A BENCHMARK: DOWNLOAD INSTRUCTIONS

Our benchmark is released at https://huggingface.co/datasets/viddiff/VidDiffBench. It has complete instructions on how to access annotations, how to download external datasets, and all licenses for our annotations and the source video datasets.

B BENCHMARK: DIFFERENCE ANNOTATION TAXONOMY GENERATION

Each dataset underwent a thorough taxonomy generation process. Details for each dataset are presented in this section. Most datasets were first processed with the Difference Proposer.

B.1 GENERATING JIGSAWS DIFFERENCE ANNOTATION TAXONOMY FOR SURGERY VIDEOS

To produce the JIGSAWS taxonomy for surgery videos, we first used the Difference Proposer to generate difference candidates. As we found many proposed variations to be irrelevant, we consulted a surgeon from ¡anonymous_hospital¿. We first showed them a variety of videos from the JIGSAWs dataset and brainstormed visually discernible differences. Next, we shared the GPT-generated variations with the surgeon and discussed which ones should be added to the brainstormed listed. Lastly, we dropped all differences that could not be annotated consistently. For instance, we removed ”Surgeon exploits robotic instrument’s range of movement more efficiently in Video A than in Video B”, as the difference is subject to interpretation.

B.2 GENERATING EGO-EXO4D DIFFERENCE ANNOTATION TAXONOMY

For datasets with expert annotation, such as Basketball, Soccer, and Music, we processed the expert commentary with GPT instead of asking for difference proposals directly. We asked GPT to summarize the expert commentary using the following prompt:

```
Below are a sequence of text strings.
The strings are written by experts who are watching videos of a
 ↪  task with this description: "{action_descriptions}".
As the experts watch each video, they pause the video and record
 ↪  verbal commentary about how that person is performing the
 ↪  task.

Return a list of text strings that summarizes what are the key
 ↪  visual cues that the expert is looking when they provide
 ↪  feedback.
Each item in the list should be specific and testable, so that
 ↪  anybody watching a single video can assess whether that
 ↪  visual cue applies to a particular video.

Your response should be a json with this structure:
{ "summary_texts" : ["text0", "text1", ...] }
The list should have at least 15 items.

Here are the texts to summarize:
{texts}
```

As GPT has no knowledge of the specifics of the dataset, we then manually parsed the proposed list and kept items that could be distinguished by non-experts based on visual information only. Some differences were not visible in our data. For instance, the ”wrist snap” in basketball was excluded, as it could not be discerned from the videos.

As we found that GPT lost a lot of information from the expert commentary, we also manually parsed the expert comments and added key visual cues that were mentioned by the experts.

B.3 Generating FineDiving difference annotation taxonomy

For the diving dataset, we used the Difference Proposer. As all videos are of experts, there is little variation between the videos. We thus used the diving score annotations given in the dataset, to find pairs of images with more variability and used those proposed differences that were discernible in these pairs.

C Benchmark: annotation details

Two annotators were provided with a list of variations and a folder containing all videos as well as concatenated videos of two actions next to each other. They were instructed to annotate differences only as such, if they were obvious. The instructions for the annotators were as follows:

```
 You'll get pairs of videos and be asked questions about how they're
  ↪  different. They'll be specific querstions:
- E.g. "Which one has a wider foot stance: (a) video A, (b) video B,
↪  or (c) they're the same.
- if it's hard to tell whether there really is a difference, then say
↪  "c".  Rule of thumb: once you've found the important point in the
↪  video, if it takes you more than 10 seconds to make a decision,
↪  then say "c".
```

C.1 Benchmark: the difference taxonomy

The difference taxonomy is available as part of the benchmark release at this link. The full list of differences can also be previewed in the analysis table 14.

C.2 Benchmark: retrieval annotation generation

For the closed-evaluation scenario, we need to temporally align the videos we wish to compare. For this alignment, we annotate retrievals, which are important and identifiable moments within an action. To generate retrievals, we used the Frame localizer, where we prompt GPT4-V to propose stages for a given action (see Section C.3 for details). We found that the stages were sometimes to coarse. We thus either manually identified key moments that helped retrieve sections of the videos important for comparing differences, or, for JIGSAWS consulted our expert. The retrieval annotations are available as part of the VidDiffBench benchmark release, and are in the folder called 'retrieval'.

C.3 Prompts for action assignment to the easy/medium/hard splits

```
I'm designing a benchmark for comparing pairs of videos of the same
↪  action.
We have many actions and each action has a list of differences we look
↪  for.
The benchmark's task is to examine differences and say whether the
↪  statement applies more to "video A" or "video B".

Below we show a dictionary where each element is a single action.
Each action has an "action_description" describing the action.
It also has "average_seconds_per_video", for the median length of videos
↪  in seconds.
Each action has a dictionary of "differences", where each difference has
↪  these keys:
- 'name' for the difference
- 'description' describing the difference

Finally, for each action, there are two unfinished options:
- 'split' which currently says 'easy|medium|hard'
- 'split_reason' which currently says '...'
Your task is to fill in these values:
```

```
- Decide whether the 'split' value is 'easy', 'medium' or 'hard'. This
↪  evaluation judges the difficulty of performing actionn difference
↪  comparison for all differences within an action. Having a high
↪  number of actions should not be considered as criteria for
↪  difficulty.
- Justify your choice in 'split_reason'.

Return the same dictionary as json, with the values of 'split' and
↪  'split_reason' populated.
Here are the actions.
{actions}
```

The `actions` field is replaced with a json with the structure that is described in the prompt.

C.4 Difficulty splits

The result of the difficulty splits is in table 6.

Table 6: The difficulty splits with action code and names

Split	Action	Action Name	Action description
easy	fitness_0	Hip circle anticlockwise	fitness exercise called standing hip circle with hands on hips, one rotation anticlockwise
easy	fitness_3	lucky cat	fitness exercise called two arm standing lucky cat starting with arms up, one repetition
easy	fitness_4	Squat Knee Raise side view	squat without weights, then knee raise on left side
easy	fitness_6	Hip circle clockwise	fitness exercise called standing hip circle with hands on hips, one rotation clockwise
medium	ballsports_0	basketball jump shot	a person is doing a basketball mid-range jump shot, starting with the ball in their hand, no defense, practice only
medium	ballsports_1	basketball mikan layup	a person does the Basketball drill called the Mikan layup where they start under the basket, do a layup with the right hand and catch it, do a left hand layup and catch it, no defense, practice only
medium	ballsports_2	basketball reverse layup	a person playing basketball does a reverse layup starting from the left side of the basket and lays it up with their right hand on the right hand side, no defense, practice only
medium	ballsports_3	soccer penalty kick	a person does a soccer drill where they do a single penalty kick, practice only, no defense, no goalie
medium	diving_0	diving	competitive diving from 10m
medium	fitness_1	Opening and closing step left side first	fitness exercise called opening and closing step on left side and then opening and closing step on right side
medium	fitness_2	car deadlift	a single free weight deadlift without any weight
medium	fitness_5	Squat Knee Raise diagonal view	a squat then a knee raise on left side
medium	fitness_7	Opening and closing step right side first	fitness exercise called opening and closing step on right side and then opening and closing step on left side
hard	music_0	piano	a person is playing scales on the piano
hard	music_1	guitar	a person is playing scales on the guitar
hard	surgery_0	Knot Tying	The subject picks up one end of a suture tied to a flexible tube attached at its ends to the surface of the bench-top model, and ties a single loop knot.
hard	surgery_1	Suturing	The subject picks up needle, proceeds to the incision (designated as a vertical line on the bench-top model), and passes the needle through the fake tissue, entering at the dot marked on one side of the incision and exiting at the corresponding dot marked on the other side of the incision. After the first needle pass, the subject extracts the needle out of the tissue, passes it to the right hand and repeats the needle pass three more times.
hard	surgery_2	Needle Passing	The subject picks up the needle (in some cases not captured in the video) and passes it through four small metal hoops from right to left. The hoops are attached at a small height above the surface of the bench-top model.

