

Black Swan: Abductive and Defeasible Video Reasoning in Unpredictable Events

Supplementary Material

Please find two videos, `dataset.mp4` and `examples.mp4`, that visually describe the dataset creation process and several qualitative examples respectively. Disclosure: An AI generated voice was used to do the voiceover for both these videos.

A. Reasoning Types

Task	Abductive	Defeasible	Commonsense
Forecaster-Gen	×	×	✓
Detective-Gen	✓	×	✓
Detective-MCQ	✓	✓	✓
Detective-Y/N	✓	✓	✓
Reporter-Gen	×	×	✓
Reporter-MCQ	×	✓	✓
Reporter-Y/N	×	✓	✓

Table 9. Types of reasoning in BlackSwanSuite.

As shown in Table 9, our benchmark evaluates three fundamental reasoning capabilities through carefully structured tasks. **Detective** assesses *abductive reasoning*, requiring models to infer the most plausible cause of post-events (V_{post}) given pre-event contexts (V_{pre}).

Both **Detective** and **Reporter** incorporate *defeasible reasoning*, presented in two complementary formats: (1) **Multiple-choice questions (MCQ)**, which assess hypothesis selection through comparative analysis, and (2) **Yes/No (Y/N) validation**, which requires direct evaluation of a hypothesis when new context emerges. While related, these formats demand distinct reasoning skills—MCQs enable relative comparisons between alternatives, whereas Y/N validation necessitates absolute judgments about specific hypotheses under evolving video contexts.

To complete the evaluation spectrum, our *generation tasks* (**Forecaster-Gen** and **Reporter-Gen**) assess open-ended prediction of unexpected events, addressing a critical gap in existing video reasoning benchmarks. Collectively, all tasks evaluate *temporal processing* and *visual commonsense*, particularly the ability to interpret and anticipate atypical events.

With **Forecaster**, **Detective**, and **Reporter** defined, we next describe our data collection process for constructing generative (**Gen**), multiple-choice (**MCQ**), and yes/no validation (**Y/N**) questions.

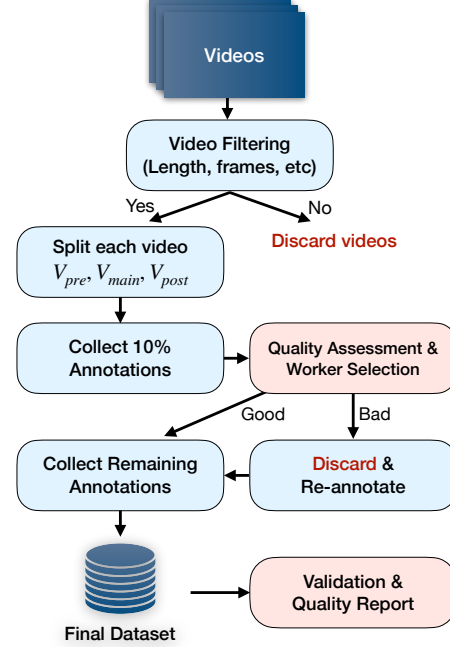


Figure 4. **Data Collection Process.** We start by filtering videos that adhere to our dataset requirements, such that they can be split into V_{pre} , V_{main} and V_{post} . With 10% of data, we collect annotations to select the best annotators. With these annotators, we collect the full dataset, and report dataset quality.

B. Data Collection and Annotations

B.1. Splitting Videos and Collecting Annotations

To automatically split a video V into its parts, V_{pre} , V_{main} , and V_{post} , we use an automatic scene splitter to clean the video clip, and use heuristics described in Algorithm 1 to perform the cut, based on the main event time, t . Note that we acquire both the video V and the main event time t from the Oops! dataset [7].

First, to ensure that the video V does not contain multiple different scenes, we use PySceneDetect’s AdaptiveDetector² with window width set to 5 frames to identify the scenes in the video. In case 4 or more scenes are found, we discard the video. If 2 or 3 scenes are found, and if the scene change is in the beginning and/or the end of the video, we trim the beginning and/or the end. Moreover, a change

²https://www.scenedetect.com/docs/latest/api/detectors.html#scenedetect.detectors.adaptive_detector.AdaptiveDetector

that coincides with the event time t can occur when major changes to the scene happen as a part of the surprising event, e.g. a light bulb going off, making the entire scene dark, is allowed. Following these steps, we get a cleaned-up and trimmed video V^* .

Algorithm 1 Video splitting

Input: video V^* , main event time t

Output: the three parts of the video $V_{pre}, V_{main}, V_{post}$

```

1: // Trim the start and end of the video, in case there are
   remnants of adjacent scenes.
2:  $V^* \leftarrow \text{trim}(V^*, 0.17 \text{ sec}, \text{start})$ 
3:  $V^* \leftarrow \text{trim}(V^*, 0.17 \text{ sec}, \text{end})$ 
4: // Cut the video into parts
5:  $V_{pre} \leftarrow [0, 0.8 \cdot t]$ 
6:  $V_{main} \leftarrow [0.8 \cdot t, 0.8 \cdot \text{length}(V^*)]$ 
7:  $V_{post} \leftarrow [0.8 \cdot \text{length}(V^*), \text{length}(V^*)]$ 
8: // Discard videos that are shorter than 1 second
9: if any( $\text{length}(v) < 1 \text{ sec} \forall v \in \{V_{pre}, V_{main}, V_{post}\}$ )
   then
10:   return Null
11: end if
12: return  $V_{pre}, V_{main}, V_{post}$ 

```

The algorithm receives the cleaned-up video and the average event time based on the annotations. To prevent parts of the previous or next video clip to be visible (due to the window width set to 5 frames), we trim the ends of the video by 0.17 seconds (lines 1-3). We then cut V^* into its three parts based on the mean event localization time t , such that V_{pre} ends shortly before t , V_{post} is the last 20% of the video, and V_{main} shows the main event (lines 4-7). Finally, we discard any videos for which at least one of the 3 parts is less than 1 second long (lines 8-11). This ensures that there is enough content in each part of the event. Through empirical analysis, we found that this method generally yields $V_{pre}, V_{main}, V_{post}$ where V_{pre} doesn't reveal exactly what unexpected event is about to happen, but allows for reasonable guesses (ideal for defeasible reasoning), and V_{post} only shows the outcome of the event, allowing for abductive reasoning.

Finally, we collect annotations using the template presented in Figure 9. We have the following conditions to participate as an annotator for our task:

- Location: Ireland, Canada, New Zealand, United Kingdom, United States, Australia
- Age: 18-65 year old
- Education: Bachelor's degree (for example: BA, AB, BS), Master's degree (for example: MA, MS, MEng, MEd, MSW, MBA), Professional degree (for example: MD, DDS, DVM, LLB, JD), Doctorate degree (for example: PhD, EdD)

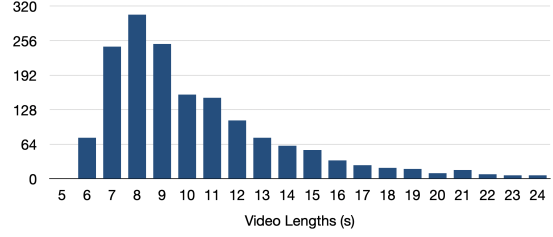


Figure 5. **Length of Videos.** The median video length is 8.83 seconds. Only a small number of videos are outliers, with 29 of them being longer than 25 seconds.

In each step of the process, annotators must write at least 8 words, and all answers are required and cannot be skipped.

