precisely localized frames for each difference.

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. Moreover, by providing the localized-frames relevant to each difference

6 Results

In this section, we present the results of evaluating large multimodal models (LMMs) and our VidDiff method and on the challenging task of video action differencing on our VidDiffBench benchmark. Our experiments show 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), Claude 3.5 Sonnet Anthropic (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 video inputs, while for GPT-4o and Claude we give frames, 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 differencer VLM is `gpt-4o-2024-08-06`. The results are results shown in Table 2 and Table 3.

Closed-Set Benchmark Performance The closed-set results are in Table 2, revealing that video action differencing is a highly challenging task. While some models surpass the random-guessing baseline of 50% – where gray shading indicates better-than-random with statistical significance – their improvements are modest, especially in the harder splits where no model performs significantly better than chance. Gemini, which has emphasized its results in video understanding, has the strongest overall performance. Our VidDiff method, which uses GPT-4o as a visual perception backbone, outperforms GPT-4o on the raw video frames and is second overall, demonstrating the value of our scaffolding for this task. LLava-Video is competitive with GPT and Claude, while Qwen2-VL performs poorly, possibly related to instruction-following challenges appendix G.4

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

	Easy	Med	Hard	Avg
GPT-4o	58.3	53.2	48.9	53.5
Gemini-1.5-Pro	**67.8**	53.6	51.7	**57.7**
Claude-3.5-Sonnet	57.1	50.5	**52.5**	53.4
LLaVA-Video	56.6	52.0	48.3	52.3
Qwen2-VL-7B	49.0	52.6	49.6	50.4
VidDiff (ours)	62.7	**56.2**	50.0	56.3

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. The closed-set results show that Gemini has lower 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 underlined.

	Easy	Med	Hard	Average
GPT-4o	45.7	**41.5**	38.0	41.7
Gemini-1.5-Pro	30.3	30.5	24.1	28.3
Claude-3.5-Sonnet	37.8	34.6	34.3	35.6
LLaVA-Video	7.8	9.0	8.5	8.4
Qwen2-VL-7B	11.2	8.8	1.6	7.2
VidDiff (ours)	**49.9**	37.9	**38.5**	**42.1**

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.

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.

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.

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	62.7

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 148 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.

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 involve fine differences, precise localization or complex actions.

7 Conclusion

In this paper, we introduce the novel task of Video Action Differencing (VidDiff), aimed at comparing actions in videos. We define this task, compile a meticulously annotated benchmark, and propose a zero-shot agent-based framework. 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 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.