C.5 Validating split generation - human study

Choosing the difficulty splits requires a holistic view of all the actions, so we decided it didn't make sense for experts to suggest them, since they are only familiar with a few actions each. On the other hand, we didn't want to rank the splits based on performance of current models since this felt like biasing towards current models; and besides, the performance for many actions in 'medium' and 'hard' is already random, so it would be hard to differentiate these actions. LLMs are a good candidate because they have a good understanding of the actions and are relatively free of the biases of this paper's authors Furthermore, human annotators could not do the ranking, because no human annotated all the actions.

To further support the choice of an LLM, we asked 3 humans to rank the action comparisons from easiest to hardest, and compared against the LLM ranking. We then computed the Spearman's rank correlation between all ranking sets, and the results are in table 7. The mean of the pairwise correlations between the humans was 0.602, while the mean of pairwise correlations between the LLM and humans was higher at 0.673. This shows (i) that there is non-negligible variability in human rankings, and (ii) that the LLM ranking is reasonable, and actually better correlated with most humans compared to several of the human annotations.

Table 7: Results on human evaluation study on choosing the splits. This is the Spearman's rank corrlation between the ranks of action dificulty, comparing our LLM approach and 3 humans.

	LLM	**Human 1**	**Human 2**	**Human 3**
LLM		53.1	68.0	80.6
Human 1	53.1		45.9	64.5
Human 2	68.0	45.9		70.3
Human 3	80.6	64.5	70.3	
Average	67.3	54.5	61.4	71.8

C.6 Further dataset construction considerations

Camera angles The change of camera angle perspective does make the task harder. For samples in the 'Fitness' category, the camera angle is the same because the source dataset has a fixed camera rig, and we chose to use the same camera angle. For samples in 'diving' and 'surgery' categories, the camera angle is approximately the same. On the other hand, samples from 'ballsports' and 'music' categories can change. A related attribute (not mentioned here) is differences in background – similarly the 'ballsports' and 'music' categories often had different backgrounds as well. Importantly, these attributes were considered when assigning the difficulty splits. This may partly explain why the fitness exercises are all in the easy and medium split.

FPS Each video pair has the same fps. In case others want to leverage our code with new videos, our code does handles the case where FPS is different. Specifically, the input configuration has a value for the target FPS for running inference, and we subsample the video to have this FPS. (If the videos cannot be subsampled to have the exact target fps, then a warning is printed).

Impact of different actor heights For annotator instructions, we clarified that all differences on things like distance should be relative to the actor's height. We gave the example of 'wider foot stance', saying that if a 5ft actor and a 6ft actor both had their legs 3ft apart, then the shorter actor has a 'wider foot stance' relative to their height. This reflects what is commonly understood by descriptions like these in skills coaching.

D Benchmark statistics

Beyond the main statistics in the main, table 8 shows further statistics broken down by difficulty splits.

Average video length is longer as the difficulty gets higher: 2.1/3.9/18.7 seconds, for easy/medium/hard. Compared to video QA datasets, the lengths are relatively shorter because we focus on fine-grained action understanding in how actions are performed. The total length of videos is 163 minutes.

Retrieval tags, temporal bias For the 'retrieval tags', we first show the number of retrieval tags – 9554 total. To give insight into their distribution within each video, each instance is normalized to the video length, and compute its 'video location'. E.g. in a squat, the starting position might

Table 8: Detailed data statistics by split

Split	easy	medium	hard	Overall
# video pairs	95	265	197	557
Avg video length (secs)	2.1	3.9	18.7	8.8
Total video length (mins)	6.5	34.7	122.5	163.7
# differences tagged	1224	4788	3542	9554
StdDev within retrieval type	8.4%	5.2%	4.1%	5.9%
StdDev across retrieval types	17.3%	25.7%	20.2%	21.0%
Difference annotations count	578	1771	2370	4719
Difference annotations A/B/C distribution	167/190/221	622/605/1143	435/452/884	1224/1247/2248

be position 0.1, the bottom of the descent 0.45, and the squat finish at 0.87. Within each retrieval type, we compute 'StdDev within retrieval type', which intuitively measures how well-aligned are the key points in the video. For example, if the average squat video records 'bottom of descent' at location 0.45, and 'within StdDev' is 0.06, then the mean distance from the average is 0.06 (so at 0.39 or 0.51). The 'within StdDev' is on average 0.059, indicating there is some variation in retrieval position, but there is temporal bias. This is expected since each video is trimmed and contains an atomic action. Future benchmarks could use untrimmed videos to make retrieval annotations less aligned, but the present benchmark is already difficult for SOTA models, so this is unnecessary now.

Retrieval tags, coverage We also measure 'StdDev across retrieval types', meaning the standard deviation of different retrieval classes within one video. Intuitively this measures how much of the video is 'covered' by retrieval keypoints. This is 0.21 on average. So if the mean of retrieval keypoints were 0.5, then the average retrieval annotations is around 0.29 or 0.71 in the video.

A/B/C distribution Additionally, we have shown the count of difference annotations and the A/B/C distribution; the 'no difference' annotation of 'C' is the most prevalent.

E BENCHMARK: RELATED VIDEO PAIR DATASETS

A small number of prior works have datasets of paired videos with some label of the difference. However none have labels for fine-grained comparison while also having a large scale.

- Nagarajan & Torresani (2024) has a large-scale dataset of video differences in instructional video – large enough to be used for instruction tuning. However their differences are very coarse-grained, for example 'cooking video A forgot to add salt'. Here, large-scale is possible because differences can be derived automatically from annotated instructional video datasets.
- (Balakrishnan et al., 2015) considers fine-grained action differences. However their dataset is very small (less than 50), and they have no labels.
- (Doughty et al., 2018) has a dataset of paired actions called EPIC-Skills2018. Here, the scale is large, but the difference label is more coarse: a binary for which video shows more skill.

F EVALUATION

The evaluation code is available at this anonymized GitHub repo https://anonymous.4open.science/r/VidDiffBench_eval-A0C1/README.md. The main file is `eval_viddiff.py`.