B.2. Data Quality Validation

We ask two students (experts) from the lab, who did not contribute to this project in any other way, to independently verify 60 randomly sampled annotation instances from our dataset. Experts are first explained all the tasks, and are provided with all parts of the video along with annotations from all three steps. We provide them with the UI in Fig. 10, where they can grade each annotation on the basis of correctness (where one mistake can indicate a deduction of one point), level of detail (do the descriptions have sufficient detail in order to easily discern people/objects and actions?), and grammar (are the descriptions reasonably well written, it does not have to be perfect). We report the average scores between the two individuals.

Furthermore, following this process, we ask the experts for feedback. We received the following feedback:

- Some annotations may contain minor mistakes, where the annotator did not look at the video carefully, and therefore marked a description that could have been valid as invalid (or vice versa). A common cause of this was the low video quality. For example, an annotator invalidated a claim that a car hit the sidewalk, though in V_{post} , the car does hit the sidewalk, but the low quality of the V_{post} made difficult to observe.
- In some cases, annotators may have missed obvious explanations of what is going on. In such cases, their descriptions may be reasonably correct, but not a true description of what is happening.
- Regarding level of detail, the experts suggested that in some cases, the sentences were too short and did not describe the scene sufficiently.
- Grammatically speaking, the only times it was marked as "no" is when there were multiple grammatical errors across multiple descriptions for the same video.

B.3. Annotator Statistics

To evaluate the diversity in the annotators of our dataset, we consider age, education level, gender and country of the annotators. Here are the statistics:

Age:

- Average Age: **37.5**
- Median Age: **36**
- Minimum Age: **22**
- Maximum Age: **61**

Education Level:

- Bachelor’s degree (for example: BA, AB, BS): **60.3%**
- Master’s degree (for example: MA, MS, MEng, MEd, MSW, MBA): **37.76%**
- Doctorate degree (for example: PhD, EdD): **1.94%**

Gender:

- Man: **48.95%**
- Woman: **39.5%**
- Not Known: **11.55%**

Country:

- USA: **73.85%**
- Canada: **13.05%**
- New Zealand: **11.5%**
- UK: **1.6%**

C. Metrics

CLIP Score. We use clip-vit-large-patch14 to embed each sentence, and use cosine similarity as the distance measure for any pair of sentences.

LLM-Based Metric We use Llama 3.1 8B (Huggingface: Llama-3.1-8B-Instruct). We chose to take the mean instead of the max of each pairwise score while comparing the ground truth set. This is because the sparsity of scores (1, 2, 3, 4 or 5) makes it very easy the LLM to rate a max score of 3 or 4 for any pair of sentence sets, yielding results that show very little distinction between different models. Instead, taking the mean allows us to also measure the diversity of generations, which indicates how aligned LLMs are with humans for all their predictions, across all samples generated.

The prompt for LLM-Match is:

You are an AI assistant tasked with evaluating how well a given response aligns with the provided ground truth. Focus on the semantic similarity between the two texts. Your assessment

should produce a single integer score between 1 and 5:

5: The response matches the ground truth perfectly.

1: The response is entirely different from the ground truth.

Please return your evaluation in the following format:

Reason: A brief, ten-word explanation for your score.

Score: Your score.

Ground Truth: ground_truth

Response to Score: model_generated

BLEU and ROUGE. Tasks in BlackSwanSuite often involve generating hypotheses, explanations, or descriptions, which are inherently open-ended. Multiple valid answers may exist, and these may differ significantly in phrasing from the reference answer. N-gram based metrics such as BLEU and ROUGE penalize such variations, despite them being semantically correct. Hence, these metrics may not accurately measure correctness of generations. However for the sake of completeness, we report them in Table 10.

Model	Forecaster		Detective		Reporter	
	BLEU	ROUGE-L	BLEU	ROUGE-L	BLEU	ROUGE-L
GPT-4o	0.70	27.10	0.80	26.06	0.90	23.40
Gemini 1.5 Pro	1.26	28.03	1.12	24.14	0.43	11.90
VideoChat 2	0.49	25.80	1.00	26.04	0.70	20.90
VideoLLaMA 2	0.67	25.20	0.11	15.42	0.52	17.30
VILA	0.80	18.77	0.83	16.59	0.60	15.30
LLaVA-Video	0.49	20.50	0.61	24.28	0.81	20.50
Human	1.18	25.41	0.98	24.74	0.00	0.45

Table 10. BLEU and ROUGE-L results for Forecaster, Detective, and Reporter Gen variants ($\times 100$).

D. Human Evaluation Setup

In order to conduct human evaluation for the generative variants, we use the template in Figure 11. We define the four criterion as:

1. Correctness (5 Point Scale) Check whether the answer correctly describes what could have happened in part 2 based on part 1.
2. Depth and Thoughtfulness (5 Point Scale) How much depth in reasoning is covered in the description? Is there sufficient reasoning to support the description?
3. Level of Detail (5 Point Scale) Does the description have sufficient level of detail to easily identify and differentiate between people/objects and actions?
4. Visual Contradiction (Yes/No) Does the description follow the content in the video, or contradict the contents of

the video? A contradiction happens when the description is impossible, given the video.

We use the same location, age and education filters as used above while collecting annotations. Annotators are always required to write reasons for their choices; accordingly, after running this evaluation for one model for 50 videos, we manually filter out evaluators who have done this task diligently. This gave us 32 annotators who were capable of doing this task well. We use these annotators on all evaluations of model and human written explanations (in the case of Forecaster-Gen and Detective-Gen, we sample a single explanation from the three generated by the model).

While we would have liked to conduct this evaluation with multiple annotators for each video, and over a larger set of videos, we believed that this would be very expensive and not a practical use of funds. We paid evaluators \$0.45 per evaluation, as each video and explanation can require 4-5 minutes to evaluate (since reasons must be written). Say we had used 50 videos, and evaluated a single explanation for each task, from all models, 3 separate times by different annotators. At this rate, for 7 models (incl. human), 50 videos, and 3 tasks and 3 annotators, and a 25% service fee, this would be over \$1770. We instead sought out to compare results qualitatively, which we do in Appendix F.2.

E. Baselines

We test several latest VLMs. Closed-source VLMs include:

- 1. OpenAI GPT-4o [21]** GPT-4o was an important choice for our video-based task because, apart from leading all kinds of VL benchmarks, it is OpenAI’s first model that is capable of accepting up to 50 pictures as input in the prompt. We use the GPT-4o model using the OpenAI Batch API³. The model was accessed between October 20th and November 14th, 2024. We feed it 10 uniformly sampled frames for every part of the video.

- 2. Google Gemini 1.5 Pro [33]** We chose to use Gemini 1.5 Pro as it is the most capable VL model that can natively accept video input. This model was accessed between October 20th and November 14th, 2024. When asking questions, we directly upload the entire video clip. In the case of Detective, the clip has the V_{main} blacked out.

Open source models include:

- 3. VideoChat2 [13]** VideoChat2 showed stronger detail and contextual understanding than models prior to it, and even beat GPT-4V at the time of its release on multiple VL tasks. We evaluate the latest VideoChat2 HD (Huggingface: videochat2.hd.mistral.7b.stage4) model. When asking questions, we directly upload the entire video clip. In the case of Detective, the clip has the V_{main} blacked out. Default settings of 16 frames per video input are used, with resolution set to 224.

³<https://platform.openai.com/docs/guides/batch/overview>

- 4. VideoLLaMA 2 [4]** VideoLLaMA 2 is capable of understanding both visual and audio signals. Though audio is not a requirement for this task, there may be cases where it can reveal important information about the events in the video. We use the VideoLLaMA2.1-7B-16F model. Again, when asking questions, we directly upload the entire video clip. In the case of Detective, the clip has the V_{main} blacked out. Default settings for resolution are used, and max number of frames are 32.

- 5. VILA [16]** VILA 1.5 is the latest in the VILA series of models, one of the first models to support multi-image understanding. We use the Llama-3-VILA1.5-8b-Fix model. We uniformly sample 5 frames for each part of the video.