F.1 CLOSED EVALUATION DESIGN CHOICE

Our closed evaluation setting has options for 'A' and 'B', but not 'c'. Our initial approach to formulating this did include an option 'C' for insignificant differences. However, the challenge of calibration made fair evaluation difficult. For example, when comparing two videos of a basketball shot to evaluate stance width, the question arises: how different is "different enough" to be both relevant for skill learning and perceptible? Different annotators may apply varying thresholds for what constitutes a significant difference, leading to inconsistencies. Introducing option 'C' further complicates evaluation because it requires calibrating not only the human annotators but also the VLMs, which may have different internal thresholds for perceiving significance. To address these challenges, we adopted the following approach:

- Annotators were instructed to choose either 'A' or 'B' only when the difference was clearly perceptible.
- We limited the evaluation of VLMs to cases where there was a very clear ground truth answer of either 'A' or 'B'.

This method ensures fairness by focusing on scenarios with unambiguous ground truth, avoiding complications introduced by subjective calibration thresholds. While we briefly discuss this in the section on annotation creation, we recognize that this is a nuanced point.

F.2 Evaluation: matching with LLMs

As described in section 3.2, we use an LLM

We use 'gpt-4o-2024-08-06' with the following prompt for matching ground truth to predicted descriptions. We do one prompt per video pair sample.

```
You are analyzing videos of people performing a specific action
↪  described as "{action_description}."

In this task, a "difference" refers to how two people might perform the
↪  same action in distinct ways.
Each difference consists of:
- A name (key) and
- A description that explains how the two performances differ visually.

You are provided with two dictionaries:
- Dictionary 0: {differences0}
- Dictionary 1: {differences1}

Your task is to match the differences from Dictionary 0 to Dictionary 1.
↪  Here's how:
1. For each entry in Dictionary 0, find the best match in Dictionary 1.
2. Each item in Dictionary 1 can only be used once.
3. Only match entries if their description strings are visually similar,
↪  even if the word choices differ. If no suitable match exists, return
↪  "None."

Your output should be a JSON object where:
- The keys are from Dictionary 0: {dict0_keys}.
- The values are a key from Dictionary 1 or "None" if no match is found.
- The available keys from Dictionary 1 are: {dict1_keys}.

Example output format:
{
    "0": "3",
    "1": "None",
    "2": "1",
    "3": "5",
    ...
    "{final}" : "0"
}
Important: the keys in this dictionary should be" {dict0_keys}
```

We replace:

- `{action_description}` with a string describing the action

- {`differences0`} with a dictionary where keys are gt difference keys `("0","1",...` and values which are strings describing differences.
- {`differences1`} is the same as {`differences0`}, except for the predicting differences.
- {`dict_keys`} are the keys in {`differences0`}
- {`dict1_keys`} are the keys in {`differences1`}
- {`final`} is the highest key in {`differences0`}

An issue with matching is that the prediction actually has the opposite value to the ground truth – e.g. "the arms are more straight" vs "the arms are more bent". If this is the case, then the prediction 'a' should be reversed to 'b' and vise versa. To identify those cases, we use this prompt, and we evaluate them in batches of 6 difference sequences at once.

```
Task:
You will be given pairs of statements. Your task is to determine the
↪  logical relationship between each pair.

Instructions:
1. Analyze Each Pair: For each pair of statements, carefully analyze
↪  their meaning and relationship.
2. Categorization:
   - Return "0" if the statements are equivalent, very similar, or
   ↪  differ only in minor details.
     Example: "X is bigger than Y" and "X is larger than Y" should both
     ↪  return "0".
   - Return "1" if the statements are direct opposites in meaning.
     Example: "X is bigger than Y" and "X is smaller than Y" should
     ↪  return "1".
3. Edge Cases:
   - Avoid returning "1" for statements that are not true opposites,
   ↪  even if they have some differences in detail or degree.
     Example: "X is much bigger than Y" and "X is slightly bigger than
     ↪  Y" should still return "0".

Output Format:
- Your response should be a JSON object with a single key "results" and
↪  an array of string values "0" or "1" as its value.
- The array should exactly match the number of statement pairs given in
↪  the input.

Input Format:
- The list of statement pairs will be provided in the following format:

{statements}

Important Requirements:
- Ensure that each value in the output array is either "0" or "1".
- The length of the "results" array must exactly match the number of
↪  input pairs.
```

We replace the {`statements`} with a list, where each element is a two element list of strings, which are matched difference descriptions.

F.3 Prompts for LMM baselines

The LMM baselines – GPT, Gemini, and Qwen – all receive the same prompts. Prompts for action assignment to the easy/medium/hard splits:

```
Here are two videos of an action with the following description:
↪  "{action_description}".
{video_representation_description}
```

```
Below is a set of identified differences that describe how the action be
↪  performed differently.
Each difference is associated with a unique key:
{differences_annotated}

Your task is to predict, for each difference, whether it is more true
↪  for video 'a' or video 'b'.
{target_out}
```

The {action_description} is replaced with a string describing the action which is the same as in table 13. The {differences_annotated} is a dictionary mapping the ground truth difference key to a difference description, and are the same stringes used in table 14. The {video_representation_description} tells the model how the video data is passed in. If 2 videos – for Gemini and Qwen – then it's:

```
We have passed 'video a' and 'video b' as videos to the prompt in that
↪  order.
```

If passing in the videos as frames – for GPT – then it's

```
We have passed a sequence of images into the prompt.
The first {vid0_nframes} are video a. The last {vid1_nframes} are video
↪  b.
The frame rate is the same and is {fps}.
```

The open prompt is:

```
Here are two videos of an action with the following description:
↪  "{action_description}".
{video_representation_description}

Return a list of 'differences' in how the action is being performed.
Each difference should have a 'description' that is a specific
↪  statements that is more true in one video compared to the other
↪  video.
Then there is a 'prediction' which is 'a' if the statement applies more
↪  to video a and 'b' if it applies more to video b.

The difference descriptions should be visual and about how the action is
↪  performed.
For example 'the jump is higher', or 'the arm is more straight'.

The difference descriptions should not refer to a specific video.
For example you would not say 'the jump in video B is higher'.
Instead, the 'description' would be 'the jump is higher', and the
↪  'prediction' is 'b'.

Suggest no more than {n_differences} differences.


Return a json like this, replacing '...' with actual content:
{
    "0" :  {
             "description" : "....",
             "prediction" : "a|b"
          },
    "1" :  {
             "description" : "....",
             "prediction" : "a|b"
          },
```

Table 9: Agreement rate of LLM and human predictions for the evaluation matching.

	LLM	Human 1	Human 2	Human 3
LLM		72.4	74.0	70.1
Human 1	72.4		75.0	78.2
Human 2	74.0	75.0		73.9
Human 3	70.1	78.2	73.9	
Average	72.2	75.2	74.3	74.0

```
    ...
}
```

F.4 Validation of matching process

We leverage LLMs in open evaluation to identify matching between the ground truth difference description strings and the predicted differences. Here we validate that this is a reasonable approach.

Robustness to multiple LLM runs The LLM evaluation is robust to random seed. We repeated the evaluation five times with different random seeds and observed a standard deviation of only 0.7 in the final evaluation score. This indicates that the results are consistent across runs. Although the prompt was specifically engineered for the GPT-4o-2024-08-06 model, we ensured consistency by fixing the model for all evaluations, treating all comparisons under identical conditions.

Comparison with Human Evaluation To measure alignment with humans, we recruited 3 human annotators to perform open evaluation matching, each with 44 video pairs and 347 individual differences. For each video pair, they were provided with a list of ground truth differences, and asked to match each one to a predicted difference from a list, or to suggest no match. We calculated inter-rater agreement across annotators and the automated LLM system. The results are in table 9. We can see semantic matching proved to be challenging for humans – the mean of pairwise rater agreement from each human to the other humans was 75.7%. Meanwhile, the mean agreement between our automated system and human annotators was 73.9%. Therefore, our LLM-based approach is on par with human annotators, while being completely automatic.

Details of Prompt for LLM Evaluation The LLM prompt was carefully developed using a prompt engineering workflow. We selected a set of four evaluation samples, covering two actions and two models, and iteratively refined the prompt based on performance in individual runs. For example, we added the instruction: ”Only match entries if their description strings are visually similar, even if the word choices differ.” This adjustment was necessary because the LLM struggled to match equivalent descriptions phrased differently (e.g., “the feet stance is wider” vs. “the legs are spread wider apart”). While this approach achieved satisfactory results, we acknowledge that the prompt could be further optimized using more systematic methods, such as DSPy Khattab et al. (2024). Exploring such techniques is a promising direction for future work.

G Results: more analyses

G.1 Results: action-level model comparison

We have performed a more thorough comparison of the different state-of-the-art LMMs on VidDIff-Bench, added a small subsection to the results, and a discussion in appendix. Specifically we look at each action, and compare the different LMMs.

First, we show the correlations in the per-action scores between models in table 10

The correlations are generally low, but there are 3 clusters of models. LLaVA-Video and Qwen2-VL are in a cluster; they are both open-source, and have the same LLM backbone. Then GPT-4o and Claude-Sonnet cluster together, and Gemini is not similar to any other model. We can speculate that for video data, Claude and GPT have similar training strategies, while Gemini’s is different.

Table 10: Correlations between models where the data is the action-level accuracy.

	GPT	Gemini	Claude	LLava-Video	Qwen2-VL
GPT-4o		0.152	0.375	0.243	0.273
Gemini-1.5-Pro	0.152		0.215	0.111	0.223
Claude-3.5-Sonnet	0.375	0.215		0.261	0.220
LLaVA-Video	0.243	0.111	0.261		0.376
Qwen2-VL-7b	0.273	0.223	0.220	0.376	

Table 11: Action-level scores for each model, and their differences compared to the average model score for that action. The model names are abbreviated and the full model names are GPT-4o, Gemini-1.5-Pro, Claude-3.5-Sonnet, LLaVA-Video-7B, Qwen2-VL-7B

Split	**Action**	**Action Name**	**Count**	**Scores GPT**	**Gemini**	**Claude**	**LLaVA-Vid**	**Qwen2**	**Avg**
easy	fitness_0	Hip circle anticlockwise	129	56.6	58.1	56.6	51.9	45.0	53.6
easy	fitness_3	lucky cat	62	53.2	58.1	43.5	58.1	45.2	51.6
easy	fitness_4	Squat Knee Raise side view	43	65.1	69.8	37.2	69.8	55.8	59.5
easy	fitness_6	Hip circle clockwise	123	58.5	76.4	69.9	56.1	52.8	62.8
medium	ballsports_0	basketball jump shot	96	55.2	57.3	52.1	57.3	61.5	56.7
medium	ballsports_1	basketball mikan layup	148	56.8	49.3	46.6	51.4	56.1	52.0
medium	ballsports_2	basketball reverse layup	125	46.4	55.2	49.6	44.0	50.4	49.1
medium	ballsports_3	soccer penalty kick	70	51.4	60.0	57.1	70.0	54.3	58.6
medium	diving_0	diving	240	53.8	52.1	53.3	50.8	54.2	52.8
medium	fitness_1	Opening and closing step left side first	186	57.5	54.8	52.7	51.1	51.6	53.5
medium	fitness_2	car deadlift	137	55.5	62.8	62.0	47.4	54.0	56.4
medium	fitness_5	Squat Knee Raise diagonal view	70	35.7	32.9	38.6	62.9	44.3	42.9
medium	fitness_7	Opening and closing step right side first	155	52.9	52.9	63.2	49.7	52.9	54.3
hard	music_0	piano	94	51.1	51.1	58.5	51.1	52.1	52.8
hard	music_1	guitar	20	55.0	40.0	45.0	50.0	65.0	51.0
hard	surgery_0	Knot Tying	237	47.7	43.5	43.5	46.4	44.3	45.1
hard	surgery_1	Suturing	309	48.9	51.5	48.2	47.6	50.2	49.3
hard	surgery_2	Needle Passing	211	51.7	46.9	49.8	48.8	50.2	49.5

Next we compare model performance within one action, and this is two large tables. table 11 is the action-level performance of each model. Then table 12 is the 'relative performance': the difference between the model score on that action compared to the mean score across all models for the action. The most significant results in the benchmark are on the easy split. Here, the improvement in score is uniform for all models. The models generally close perform similarly each other. The relative performance is usually less than 10 points – when it is higher, the sample size is very small.

By comparing models at the level of actions, we are considering smaller sample sizes than in the main results, which compare models at the level of easy/medium/hard splits. There is therefore lower statistical power to identify significant result differences, so the results are less certain. We elected not to compare model performance at the level of action differences, because here the sample sizes are very small, so any correlations would not meet significance thresholds.

G.2 Detailed difference analysis

In Section 6.3, we discuss an analysis of the accuracy at the difference level. The vary long table 14 gives the per-difference accuracies and p-values compared for the accuracy against a random guessing baselines. Each difference is associated with an action key, whose description is in table 13.

Table 12: Action-level difference scores for each model relative to the mean model score on that action. This is the difference with respect to the table 11. The model names are abbreviated and the full model names are GPT-4o, Gemini-1.5-Pro, Claude-3.5-Sonnet, LLaVA-Video-7B, Qwen2-VL-7B

Split	Action	Action Name	Count	Differences				
				GPT	Gemini	Claude	LLaVA-Vid	Qwen2
easy	fitness_0	Hip circle anticlockwise	129	2.9	4.5	2.9	-1.7	-8.7
easy	fitness_3	lucky cat	62	1.6	6.5	-8.1	6.5	-6.5
easy	fitness_4	Squat Knee Raise side view	43	5.6	10.2	-22.3	10.2	-3.7
easy	fitness_6	Hip circle clockwise	123	-4.2	13.7	7.2	-6.7	-9.9
medium	ballsports_0	basketball jump shot	96	-1.5	0.6	-4.6	0.6	4.8
medium	ballsports_1	basketball mikan layup	148	4.7	-2.7	-5.4	-0.7	4.1
medium	ballsports_2	basketball reverse layup	125	-2.7	6.1	0.5	-5.1	1.3
medium	ballsports_3	soccer penalty kick	70	-7.1	1.4	-1.4	11.4	-4.3
medium	diving_0	diving	240	0.9	-0.8	0.5	-2.0	1.3
medium	fitness_1	Opening and closing step left side first	186	4.0	1.3	-0.9	-2.5	-1.9
medium	fitness_2	car deadlift	137	-0.9	6.4	5.7	-8.9	-2.3
medium	fitness_5	Squat Knee Raise diagonal view	70	-7.1	-10.0	-4.3	20.0	1.4
medium	fitness_7	Opening and closing step right side first	155	-1.4	-1.4	8.9	-4.6	-1.4
hard	music_0	piano	94	-1.7	-1.7	5.7	-1.7	-0.6
hard	music_1	guitar	20	4.0	-11.0	-6.0	-1.0	14.0
hard	surgery_0	Knot Tying	237	2.6	-1.6	-1.6	1.4	-0.8
hard	surgery_1	Suturing	309	-0.4	2.2	-1.0	-1.7	0.9
hard	surgery_2	Needle Passing	211	2.2	-2.6	0.3	-0.7	0.8