- 6. LLaVA-Video [43]** LLaVA was first introduced as a multimodal model with performance rivaling GPT-4. Following the release of LLaVA 1.5 and LLaVA-NeXT, LLaVA-Video is the most advanced version of this model, specifically trained on a new 178K video dataset. Specifically, we use the LLaVA-Video-7B-Qwen2 model specification. When asking questions, we directly upload the entire video clip. In the case of Detective, the clip has the V_{main} blacked out. 32 frames are used. We also test the 72B LLaVA-Video (LLaVA-Video-72B-Qwen2) model; however, due to compute limitations and setup issues, we were only able to run that model on a subset of the data. The results are detailed in Appendix F.3.

Our implementation of all open-source models follows their respective instructions on GitHub and Huggingface, and we tried our best to recreate the same environment as the original developers for each model. We do not do batch inference on any of these models, and instead query them iteratively in chat mode only (each chat conversations for every question is independent). For the generative variant, in order to obtain different results each time, we turn sampling on.

We use the following prompts for each task:

Forecaster-Gen Describe what could happen next, by explaining the sequence of actions leading to the outcome.

Detective-Gen What happened in the missing frames (in black) of the video?

Detective-MCQ Select the description that indicates what happened in the hidden (black) frames of the video: A. <Option A> B. <Option B> C. <Option C>

Detective-Y/N Hypothesis: <hypo>
Given the video clip, does this hypothesis hold? Answer yes or no.




<div>Pre-event: V_{pre}</div> <div></div>		<div>Main event: V_{main}</div> <div></div>		<div>Post-event: V_{post}</div> <div></div>	
Generative task outputs for the above video:					
Task:	GT	GPT-4o	Llava-Video	VideoChat2	Human
Forecaster-Gen	The monkey will land on the steering wheel causing an accident	The monkey decides to climb onto the driver's lap, causing the driver to carefully adjust his seating to avoid any disturbance while driving.	The monkey, having successfully grabbed the driver's attention and hand, might attempt to climb into the vehicle, potentially causing a distraction or interruption in the driver's focus on the road.	The monkey, having been given the steering wheel , could potentially take control of the vehicle, leading to an unpredictable outcome. (a)	The monkey moves past the camera onto the bus because it is curious and is exploring.
Detective-Gen	The monkey will enter the vehicle, steal something, and then run out	The monkey might playfully interact with the driver and then exit the vehicle, prompted by being offered a treat.	The missing frames in the video, ..., likely indicate a transition or a cut in the footage... [incorrect] (b)	The monkey, in the missing frames, is seen falling from the vehicle.	Its possible the monkey left the truck because the driver seems to be looking for the animal.
Reporter-Gen	The monkey will enter the vehicle, steal something, and then run out	A monkey rides inside a vehicle with a driver, explores the dashboard , and eventually hops out of the vehicle. (c)	A monkey is sitting on the dashboard of a bus and interacting with the driver. (d)	A monkey is seen sitting on the driver's lap and steering the vehicle while the driver is wearing a headset and appears to be in a state of surprise... (e)	A monkey <u>grabbed a plastic bag</u> and jumped out the window of a moving bus. (f)

Figure 6. **Qualitative results on Gen variants.** Due to space constraints, only one sample from each model is shown.

Reporter-Gen Explain what is happening in the video.

Reporter-MCQ Select the description that correctly explains what happens in this video: A. <Option A> B. <Option B> C. <Option C>

Reporter-Y/N Hypothesis: <hypo>
Given the video clip, does this hypothesis hold? Answer yes or no.

Depending on the model, we may vary the prompt in small ways. For example, we may use the default system instruction for each model, or ask models to return an answer in only one sentence. In the case of multi-frame models, when we feed frames from the beginning and the end, we specify it as such:

Here is the beginning of the video: <image tokens for V_{pre} >
Here is the end of the video: <image tokens for V_{post} >
<Question...>

F. Results

F.1. Human Baseline

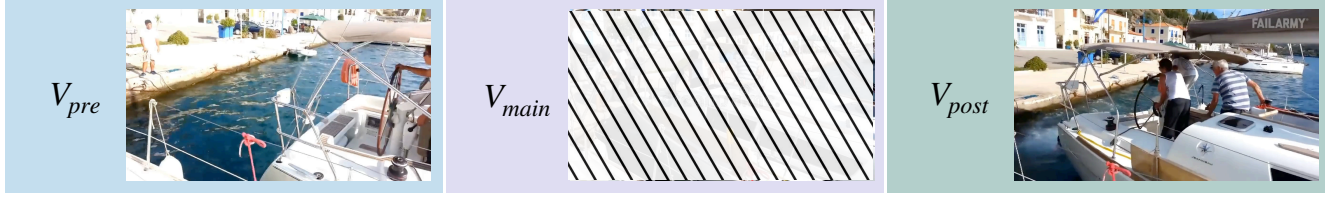
We compare models to human performance. We ask humans to do the same tasks as models. For the generative variant, we use the template in Figure 12 to collect annotations (we build similar templates for other generative variants, with more parts of the videos shown). For the MCQ

variant, an example template is shown in Figure 13. Similarly, for the Y/N task, an example template is shown in Figure 14.

For each of the MCQ and Y/N variants of **Detective** and **Reporter**, we ask two students from the lab to answer 100 randomly sampled questions for MCQ and 150 randomly sampled questions for Y/N and report the maximum score achieved across the two annotators. We do this since we want to measure the maximum achievable score by a human expert on a given task, as an upper bound. For the generative variant of all three tasks, we crowd source new annotations by a single annotator (same CloudResearch setup as before), and compute metrics on the obtained annotations.

F.2. Additional Qualitative Results

Figure 6 demonstrates the outputs for the generative task variants. Models are reasonably good at estimating what could happen next (**Forecaster-Gen**), albeit with some uncertainty. For example, in (a), VideoChat2 hallucinates the action of having been given the steering wheel, and expresses uncertainty (“leading to an unpredictable outcome”) Having uncertainty, by itself, is technically not wrong, but it is often a way for the model to avoid the question. Uncertainty is better than hallucination or returning an entirely wrong answer. In part (b) (**Detective-Gen**), LLaVA-Video misunderstands the instruction to hypothesize about what is happening in the missing frames and instead describes why the missing frames are not present. In **Reporter-Gen** (parts c–e), all models failed to capture the instantaneous event where the monkey picks up the plastic bag and leaves, which lasts less than 2 seconds. The most clear response, in this case, is from the human, who clearly stated that the



Question: What could have happened in between?

- A. The man in the striped shirt tries to get off the boat, but falls into the water.
- B. The boat ends up drifting backwards and then slams into the dock again.
- C. The boat bounces off the dock so hard that it drifts backwards and the men are unable to dock.

Perception:

V_{pre} : A boat is approaching a pier too fast. Two people are on the boat, one is a woman who is attempting to steer the boat and stop it. As the boat is about to collide with the pier three men approach the boat to try to stop it.

V_{post} : Someone is docking their boat. Someone else is recording on their other boat. The docking boat is going in reverse. An elderly man appears to want to board the other boat, but is stopped.

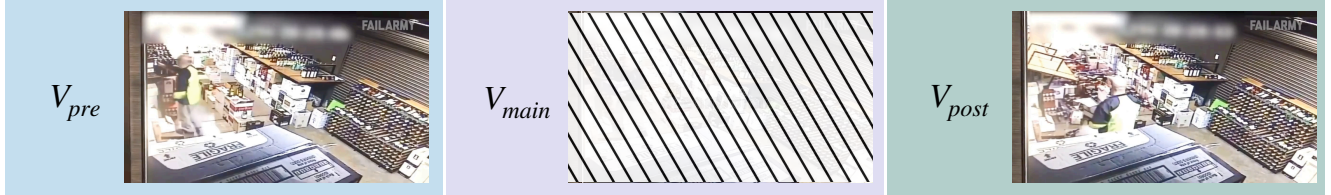
Comprehension:

In the beginning the woman steering the boat tries to slow it down before it hits the dock. The men standing on the dock immediately try to kick the boat to slow it down and push it back. The end video only shows the couple on the boat going backwards for a few seconds.