Table 13: Actions keys and their descriptions

Action	Action description
ballsports_0	a person is doing a basketball mid-range jump shot, starting with the ball in their hand, no defense, practice only
ballsports_1	a person does the Basketball drill called the Mikan layup where they start under the basket, do a layup with the right hand and catch it, do a left hand layup and catch it, no defense, practice only
ballsports_2	a person playing basketball does a reverse layup starting from the left side of the basket and lays it up with their right hand on the right hand side, no defense, practice only
ballsports_3	a person does a soccer drill where they do a single penalty kick, practice only, no defense, no goalie
diving_0	competitive diving from 10m
fitness_0	fitness exercise called standing hip circle with hands on hips, one rotation anti-clockwise
fitness_1	fitness exercise called opening and closing step on left side and then opening and closing step on right side
fitness_2	a single free weight deadlift without any weight
fitness_3	fitness exercise called two arm standing lucky cat starting with arms up, one repetition
fitness_4	squat without weights, then knee raise on left side
fitness_5	a squat then a knee raise on left side
fitness_6	fitness exercise called standing hip circle with hands on hips, one rotation clockwise
fitness_7	fitness exercise called opening and closing step on right side and then opening and closing step on left side
music_0	a person is playing scales on the piano
music_1	a person is playing scales on the guitar
surgery_0	The subject picks up one end of a suture tied to a flexible tube attached at its ends to the surface of the bench-top model, and ties a single loop knot.
surgery_1	The subject picks up needle, proceeds to the incision (designated as a vertical line on the bench-top model), and passes the needle through the fake tissue, entering at the dot marked on one side of the incision and exiting at the corresponding dot marked on the other side of the incision. After the first needle pass, the subject extracts the needle out of the tissue, passes it to the right hand and repeats the needle pass three more times.
surgery_2	The subject picks up the needle (in some cases not captured in the video) and passes it through four small metal hoops from right to left. The hoops are attached at a small height above the surface of the bench-top model.

Table 14: Difference-level accuracy scores for VidDiff. The 'action' values can be looked up at table 13. The grayed columns indicate a p-value < 0.05 for the two-tailed binomial significance test

Split	Action	Difference Description	Mean score	Num Samples	p-value
easy	fitness_6	the head remains more vertical during the rotation	1	13	0
medium	fitness_2	the gaze is forward at the bottom of the deadlift	1	7	0.016
medium	ballsports_2	they gather the ball with both hands	1	2	0.5
hard	music_1	The player uses a plectrum.	1	4	0.125
medium	fitness_7	the motion is faster	0.93	14	0.011
medium	ballsports_1	uses the non-shooting hand (the guide hand) for stabilizing the ball during the shot in the 1st shot	0.9	10	0.02
medium	fitness_2	the knees bend less at the bottom of the deadlift	0.89	19	0.004
easy	fitness_6	the toes are more pointed out	0.88	16	0.004
medium	ballsports_1	they jump higher on the first shot	0.83	12	0.107
easy	fitness_6	the feet stance is wider	0.83	24	0.005
medium	fitness_2	the feet stance is wider	0.8	15	0.028
easy	fitness_3	the feet stance is wider	0.8	5	0.313
easy	fitness_0	the feet stance is wider	0.8	25	0.003
medium	ballsports_1	uses the non-shooting hand (the guide hand) for stabilizing the ball during the shot in the 2nd shot	0.8	10	0.088
medium	ballsports_0	the shooter's arm is more extended towards the basket	0.79	14	0.122
medium	fitness_7	the arms are elevated in an uneven way on the first step	0.75	20	0.03
hard	surgery_1	The left graser supports the right grasper, by pressing down on the tissue.	0.75	4	0.5
medium	ballsports_0	as the shooter begins extending the elbow to shoot, the non-shooting hand (the guide hand) is on the side of the ball and does not influence the balls trajectory	0.75	12	0.107
medium	ballsports_3	they kick the ball harder	0.75	12	0.107
medium	ballsports_2	the non-jumping leg has a more elevated knee	0.73	15	0.183
medium	fitness_2	the gaze is forward at the start of the motion	0.73	11	0.322
hard	surgery_2	The instrument tips are never out of view (occluded by instruments, or out of frame)	0.73	11	0.322
medium	fitness_7	the arms reach higher on the first step	0.71	17	0.189
hard	surgery_2	The second grasper is used to stabalize the target.	0.69	26	0.093
medium	fitness_2	the hands are lower at the bottom of the deadlift	0.68	19	0.192
easy	fitness_3	the upper arms are parallel to the ground at the start of the motion	0.68	19	0.192
medium	fitness_1	the torso moves out further closer to the foot during the second step out	0.67	21	0.194
easy	fitness_6	the range of motion in the hips is larger	0.67	24	0.156
easy	fitness_6	The upper body rocks more in a forward-backward way	0.67	9	0.492
medium	fitness_2	the body is more locked out at the top of the deadlift	0.67	6	0.625
medium	fitness_1	the arms reach higher on the first step	0.67	18	0.243
medium	ballsports_2	the ball is closer to the corner of the square on the backboard	0.67	9	0.492
easy	fitness_0	The upper body rocks more in a forward-backward way	0.67	18	0.243
medium	ballsports_3	they rotate their hips more during the strike	0.67	18	0.243
medium	ballsports_3	the head is facing the ball leading up to the strike	0.67	6	0.625
hard	music_0	Rhythmic consistency is better maintained in Video A than in Video B.	0.67	12	0.387
easy	fitness_0	the toes are more pointed out	0.65	17	0.297
medium	diving_0	The size and volume of the splash created upon entry is larger for video A than video B.	0.65	48	0.052
easy	fitness_6	the speed of hip rotation is faster	0.64	25	0.122
medium	ballsports_1	they jump with two feet on the 1st shot	0.64	11	0.451
medium	diving_0	Diver enters the water at an angle closer to 90 degrees in video A than in video B.	0.63	30	0.161
medium	fitness_1	the arms are elevated in an uneven way on the second step	0.63	19	0.192
easy	fitness_0	the range of motion in the hips is larger	0.63	19	0.192
hard	surgery_0	The tube in Video A moves more than in Video B.	0.63	27	0.194
easy	fitness_4	the squat is deeper, measured by angle of the thigh to the ground	0.63	16	0.244
easy	fitness_3	the toes are more pointed out	0.63	8	0.438
medium	fitness_7	the toes are more pointed outwards on the first step out	0.63	8	0.438
hard	music_0	Forearm movement is more controlled and minimal in Video A than in Video B.	0.62	13	0.419
medium	fitness_1	the second step out is wider	0.62	13	0.419
medium	fitness_1	the arms reach higher on the second step	0.61	18	0.334
easy	fitness_0	the head remains more vertical during the rotation	0.61	18	0.334
medium	ballsports_1	the non-jumping leg has a more elevated knee in the 2nd shot	0.6	5	0.625
medium	ballsports_0	the shooter's jump is more vertical than forward	0.6	5	0.625
medium	ballsports_2	the gaze is more up and forward instead of down on the 2nd last step	0.6	10	0.41
medium	ballsports_1	the arm is more fully extended towards the basket in the follow through in the 2nd shot	0.6	15	0.305
medium	ballsports_0	the shooter's feet stance is wider when starting the shooting motion	0.6	10	0.41
medium	ballsports_1	the body moves more forward during the 2nd shot, rather than up or back	0.6	10	0.41
medium	ballsports_3	the non-kicking foot is planted closer to the ball	0.6	10	0.41
hard	surgery_0	The tension on the suturing material and the tissue is better controlled in Video A than in Video B.	0.59	32	0.162
easy	fitness_6	the hand position is higher on the body	0.58	12	0.451
medium	ballsports_3	the non-kicking foot is planted more next to the ball and less behind the ball	0.58	12	0.451
medium	ballsports_1	they catch the ball in a higher position on the 1st shot	0.58	12	0.451
easy	fitness_0	the hand position is higher on the body	0.58	12	0.451
hard	music_0	The speed of playing is higher in Video A than in Video B.	0.58	19	0.352
hard	surgery_1	The instrument tips are never out of view (occluded by instruments, or out of frame)	0.58	19	0.352
hard	surgery_1	The suturing speed is higher in Video A than in Video B	0.58	52	0.157
hard	surgery_2	The movement of the needle through the hoop is more radial in Video A than in Video B.	0.58	26	0.288
hard	surgery_1	The tension on the suturing thread is lower in Video A than in Video B.	0.57	28	0.279
medium	fitness_7	the arms are elevated in an uneven way on the second step	0.56	16	0.349
medium	fitness_1	the arms are elevated in an uneven way on the first step	0.56	16	0.349
hard	surgery_2	The number of movements to arrange the needle before threading is lower in Video A than in Video B.	0.56	41	0.223
hard	surgery_1	The movement is more fluid in Video A than in Video B.	0.56	43	0.218
hard	surgery_1	The grasper in Video A is more quickly positioned on the needle than in Video B.	0.56	43	0.218
medium	fitness_7	the toes are more pointed outwards on the second step out	0.56	9	0.492
medium	diving_0	Diver rotates forward relative to themselves.	0.56	27	0.299
medium	ballsports_2	they get deeper knee bend before jumping	0.56	9	0.492
easy	fitness_0	the speed of hip rotation is faster	0.55	20	0.32
medium	ballsports_1	they jump higher on the second shot	0.55	11	0.451
easy	fitness_4	the toes are more pointed out	0.54	13	0.419
medium	ballsports_0	the shooter's feet position are more staggered, meaning the feet are at a different distance from the basket	0.54	13	0.419