GT Ans: C

Baseline: B ❌ | +Perception: C ✅ | +Perception+Comprehension: C ✅

(a) Example where LLaVA-Video benefits from Perception



Question: What could have happened in between?

- A. As the guy carries the box of wine bottles, he begins to slip around while still carrying them.
- B. The guy throws the box of wine bottles in the air out of frustration and lets the bottles crash onto the floor all around him.
- C. As the man removes a box of wine bottles from the table, the table starts to wobble, causing the other boxes still on the table to start falling to the floor.

Perception:

V_{pre} : A man is removing a box of wine bottles from a shelf in a liquor storage area or liquor store. The area is closed up and presumably not open to the public or not a retail store.

V_{post} : A man is standing with his back to the camera. Surrounding him are many shelves and boxes with what appear to be wine and liquor bores. Directly behind the man is a box labeled "Fragile".

Comprehension:

In the beginning, a bald man wearing tan pants, a black shirt, and a yellow vest appears to be taking boxes off a shelf on the left-side wall of a warehouse or brewery. In end, the man is seen facing away from the camera looking at the shelf he originally took the box from. The shelf appears to have tipped, as it's leaning sideways and its contents are all over the floor.

GT Ans: C

Baseline: B ❌ | +Perception: B ❌ | +Perception+Comprehension: C ✅

(b) Example where LLaVA-Video benefits from Comprehension

Figure 7. **Perception and Comprehension with LLaVA-Video.** (a) shows a case where human-written Perception helps. (b) shows a case where both human-written Comprehension helps.

monkey grabbed the plastic bag (f).

For additional examples, please watch the video examples.mp4 in the supplementary zip file.

F.3. Additional Quantitative Results

Model	Detective		Reporter	
	MCQ	Y/N	MCQ	Y/N
Open Source				
LLaVA-Video-7B	55.9	59.3	69.6	55.1
LLaVA-Video-72B	59.91	56.93	74.91	54.47
Human				
Human	90.0	85.3	95.3	92.0

Table 11. Results on MCQ and Y/N variants of Forecaster and Detective on 72B variant of LLaVA-Video on 20% of the data.

Model	Forecaster		Detective		Reporter	
	CLIP	LLM-M	CLIP	LLM-M	CLIP	LLM-M
Open Source						
LLaVA-V-7B	0.64	1.57	0.58	<u>1.70</u>	0.57	<u>2.36</u>
LLaVA-V-72B	<u>0.69</u>	<u>1.67</u>	<u>0.60</u>	1.68	<u>0.59</u>	2.15
Human						
Human	0.78	1.98	0.77	1.92	0.76	3.23

Table 12. Results on Gen variants with 72B version on 10% of data. Metrics are grouped by tasks, each with CLIP and LLM-M scores.

In Table 11 we include the results on a subset of 20% randomly sampled MCQ and Y/N questions on the 72B LLaVA-Video model, and compare it against the 7B model. We observe that the 72B version significantly outperforms the 7B variant and reaches an accuracy on both MCQ and Y/N close to the best-performing closed-sourced models shown in Table 2.

Table 12 shows CLIP Score and LLM-Match on the three tasks for the 72B variant, on 10% of the data. We only generate a single explanation for each task (in general, we generate three responses for Forecaster and Detective for all other models). We observe that the 72B version performs very similarly to the 7B version.

G. Leaderboard Results

We release our data with two splits: a validation split, where we make ground truth labels accessible, and a test split, where we hide ground truth labels. This is to prevent misuse of our data. The validation set contains 827 videos (50% of data), of which 95 are from the hard subset (Sec. 7.3). The test set contains 828 videos, of which 224 are from the hard

Model	Detective		Reporter	
	MCQ	Y/N	MCQ	Y/N
Closed Source				
GPT-4o	<u>67.2</u>	59.8	<u>80.3</u>	63.7
Gemini 1.5 Pro	60.4	<u>64.7</u>	71.5	57.0
Open Source				
VideoChat2	31.3	63.9	53.8	51.3
VideoLLaMA 2	53.8	61.3	53.5	56.1
VILA-1.5	53.3	58.7	57.0	52.4
LLaVA-Video	57.4	61.6	68.7	58.9

Table 13. Validation Set results on MCQ and Y/N variants of Forecaster and Detective. The best result for each task is shown in bold, and the best model result is underlined.

Model	Detective		Reporter	
	MCQ	Y/N	MCQ	Y/N
Closed Source				
GPT-4o	<u>63.2</u>	<u>65.7</u>	<u>78.5</u>	<u>55.9</u>
Gemini 1.5 Pro	57.1	62.0	70.6	49.1
Open Source				
VideoChat2	28.5	63.8	49.7	42.3
VideoLLaMA 2	53.3	56.6	53.0	52.1
VILA-1.5	49.4	56.5	52.2	48.7
LLaVA-Video	54.5	57.1	70.4	51.2

Table 14. Test Set results on MCQ and Y/N variants of Forecaster and Detective. The best result for each task is shown in bold, and the best model result is underlined.

Model	Forecaster		Detective		Reporter	
	CLIP	LLM-M	CLIP	LLM-M	CLIP	LLM-M
Closed Source						
GPT-4o	0.77	1.63	0.78	2.07	0.65	3.08
Gemini 1.5 Pro	0.79	1.64	0.74	2.24	0.53	2.55
Open Source						
VideoChat2	0.71	1.17	0.68	1.81	0.62	2.16
VideoLLaMA 2	0.78	1.45	0.61	1.16	0.66	2.13
VILA-1.5	0.66	1.53	0.50	2.16	0.40	2.07
LLaVA-V-7B	0.65	1.66	0.59	1.71	0.58	2.44

Table 15. Validation Set Results on Gen variants. Metrics are grouped by tasks, each with CLIP and LLM-M scores.

subset. This makes the test set slightly more challenging than the validation set.

Results on MCQ and Y/N variants for the validation and test sets are in Tables 13 and 14 respectively.

Model	Forecaster		Detective		Reporter	
	CLIP	LLM-M	CLIP	LLM-M	CLIP	LLM-M
Closed Source						
GPT-4o	0.76	1.66	0.77	2.09	0.64	2.92
Gemini 1.5 Pro	0.77	1.57	0.72	2.05	0.48	2.45
Open Source						
VideoChat2	0.70	1.62	0.68	1.98	0.60	2.08
VideoLLaMA 2	0.77	1.46	0.62	1.19	0.65	1.97
VILA-1.5	0.66	1.48	0.49	2.20	0.39	2.09
LLaVA-V-7B	0.64	1.50	0.58	1.70	0.57	2.29

Table 16. **Test Set Results on Gen variants.** Metrics are grouped by tasks, each with CLIP and LLM-M scores.

H. Analysis

H.1. Perception-Comprehension

We collect human perception and comprehension descriptions using the templates in Figure 15 and Figure 16 respectively. Once again, we use the same location, age and education criterion to select annotators as before. We pay \$0.26 per annotation.

For perception input to LLaVA-Video, we modify the prompt for **Detective-MCQ** to:

The beginning of the video shows <p.preevent>. The end of the video shows <p.postevent>.

Which of the following descriptions indicate what happened in the hidden (black) frames of the video? A. <Option A> B. <Option B> C. <Option C>

For perception and comprehension, we further modify the prompt to:

The beginning of the video shows <p.preevent>. The end of the video shows <p.postevent>. The two parts differ in the following way: <pc.comp>.

Which of the following descriptions indicate what happened in the hidden (black) frames of the video? A. <Option A> B. <Option B> C. <Option C>

For examples with full video clips, please watch the video `examples.mp4` in the supplementary zip file. In Figure 7, we show an example how perception and comprehension have helped the baseline model arrive at correct answers.

H.2. CoT examples

Because of the low-quality of LLaVA-Video on this task, we focus on GPT-4o. In Figure 8, we show examples of how step-by-step reasoning affects answer choice selection for GPT-4o. In many cases, these reasoning steps are correct (e.g. in part (a)). On the other hand, preconceptions about the scene can influence the reasoning process, leading to mistakes (e.g. in part (b)). Again, for examples with full video clips, please watch the video `examples.mp4` in the supplementary zip file.

H.3. Challenging subset examples

Hard-subset examples with video clips are shown in `examples.mp4` in the supplementary zip file. These are marked with an orange asterisk on the top right corner.

I. Limitations and Ethical Considerations

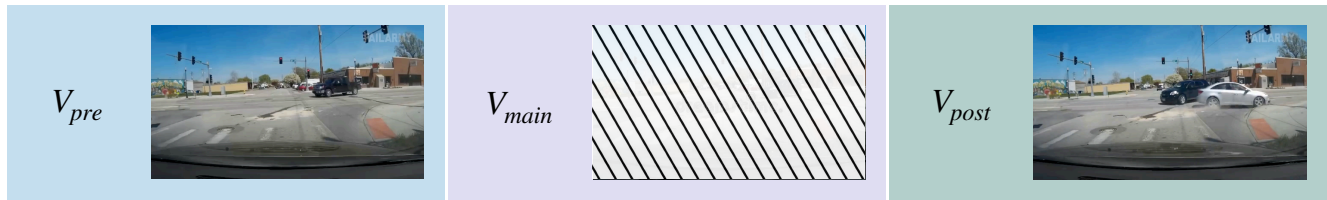
Data Source: As we use videos from the test set of Oops! [7], it is possible that some VLMs trained on extensive datasets may already be familiar with similar content. However, our benchmark can be extended to include newer videos in future.

Evaluation Metrics: Although we defined our MCQ and Y/N tasks to challenge models to perform more complex reasoning, quantitative metrics like accuracy might not reflect the depth of reasoning or the logical processes involved. For generative tasks, current metrics struggle to capture the nuances involved. We address this by asking humans to evaluate the thoughtfulness and visual contradictions in reasoning, but further research is needed to automatically evaluate explanations generated by models. Furthermore, given the open-ended nature of our generative tasks, our human evaluation may be subject to bias. Our MCQ and Y/N variants, on the other hand, do not suffer from this limitation. Finally, we conduct all our human annotations according to ethics policies.

Pre-training strategies: The models evaluated are primarily trained on language modeling, and may not have been explicitly trained for abductive and defeasible reasoning, potentially limiting their performance. Further research is required to study how different pretraining or finetuning approaches may perform on this task.

Explanation complexity: Our current annotation process focuses on free-form explanations for the unexpected scenarios, however, it may be interesting to study how more scientific reasoning (e.g., using intuitive physics to explain a fall) may influence the performance of models.

Size and diversity: Although BlackSwanSuite includes 15,469 questions, the size and diversity might still be insufficient to generalize findings across all types of reasoning challenges and rare and highly uncommon events might still be underrepresented.



Question: Explain what could have happened in between.

- A. A car from the left comes speeding through almost hitting the black car.
- B. The oncoming car speeds through a red light
- C. The black suv turns and runs into the car that is filming

GT Ans: A

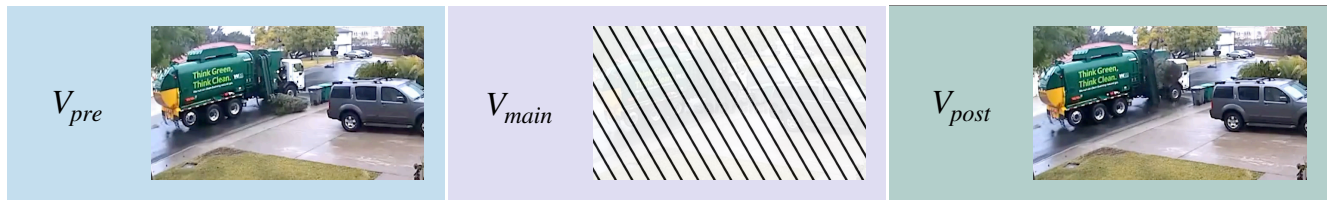
GPT-4o Answer:

Reason:

- Step 1: In the beginning frames, a black SUV is making a left turn at an intersection,
- Step 2: The SUV completes its turn without hitting anything visible,
- Step 3: In the end frames, a silver car is seen, indicating a sudden appearance,
- Step 4: The silver car seems to have come from the direction of the left, suggesting it sped through
- Step 5: There is no visual evidence of a collision with the filming car

Answer: A. ✓

(a) Example where GPT-4o gets it right



Question: Explain what could have happened in between.

- A. The garbage truck will park on the side of the road and collect trash from a green bin.
- B. The truck will malfunction, dropping the tree on the ground.
- C. The tree will be sent flying in the air and damage a nearby car.

GT Ans: B

GPT-4o Answer:

Reason:

- Step 1: The initial frames show a garbage truck lifting a tree,
- Step 2: The final frames show the tree being held higher by the truck,
- Step 3: There is no visible damage to the surrounding area,
- Step 4: Green bins are present, suggesting routine collection

Answer: A. The garbage truck will park on the side of the road and collect trash from a green bin. ✗

(b) Example where GPT-4o gets it wrong

Figure 8. CoT Examples with GPT-4o. (a) has a case where GPT-4o shows good step-by-step reasoning to arrive at the correct answer. (b) shows that GPT-4o sometimes relies of preconceived notions in its reasoning steps, and therefore makes a mistake.

Welcome!

Welcome to this task! Please begin by reading the full instructions and the guidelines.

Full Instructions [\(Expand/Collapse\)](#)

Task Instructions

Thanks for participating in this HIT! Each video is divided into three parts Part 1, Part 2 and Part 3, described below.

Part 1: The beginning or setup of the scene in the video.

Part 2: The interesting event in the video.

Part 3: The result or outcome of the event in the video.

Tasks:

1. **Task 1:** First, you will watch **Part 1** of the video, and **Part 2** and **Part 3** will be hidden. You need to write **diverse** descriptions of what could have happened in **Part 2** (the middle part).
2. **Task 2:** Next, you will be revealed **Part 3** of the video. You need to validate whether your descriptions from Task 1 can answer the question *What happens in between Part 1 and Part 3*?. If marked invalid, you **must** update a new one based on the video. Update descriptions for any changes, even if they are minor.
3. **Task 3:** Finally, you will get to watch **Part 2** in the video. Evaluate if any of your previous descriptions align with **all three parts of the video**. You also need to provide a new description, based on the ground truth video shown to you.

In all the tasks, please make sure your descriptions capture the *visual* context in the video accurately. As we provide more visual context and expose you to additional parts of the video, we expect more and more visually grounded descriptions.

Example

Task 1: Given Part 1, guess Part 2

Part 1:



Part 2 (Hidden)

Part 3 (Hidden)

Think about what could happen in Part 3 (the outcome of the video).
Write three possible descriptions on what will happen in Part 2 to lead to any outcome in Part 3.


1. The skateboarder in the back hits the person skating in the front and they both lose balance.
2. The skateboarder in the back moves closer to the person skating in front but goes around him, avoiding an accident.
3. The skateboarder in the back hits the person skating in the front, and they both hit a car.