medium	diving_0	Duration from jump off the board to water entry in longer in video A than in video B.	0.54	39	0.251
medium	ballsports_0	the shooter's knees are more bent before taking the shot	0.54	13	0.419
medium	ballsports_2	they use the non-shooting hand (the guide hand) for stabilizing the ball during the shot	0.53	17	0.371
medium	fitness_5	the squat is deeper, measured by angle of the thigh to the ground	0.53	17	0.371
hard	surgery_2	The needle is grasped closer to the tip in Video A than in video B.	0.52	21	0.336
hard	surgery_2	The force on the target is lower in Video A than in Video B	0.51	37	0.257
medium	diving_0	Diver's body is more straight in video A than in video B.	0.51	39	0.251
medium	ballsports_1	they have better balance when landing on the 1st shot	0.5	2	1
medium	diving_0	Speed at which divers rotate during the dive in larger in video A than in video B.	0.5	38	0.257
medium	fitness_7	the arms reach higher on the second step	0.5	14	0.419
medium	ballsports_1	they have better balance when landing on the 2nd shot	0.5	2	1
hard	music_1	Smooth transitions between strings with minimal disruption to the rhythm or tempo. Transitions in Video A are smoother than in Video B.	0.5	2	1
hard	music_1	Guiatrist uses finger vibrato.	0.5	2	1
medium	ballsports_3	the body (or torso) is more facing the net, or more 'square' to the net	0.5	12	0.451
medium	fitness_7	the first step out is wider	0.47	17	0.297
hard	surgery_0	The movements in Video A are more precise than in Video B.	0.47	34	0.216
easy	fitness_3	the speed of the arms is faster	0.47	17	0.297
medium	ballsports_2	they land on two feet	0.47	17	0.297
medium	ballsports_2	they follow through more and towards the basket	0.47	15	0.305
medium	fitness_2	the toes are more pointed out	0.47	15	0.305
medium	fitness_2	there is a pause at the bottom of the deadlift	0.47	15	0.305
medium	fitness_7	the second step out is wider	0.46	13	0.314
hard	surgery_1	The needle is inserted in the fabric more perpendicular to the incision.	0.46	13	0.314
medium	fitness_1	the toes are more pointed outwards on the first step out	0.46	13	0.314
hard	music_0	The smoothness of thumb crossing is more evident in Video A than in Video B.	0.46	13	0.314
easy	fitness_3	the upper arms more stable through the entire motion	0.46	13	0.314
medium	ballsports_1	they jump off the right foot for right-hand shot (the 2nd shot)	0.45	11	0.322
medium	ballsports_2	before raising the ball to shoot, the ball is more to the right side of the hip	0.45	11	0.322
hard	surgery_1	The force is applied in a more radial way in Video A than in Video B.	0.45	31	0.192
medium	fitness_7	the torso moves out further closer to the foot during the first step out	0.44	18	0.243
medium	fitness_2	the arms are in front of the body at the bottom of the deadlift	0.44	9	0.328
hard	surgery_2	The passage of the needle between two hands is more fluid in Video A than in Video B.	0.44	25	0.266
medium	fitness_5	the feet stance is wider	0.44	16	0.349
hard	music_0	The wrist should be straight and not dipped or raised, facilitating fluid motion and avoiding strain. The wrist position is more appropriate in Video A than in Video B.	0.43	14	0.244
medium	fitness_2	the entire motion is faster	0.43	21	0.194
hard	surgery_0	The movements in Videos A are faster than in Video B.	0.43	42	0.116
hard	surgery_0	More errors are corrected in Video A than in Video B.	0.42	31	0.131
hard	surgery_2	The thread is more efficiently managed in Video A than in Video B.	0.42	24	0.156
medium	fitness_1	the toes are more pointed outwards on the second step out	0.41	17	0.189
hard	surgery_1	The instrument applies more force to the tissue and needle in Video A than in Video B	0.41	44	0.078
hard	surgery_0	The movements in Video A are more efficient than in Video B.	0.4	42	0.076
medium	ballsports_1	the body moves more forward during the 1st shot, rather than up or back	0.4	5	0.625
medium	ballsports_0	the shooter's feet are oriented more square to the basket when starting the shooting motion, meaning the feet point more forward	0.38	13	0.175
medium	ballsports_0	as the shooter begins extending the elbow to shoot, the ball is more in front of the body, rather than behind the head	0.38	16	0.244
hard	surgery_1	The dot is more accurately hit in Video A than in Video B	0.36	14	0.122
hard	music_0	The body is closer to the piano in Video A than in Video B.	0.35	17	0.094
medium	fitness_5	the speed of the whole motion is faster	0.35	17	0.094
medium	ballsports_2	they release the ball at a higher position	0.35	20	0.148
medium	ballsports_1	the non-jumping leg has a more elevated knee in the 1st shot	0.33	9	0.141
hard	surgery_0	Both graspers are used efficiently.	0.33	9	0.141
hard	surgery_0	The surgeon in Video A stops more often to plan next steps than the surgeon in Video B.	0.33	3	0.25
hard	music_0	There are more wrong note corrections in Video A than in Video B.	0.33	6	0.188
medium	fitness_1	the motion is faster	0.33	18	0.065
medium	ballsports_1	they have more fluid motion in moving between the shots	0.33	12	0.107
hard	music_1	The left fingers in Video A are more curved / less collapsed than in video B.	0.33	3	0.25
medium	fitness_7	the torso moves out further closer to the foot during the second step out	0.33	9	0.141
medium	diving_0	Diver faces the water at jump off.	0.32	19	0.044
medium	fitness_1	the first step out is wider	0.31	16	0.133
medium	fitness_5	during the squat descent, the knees cave inwards, instead of tracking over the feet	0.3	20	0.074
medium	fitness_1	the torso moves out further closer to the foot during the first step out	0.29	17	0.036
hard	surgery_1	The grasper grasps the needle approximately 2/3 from the needle tip. The needle is grasped more precisely in Video A than in Video B.	0.29	34	0.006
easy	fitness_4	the feet stance is wider	0.29	14	0.044
hard	music_1	The unused left finger tips in Video A stay closer to the board than in video B.	0.25	4	0.5
hard	surgery_0	The suturing thread tangles.	0.24	17	0.01
medium	ballsports_1	the arm is more fully extended towards the basket in the follow through in the 1st shot	0.18	11	0.011
hard	music_1	Fingers should press strings at the center of the frets, avoiding the metal fret bars for clear sound production. Video A shows more accurate finger placement on the fretboard than Video B.	0	2	0.5
hard	music_1	Only one finger of the left hands rests on a string at a time.	0	3	0.25