(a) 1 of 6

Figure 9. Black Swan Annotation Template

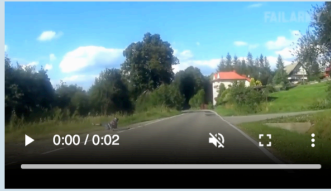
Task 2: Given Part 1 & 3, update Part 2

Part 1:



Part 2 (Hidden)

Part 3:



After watching Part 3, you need to update your descriptions for Part 2.
 Validate what could have happened in this part.
 Please write a new description for each invalid one.

1. *Description:* The skateboarder in the back hits the person skating in the front and they both lose balance.
☒ Valid

2. *Description:* The skateboarder in the back moves closer to the person skating in front but goes around him, avoiding an accident.
☒ Invalid (the skateboarder does fall)
Updated Description: The skateboarder in the back tried to go around the skateboarder in front, but he still lost balance and fell on the roadside.

3. *Description:* The skateboarder in the back hits the person skating in the front, and they both hit a car.
☒ Invalid (There is no car)
Updated Description: The skateboarder in the back hits the person skating in the front causing him to turn and fall in the middle of the road.


Figure 9. (Continued) Black Swan Annotation Template

Task 3: Given all parts, validate Part 2

Part 1:



Part 2:



Part 3:



Select the valid descriptions for this part. If none are valid, please write a new description.

- Description:* The skateboarder in the back hits the person skating in the front and they both lose balance.

✗ Invalid (the skateboarder in the front does not lose balance/fall.)
- Description:* The skateboarder in the back tried to go around the skateboarder in front, but he still lost balance and fell on the roadside.

✓ Valid
- Description:* The skateboarder in the back hits the person skating in the front causing him to turn and fall in the middle of the road.

✗ Invalid (the skateboarder in the front does not get hit by the skateboarder in the back.)

Guidelines:

- **Visually grounded** - Your descriptions should be grounded to the visual context - objects, scenes and activities in the video.
- **Plausible** - Your descriptions should be plausible. Please don't write descriptions that break physical rules or are against common sense.
- **Persuasive** - Your descriptions should be persuasive. It should add sufficient information so that the outcome happened given the initial context of the video.
- **Non-contradictory** - Do not contradict the information in the videos.

Figure 9. (Continued) Black Swan Annotation Template

Task 1: Given Part 1, describe Part 2

Part 1:

▶ 0:00

Part 2 (Hidden)

Part 3 (Hidden)

Think about what could happen in Part 3 (the outcome of the video).

Write three possible diverse descriptions on what will happen in Part 2 to lead to any outcome in Part 3.

Describe WHY and HOW (rationale) to reach Part 3.

Describe WHY and HOW (rationale) to reach Part 3.

Describe WHY and HOW (rationale) to reach Part 3.

Please write at least 8 words per description. All descriptions should be unique. Please avoid using "it", "he", "she" or "they"; instead use specific descriptions like "the cat" or "the boy on the skateboard".

Think about what could happen in the outcome here!

You should consider many possible outcomes, then write the descriptions in Part 2.

For an example of how each part looks, please see the instructions above.

Please check that your descriptions are:

Visually grounded

Plausible

Persuasive

Non-contradictory

Proceed to Task 2

(d) 4 of 6

Figure 9. (Continued) Black Swan Annotation Template

Task 2: Given Part 1 & 3, describe Part 2

Part 1:

▶ 0:00

Part 2 (Hidden)

Part 3:

▶ 0:00

After watching Part 3, are your answers from Task 1 still possible?

Previous:

Is this description still possible?

☒ Yes, possible

☐ No, impossible

Previous:

Is this description still possible?

☒ Yes, possible

☐ No, impossible

Previous:

Is this description still possible?

☐ Yes, possible

☒ No, impossible

Please write a new description to replace the impossible ones. No duplicates please, you can be creative!

Describe what could happen in Part 2 to lead to Part 3

Please write at least 8 words per new description. Also, do NOT describe the outcome of the event that you see in Part 3; only write about Part 2.

Please check that your descriptions are:

Visually grounded

Plausible

Persuasive

Non-contradictory

Proceed to Task 3

(e) 5 of 6

Figure 9. (Continued) Black Swan Annotation Template

Task 3: Given all parts, validate Part 2

Part 1:

▶ 0:00

Part 2:

▶ 0:00

Part 3:

▶ 0:00

Select the description, that are correct/possible after watching the video.

Previous:
Is this description still correct/possible?

☐ Yes, possible
☐ No, impossible

Previous:
Is this description still correct/possible?

☐ Yes, possible
☐ No, impossible

Previous:
Is this description still correct/possible?

☐ Yes, possible
☐ No, impossible

Finally, please write a new description based on the true video. Make sure your description is specific and detailed, and accurately describes the events in Part 2 and Part 3:

Final detailed description. Be more specific and accurate than your previous descriptions.

Required Feedback: Are there more than one complete events/clips/videos in the video (*spanning Part 1, 2 & 3*).

☐ Yes
☐ No

[In general, this should be "No", but if you find a case where there are two complete sub-videos in one video, please report "Yes" here.]

Please check that your descriptions are:

Visually grounded

Plausible

Persuasive

Non-contradictory

☐ *Optional Feedback:* Something about the HIT is unclear/These videos have hateful/offensive content/You have additional feedback:

Submit

(f) 6 of 6

Figure 9. (Continued) Black Swan Annotation Template

Black Swan Validation


10

Get Video


Next Video

ID: 10,oops_val_v18,1


Part 1:



Part 2:



Part 3:



Task 1 (with Part 1): Given the beginning, what do you think happened next?

1. A pile of snow suddenly falls on top of the driver and passenger.

Valid

2. A flock of birds appear and fly into the faces of the driver and passenger.

Invalid

3. The car suddenly breaks down and stops in the middle of the road.

Invalid

Task 2 (with Part 1 and Part 3): Given the beginning and end, what do you think happened in between?

1. A pile of snow suddenly falls on top of the driver and passenger.

Invalid

2. The guy in the passing truck blares his horn at the driver and passenger.

Invalid

3. A strong wind blows in the faces of the driver and passenger... causing the driver lean back in his seat.

Invalid

Task 3 (all parts): Given the full video, explain what happened.

1. As the car turns the corner... the truck splashes water from the ground onto the car and in the faces of the driver and passenger as it goes by.

Feedback Form

Rate the correctness:

5

Rate the reasoning quality or depth/thoughtfulness:

5

Is it grammatically correct?

Yes

No

Submit Feedback

Figure 10. Template for Dataset Validation.

Welcome!

Welcome to this task! Please begin by reading the full instructions and the guidelines.

Full Instructions (Expand/Collapse)

Task Instructions

Thanks for participating in this HIT! In this task, you will evaluate an inference derived from a part of a video. Each video is divided into three parts Part 1, Part 2 and Part 3, described below.

Part 1: The beginning or setup of the scene in the video.

Part 2: The interesting event in the video.

Part 3: The result or outcome of the event in the video.

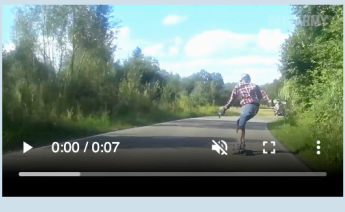
In this task, you will watch only **Part 1** of the video, and **Part 2** and **Part 3** will be hidden. You need to evaluate a description that answers the question, "What could have happened in **Part 2** (the middle part) after seeing Part 1?" based on the criteria listed below. Please refer to the example to understand the scoring better.