Table 15: Evaluating 'easy' split with variable video fps for three models. Our evaluation protocol chooses 4fps.

	1 fps	2 fps	4 fps	8 fps	average
GPT-4o	58.0	59.4	58.8	59.10	58.8
Gemini-1.5-Pro	59.7	66.9	65.8	66.9	64.8
Claude-3.5-Sonnet	58.1	58.5	56.6	52.9	56.5

G.3 Results: dependence on FPS

The frame sampling rate, fps, is an important consideration for evaluating fine-grained actions. While typical video benchmarks like Video-MME Fu et al. (2024) sample videos at 1fps, we have sampled at a higher rate depending on category. The categories with shorter videos were sampled at a higher rate: 4fps for 'fitness', 5fps for 'ballsports', and 6fps for 'diving' (they are slightly different so they can be compatible with fps in the source dataset). We chose this relatively higher rate because we are interested in more fine-grained differences, while prior benchmarks are more coarse-grained; however we did not sample at even higher due to practical cost constraints of processing too many frames. The longer videos 'surgery' and 'music' were sampled at 1fps: these are longer videos where differences are discernible at lower sampling rates, and where the longer videos make high-fps sampling impractical.

To show that our fps is reasonable, we tested the three closed-source models on a range of fps levels on the 'easy' subset of closed evaluation. We chose this set because this is where statistically significant differences were clear. The results are in table 15.

Across all models, the sampling rate that we use, 4 fps, has reasonable scores, either at or above the average over the other fps values. For all models, the variability is low: GPT's scores are within 0.8 points of the average; all other models have scores within 2.1 points of the average (except for the low sampling rate of 1fps in Gemini, where it degrades by 5.2 points). Moreover, the optimal fps is different for different models.

To help explain the results, we refer to the qualitative examples in the main results sections. The only 'success cases' for all our models were those having easy localization, and coarse differences. We hypothesize that fps is not important for these cases. Where fps is likely important -– fine-grained multiframe reasoning – the current LMMs cannot perform better than random. So although 2fps currently has good performance, we believe that as LMMs improve, they will perform better on subtle motions and using a higher fps will be important.

G.4 Results: Qwen2-VL open evaluation

Qwen2-VL performs especially poorly in open evaluation, which we investigate here. The key issue is that Qwen2-VL-7b was failing to follow the evaluation prompt, while the other compared models did follow it. We sampled 3 video pairs for each action and manually inspected Qwen's responses, identifying multiple key issues. Below, we list each issue, and provide a quantitative estimate for the prevalence of each issue.

- (45% of differences) Proposing differences not relevant to *how* to perform actions, but instead are visual things like "The person in video a is wearing a blue jacket, while the person in video b is wearing a plaid shirt." We estimated prevalence by using a gpt-4o query that we manually prompt engineered.
- (26% of differences) Proposing a difference that is actually not a difference, e.g. "The person in video a is performing the exercise with their arms out to the sides, while the person in video b is performing the exercise with their arms out to the sides." We estimated prevalence by using a gpt-4o query that we manually prompt engineered.
- (56% of differences) are repeated, meaning when trying to propose multiple differences, it proposes the same difference multiple times. We could directly measure this prevalence exactly.

- (23% of actions) Proposing only a small number of differences – less then half as many as what is prompted for. We could directly measure this prevalence exactly.
- ($\leq$5% of differences) Proposing vague differences that are harder to interpret visually like "The player in video a has a more versatile and adaptable skill set than the player in video b". We estimated prevalence by using a gpt-4o query that we manually prompt engineered.

Overall, only 31.9% of proposed differences by Qwen did not suffer from any of these errors. (Note that some differences suffered from multiple errors at the same time)

G.5 No multiple choice bias in closed evaluation

In multiple choice benchmarking, models may be biased towards one particular option, which can impact evaluation robustness. We find no evidence of this. Firstly, in closed evaluation, the A/B ratio is 0.493/0.507. Second, we test the impact of video order on GPT-4o for the 'fitness' category, which has samples in the easy and medium subsets(sample size 193). We test flipping the order of videos which flips the A/B answer. The performance is 54.8% in the original evaluation, and reversing the order of videos gives performance of 55.5%, showing a 0.7% difference. This result suggests that the performance on VidDiffBench is not significantly sensitive to video order.