Criteria:

- Correctness (5 Point Scale)** Check whether the answer correctly describes what could have happened in part 2 based on part 1.
- Depth and Thoughtfulness (5 Point Scale)** How much depth in reasoning is covered in the description? Is there sufficient reasoning to support the description?
- Level of Detail (5 Point Scale)** Does the description have sufficient level of detail to easily identify and differentiate between people/objects and actions?
- Visual Contradiction (Yes/No)** Does the description follow the content in the video, or contradict the contents of the video? A contradiction happens when the description is impossible, given the video.

Example

Given Part 1, what happens next? (in Part 2)

Part 1:


Part 2 (Hidden)

Part 3 (Hidden)

(a) 1 of 3

Figure 11. Human Evaluation for Generative Tasks

Correctness:

Description	Score	Rubric	Reason
The skateboarder in the back goes very close to the one in the front, and it looks like they might collide and lose balance.	5	Very likely correct	Matches the situation well, and the inference is valid based on the video context.
The car passes by, and the skateboarder in the back gets too close to the one in the front. There might be an accident.	4	Mostly correct	Mostly correct but lacks clear mention of balance or specific consequences.
The skateboarder in the back skates faster and overtakes the one in the front. It could lead to a fall.	3	Partly incorrect	Partially accurate; overtaking is not visible, but inferring a fall is plausible.
The skateboarder in the back is far away from the front one and continues skating without any issues.	2	Possibly incorrect	Incorrect inference; the video shows the skateboarders close to each other.
The skateboarders stop skating and start walking after the car passes by.	1	Very likely incorrect	Completely incorrect as it contradicts the visible actions in the video.

Depth and Thoughtfulness:

Description	Score	Rubric	Reason
The skateboarder in the back seems to miscalculate the speed and distance from the front skateboarder after the car passes. This could lead to a loss of balance, especially as the road seems narrow.	5	Very thoughtful reasoning	Thorough and thoughtful with specific details about speed, distance, and road condition.
The skateboarder in the back almost collides with the one in the front, which could lead to them both falling.	4	Thoughtfully written, but could be better.	Thoughtful, but missing some details like the speed or the road condition.
The skateboarders are skating close to each other, and the one in the back might fall if they collide.	3	Mediocre level of reasoning depth	Provides some thought but lacks depth or meaningful insight.
The skateboarder in the back gets very close to the front one, and something bad might happen.	2	Somewhat vague reasoning	Too vague and lacks meaningful analysis or details.
Both skateboarders keep skating without any issues after the car passes.	1	No depth in reasoning	No depth or thought in the inference; it's a basic and shallow observation.

Level of Detail:

Description	Score	Rubric	Reason
The skateboarder in the red checkered shirt miscalculates the distance and falls to the grass ahead of him while trying to avoid hitting the one in front.	5	The level of detail is sufficient to perfectly identify the people, objects and cause of actions.	"skateboarder in the red checkered shirt" has enough level of detail to differentiate between the two skateboarders and the cause is specifically written.
The skateboarder in the back of the other skateboarder slows down but loses control and falls on the grass.	4	The level of detail is good enough to identify the people, objects and cause of actions, but may not be specific in some aspects.	"skateboarder in the back" has enough level of detail to differentiate between the two skateboarders, but additional information like "grass ahead of him" could help.
One of the skateboarders slows down but loses control, hits the other skateboarder, and falls on the grass.	3	Level of detail may be just sufficient for answering the question.	"One of the skateboarders" implies that there are two skateboarders, but is not enough level of detail to differentiate between the skateboarders clearly.
The young skateboarder is full of energy, and is going really fast on the roads, but accidentally veers off the road and falls.	2	Poor level of detail, missing important aspects to differentiate between different people/objects and actions.	"young" and "full of energy" are not really important for differentiating between the two skateboarders, in fact, there is no mention of the other skateboarder
A skateboarder hits another skateboarder.	1	Very poor level of detail. Does not describe the people/objects or actions sufficiently.	Does not differentiate between the skateboarders and does not describe the cause of the fall.

Please keep in mind that the level of detail is specific to the video. For example, in a video where there is only one person, a detailed description may only contain "The person" as that is sufficient information.

Visual Contradiction:

Description	Does the description reasonably follow or contradict what is happening in the video?
The skateboarder in the back is close to the front one, and it seems likely they might collide.	Follow
The car passes, and the skateboarders are close to each other; a collision might occur.	Follow
The skateboarders are far apart from each other, and no collision seems possible.	Contradict
The skateboarders suddenly stop and walk away from the scene.	Contradict

Figure 11. (Continued) Human Evaluation for Generative Tasks

Given Part 1, what happens next? Validate this answer.

Part 1:



Part 2 (Hidden)

Part 3 (Hidden)

Here is an answer that describes "What happens next (in Part 2)?"

{{task.row_data['reference_text']}}

Evaluate the description (in green) based on the below criteria:

1. Correctness (Rate 1-5):

Does the description (in green) correctly answer the question (in red)?



Score: 3

Please explain why you rated this score:

2. Depth and Thoughtfulness (Rate 1-5):

How much depth in reasoning is covered in the description? Is there sufficient reasoning to support the description?



Score: 3

Please explain why you rated this score:

3. Level of Detail (Rate 1-5):

Does the description have sufficient level of detail to easily identify and differentiate between people/objects and actions?



Score: 3

Please explain why you rated this score:

4. Visual Contradiction:

Does the description reasonably follow or contradict what is happening in the video?

A description is a contradiction only if it is impossible for that scenario to happen, based on Part 1.



Please explain why you rated this score:

☐ *Optional Feedback:* Something about the HIT is unclear/These videos have hateful/offensive content/You have additional feedback:

Submit

Figure 11. (Continued) Human Evaluation for Generative Tasks

Instructions

Welcome to this task! Please read the question carefully and answer it. All parts must be completed before submitting.

Given the beginning of a video, what what could happen next?

Beginning:

▶ 0:00

Rest of the video (Hidden)

Explain what will happen next.
Write three possible diverse explanantions:

What happens next? Also describe WHY and HOW (rationale).

What happens next? Also describe WHY and HOW (rationale).

What happens next? Also describe WHY and HOW (rationale).

Please write at least 8 words per description. All explanations should be unique. Please avoid using "it", "he", "she" or "they"; instead use specific descriptions like "the cat" or "the boy on the skateboard".

☐ *Optional Feedback:* Something about the HIT is unclear/These videos have hateful/offensive content/You have additional feedback:

Submit

Figure 12. **Sample template for Human Baseline for the generative variant.**

Welcome!

Welcome to this task! Your goal is to watch the provided videos and choose the most accurate option from the given set of options.

Full Instructions (Expand/Collapse)

Task Instructions

Thanks for participating in this HIT! This is a task where you will look at videos and choose the right option for the given question.

Guidelines:

- Please read the options carefully.
- Please watch the video(s) completely.
- Be as reasonable as you can, and make a smart guess if you are unsure.

Task: Multiple Choice Question Answering

Part 1:


Part 2 (Hidden)

Part 3:


Question: Which of the following could have happened in between the two videos (in Part 2) shown here?

If no option seems correct, pick the most plausible (but not impossible) option.

- ☐ \${option1}
- ☐ \${option2}
- ☐ \${option3}

Please read all options carefully, as the options may vary in very specific ways.

Submit

Figure 13. Sample template for Human Baseline for the MCQ variant.

Welcome!

Welcome to this task! Your goal is to watch the provided videos and choose the most accurate option from the given set of options.

Full Instructions [\(Expand/Collapse\)](#)

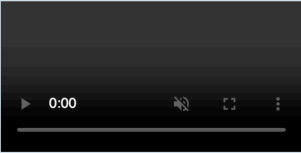
Task Instructions

Thanks for participating in this HIT! This is a task where you will look at videos, read a hypothesis, and choose the right option for the given question.

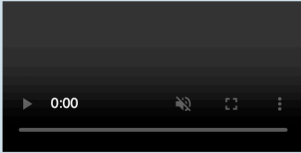
Guidelines:

- Please read the hypothesis carefully.
- Please watch the video(s) completely.
- Be as reasonable as you can, and make a smart guess if you are unsure.

Task: Multiple Choice Question Answering

Part 1:


Part 2 (Hidden)

Part 3


Hypothesis: `{{ task.row_data['exp'] }}`

Question: Given the beginning and end of a video, does this hypothesis hold?

Please watch the video carefully before answering.

☐ Yes

☐ No

Submit

Figure 14. Sample template for Human Baseline for the Y/N variant.

Welcome!

Welcome to this task! Please begin by reading the full instructions and the guidelines.

Full Instructions [\(Expand/Collapse\)](#)

Task Instructions

Thanks for participating in this task. Your goal is to describe a very short video clip (can be a small part of a longer video) to the best of your ability.

1. **Step 1:** First, watch the video carefully and think about the scene, the objects and people, and their actions.
2. **Step 2:** Next, describe the video in a few sentences. You must write at least 25 words, but longer descriptions are welcome.
3. **Step 3:** Check to ensure that you are not making any **assumptions** about the contents in the video, and that your description contains all aspects of the scene.

Example 1



Watch this clip and describe everything about the contents of the video:

A person is walking towards a body of water, may be a lake on a board walk with a fishing stick. Then, the person sees a silver sedan moving towards the water on his left. The person steps off the boardwalk towards the car hastily. The video is in the perspective of the person walking.

(a) 1 of 3

Figure 15. Analysis: Collection of Human Perception

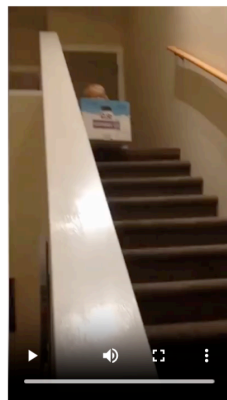
Example 2



Watch this clip and describe everything about the contents of the video:

A man walks down a small mud hill towards the camera. There appears to be an upside-down wheelchair to the right of the scene. The scene is set in a farm-like area, with a shed in the back and mostly grassy fields.

Example 3



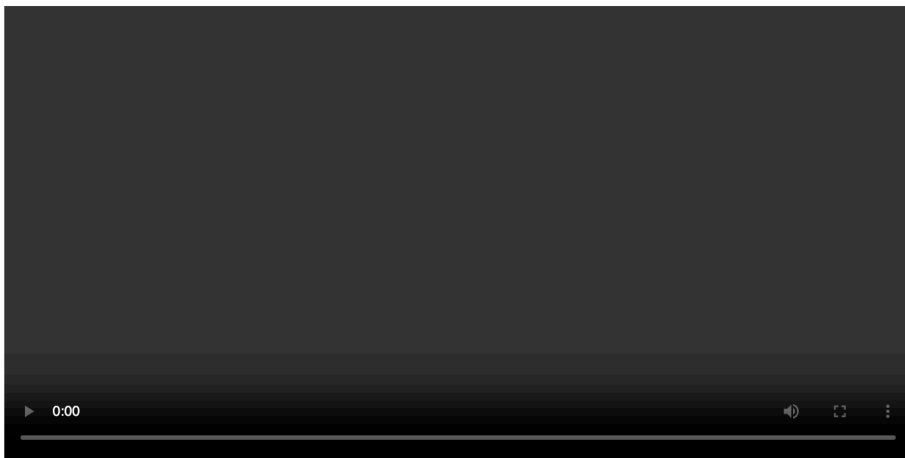
Watch this clip and describe everything about the contents of the video:

Two kids are sitting in a cardboard box, and they have started going down the stairs of the house while riding in the box. The camera is pointed looking up towards the stairs.

Guidelines:

- **Objects** - Describe the objects present in the scene and the state/location of the objects.
- **People** - Identify and describe the people in the video (eg. young child, old lady...).
- **Actions** - Describe the movements or actions in the video.
- **Scene** - Describe the scene and setting, and where the camera is pointed.

Task: Describe the events in this video



Watch this clip and describe what is happening in the video.

Please check that your description contains information about:

Scene **Objects** **People** **Actions**

Submit

(c) 3 of 3

Figure 15. (Continued) Analysis: Collection of Human Perception

Welcome!

Welcome to this task! Please begin by reading the full instructions and the guidelines.

Full Instructions [\(Expand/Collapse\)](#)

Task Instructions

Thanks for participating in this task. Your goal is to **describe the differences** between two short video clips, the beginning and end of a video. **Please note that the middle of the clip is hidden on purpose, so you should not make any assumptions about the contents of what happened in the middle.**

1. **Step 1:** First, watch both the beginning and the end video carefully and think about the scene, the objects and people, and their actions.
2. **Step 2:** Next, describe the differences you observe in the scene, objects, people and actions between the beginning and end of the video (in at least 15 words).
3. **Step 3:** Check to ensure that you are not making any **assumptions** about the contents in the videos, and that your description contains all aspects of the scene.

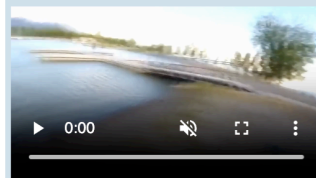
Example 1

Beginning:



Middle (Hidden)

End:



What are the differences in the scene between the beginning and the end?

In the beginning the car is on the beach whereas in the end, the car is in the water.

(a) 1 of 3

Figure 16. Analysis: Collection of Human Comprehension

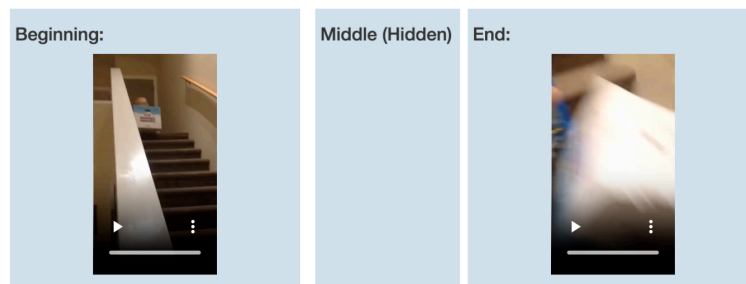
Example 2



What are the differences in the scene between the beginning and the end?

In the beginning, the man and the woman in the wheelchair are on top of a mud hill, whereas in the end, the wheelchair is upside down on the right of the frame, and the woman is no longer visible.

Example 3



What are the differences in the scene between the beginning and the end?

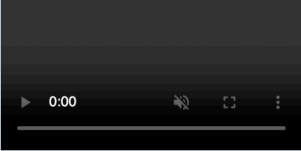
The two kids are in a box on the top of the stairs in the beginning. In the end, the kids are on the floor at the end of the staircase, and both look like they took a fall. The box seems to be upside down.

Guidelines:

- **Objects** - Describe the changes to the objects present in the scene and the state/location of the objects.
- **People** - Identify and describe what happened to the people in the video.
- **Actions** - Describe the movements or actions in the video (but do not make assumptions about what happened in the middle).
- **Scene** - Describe the scene and setting, and where the camera is pointed, before and after, if changed.

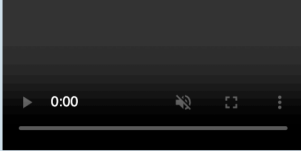
Task: Multiple Choice Question Answering

Beginning:



Middle (Hidden)

End:



What are the differences in the scene between beginning and end?

Please ensure you have read the examples in the instructions, to understand what is expected, before proceeding.

Please check that your description contains information about:

Scene Objects People Actions

Submit

(c) 3 of 3

Figure 16. (Continued) Analysis: Collection of Human Comprehension