G.6 Experiment on duplicating video

One idea to validate the reasonable-ness of the benchmark is to check what happens when passing an identical video as A and B to the system – we should expect that in closed evaluation, the predictions should be A/B 50% of the time. We did this experiment on the closed setting, for the 'easy' subset for GPT-4o. Over two random seeds, the results were 49.3 and 50.2. This is an interesting validation check that the benchmark passes.

G.7 No multiple choice bias in closed evaluation

In multiple choice benchmarking, models may be biased towards one particular option, which can impact evaluation robustness. We find no evidence of this. Firstly, in closed evaluation, the A/B ratio is 0.493/0.507. Second, we test the impact of video order on GPT-4o for the 'fitness' category, which has samples in the easy and medium subsets(sample size 193). We test flipping the order of videos which flips the A/B answer. The performance is 54.8% in the original evaluation, and reversing the order of videos gives performance of 55.5%, showing a 0.7% difference. This result suggests that the performance on VidDiffBench is not significantly sensitive to video order.

G.8 Sample retrievals

In our methods, we leverage a frame localizer to find either a single frame or a small sequence of frames. These localized frames are then passed to the next stage. In fig. 4, we show a few examples of predictions vs ground truth for some frames

H VidDiff method

The code for VidDiff method is available at this anonymized GitHub repo https://anonymous.4open.science/r/VidDiffBench_eval-A0C1/README.md under the folder`viddiff_method.py`.

Here we show the prompts used in the different components.

Proposer stage . Part 1 chooses candidate differences (Open setting only):

```
I have two videos of an action with the following description:
↪  "{action}".

Propose a set of 'differences' in how this action may be performed
↪  between the two videos.
```

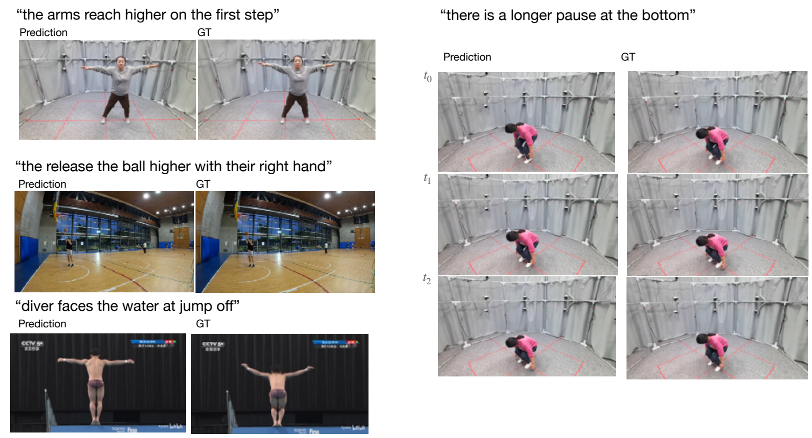

Figure 4: Sample frame localizations: prediction vs ground truth.

```
For each difference, give a 'name', 'description', 'query_string', and
↪  'num_frames'.

The 'name' is a very short name that can be used as a key.

The 'description' is a slightly longer description of the difference.
The 'query_string' is the same as 'description'.
The  descriptions should be visual and about how the action is
↪  performed. For example 'the jump is higher', or 'the arm is more
↪  straight'.
The difference descriptions should not refer to a specific video. For
↪  example you would not say 'the jump in video B is higher'.
Instead, the 'description' would be 'the jump is higher', and the
↪  statemet could be more true of one video or the other.

Now suppose you had to judge whether the difference was stronger in one
↪  video, but you could only look at individual frames.
        - What's the smallest number of frames you need, assuming the
        ↪  frame is well-chosen? Put the answer in 'num_frames'. The
        ↪  answer should be '1' or 'gt_1' (meaning 'greater than 1').
        - Once you have the frames to compare, create a 'query_string'
        ↪  that is a simple statement about this difference that is
        ↪  more true in one video vs the other video (based on the
        ↪  frames only). For example "the left leg is higher" or
        ↪  "movement is faster".

List {n_differences} differences.
Return a json like this, where the differences keys are stringified ints
↪  starting from 0.
{
        '0' : {
                "name" : "...",
                "description" : "...",
                "query_string" : "...",
                "num_frames": "1|gt_1",
        },
        ...
}
```

Proposer part 2 estimates sub-action stages:

```
I have two videos of an action with the following description:
↪  "{action}".

Provide a 'stage transcript' as a list. These are sub-actions that make
↪  up that action.
Give 5 steps or fewer in the action transcript.

For each stage, give a 'name' for the stage, and a 'description' of that
↪  stage.

For each stage, give a list of 'retrieval_strings'.
These are strings that describe what is visible in the frame.
Only describe the visual features. Only describe what is visible in a
↪  single frame. Focus on appearance. Focus on pose. Do not use the
↪  name of the action. Start each string with something similar to "A
↪  photo of a ...".
Give at least {n_retrieval_keys} retrieval strings per stage.

Return a json like this:
{
        "stages" : [
                {
                "name : "",
                "description" : "...",
                "retrieval_strings" : ["A photo of a ...", ...],
                },
                ...
]}
```

And proposer part 3 does linking between those stages

```
I have two videos of an action with the following description:
↪  "{action}".

Here are a list of stages or subactions that make up that action:
{stages}

We also have differences in how this action may be performed between the
↪  two videos.
The differences are specific statements that are more true in one video
↪  vs another.
Here they are:
{differences}

Now we need to match each differences to a stage.
Return a list of the stages using their names.
If a difference is relevant to a particular stage, put its name in the
↪  'difference' list.
It's okay for a 'difference' to be visible in multiple stages.
It's okay for some stages to have no difference.
Refer to stages and differences by their 'name' attribute.

Return a json like this:
[
        "<stage_name0>" : ["<difference_name0>", "<difference_name1>",
        ↪  ...],
        "<stage_name1>" : [],
        "<stage_name2>" : ["<difference_name1>", "<difference_name2>",
        ↪  ...],
        ...
]
Please be careful. Every difference must appear at least once
```

Frame Differencer

```
I have two videos of people performing an action with description:
↪  "{action}".
The first {num_frames} frames are from video A and the last {num_frames}
↪  frames are from video B.
For each video, the frames are very close together in the video: they
↪  are {time_diff} seconds apart.
Which one shows more of the variation with this description:
↪  "{query_string}"?
(a) video 1, (b) video 2, (c) similar or can't tell.
Answer in json:  {'answer_detailed' : "...", 'answer':"a|b|c"}""",
```

I COMPUTATIONAL COSTS FOR VIDDIFF METHOD

Our method's runtime is less than one minute per video pair using an A6000 GPU for running CLIP inference Radford et al. (2021). Additionally, we utilize the GPT API at an average cost of $0.2 per sample Achiam et al. (2023). Notably, over 95% of the GPT cost arises from verbose VLM responses. Attempts to prompt for shorter responses resulted in degraded performance.

Methodologically, we rely on pre-trained zero-shot models, which limits their applicability in specialized domains, as discussed in our results. For evaluation, the Open setting formulation necessitates an LLM in the evaluation pipeline. One challenge is the subjectivity in annotations: determining which differences are relevant and what magnitude of difference is significant, though we thoroughly discuss this problem and mitigations.