

Appendix

A Benchmark Construction Details

A.1 Dataset Statistics

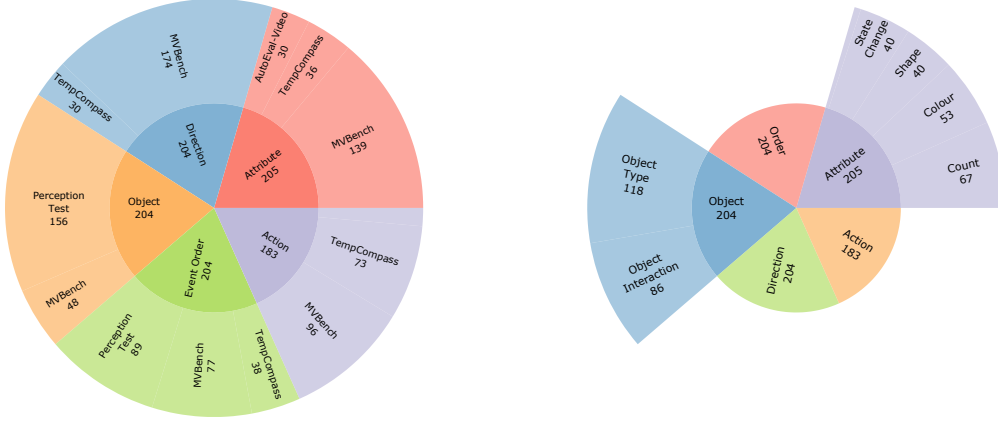


Figure 1: Distribution of visual instances in VIDHAL by (Left) public dataset source, categorized by the five temporal aspects, and (Right) temporal aspects and their sub-aspects.

Figure 1 presents the distribution of visual instances in VIDHAL by public dataset sources and temporal aspects. Additionally, Figure 2 further shows the distribution of ground truth answers for the MCQA and caption ordering tasks. One can observe that both temporal aspects and ground truth options are uniformly distributed across our benchmark. The distribution of video caption lengths and video durations is also presented in Figure 3.

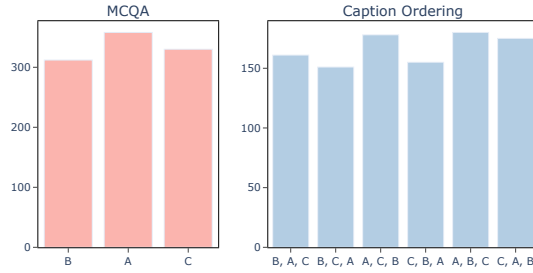


Figure 2: Distribution of (Left) correct answer options for the MCQA task and (Right) optimal option orders for the caption ordering task.

A.2 Dataset Development Pipeline

Visual Instance Selection To ensure a rich coverage of temporal aspects and visual diversity, we methodically selected video instances from four public datasets: TempCompass [5], Perception Test [7], MyBench [4], and AutoEval Video [1]. Given the unique characteristics of each dataset, we outline the specific guidelines adopted for each dataset below:

- **TempCompass** encompasses five temporal aspects: *Action*, *Speed*, *Direction*, *Event Order*, and *Attribute Change*. As most of these aspects align with those chosen to construct VIDHAL, we retain all video instances except those related to speed. TempCompass includes four evaluation tasks: *MCQA*, *Yes/No QA*, *caption matching*, and *caption generation*. Given the conciseness of captions in the latter two tasks, their information can often be subsumed within the more detailed

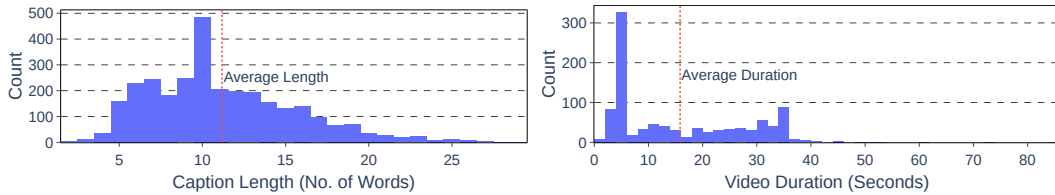


Figure 3: Distribution of (Left) caption lengths with an average of 11.2 words, and (Right) duration of videos in VIDHAL with an average of 15.8s.

Object Recognition [Object]: What object does the person use to hit other objects? What ingredients did the person put in the bowl or on the plate? Which object was removed by the person from the tabletop? What geometric shapes did the person put on the table? What objects did the person hit? What is the order of the letters on the table at the end? What letters did the person type on the computer in order? Distractor Action [Action]: What is the person preparing? Motion [Action]: What happens with the object after being placed on the slanted plane? What happened once the person removed an object from the tabletop?	Action Recognition [Action]: What object does the person use to hit other objects? What objects did the person hit? What is the person preparing? Which statement describes better the actions done by the person? Sequencing [Event Order]: What letters did the person show in order? What is the order of the letters at the end?", In what order did the person put the objects in the backpack? What is the order of the letters on the table at the end?
---	---

Figure 4: Specific skills and corresponding questions from the Perception Test dataset chosen for VIDHAL instance selection, with the matched aspects indicated in brackets.

25 QA-based annotations. Therefore, we focus exclusively on MCQA and Yes/No QA annotations to
 26 create an informative anchor caption.

- 27 • **Perception Test** spans various skill and reasoning domains to thoroughly evaluate VLLMs’
 28 perception and understanding abilities. Our inspection of these evaluation dimensions re-
 29 veals alignment between the *semantics*, *physics*, and *memory* skill areas, as well as *de-*
 30 *scriptive* and *explanatory* reasoning dimensions, with the temporal aspects of action, or-
 31 der, and event order. Accordingly, we limit our video selection in Perception Test to
 32 these specific pillars. Additionally, we review the question templates adopted in these ar-
 33 eas and select video instances with question-answer pairs that support VIDHAL’s evalua-
 34 tion objectives. The specific skills and associated questions chosen are detailed in Figure 4.
- 35 • **MVBench** includes twenty video understanding tasks with question-answer pairs designed to
 36 challenge the reasoning capabilities of VLLMs. Similar to the Perception Test, we identify the
 37 tasks relevant to the temporal aspects in VIDHAL and focus on collecting videos belonging from
 38 these tasks. The specific tasks for each aspect are presented in Figure5. We observe that MVBench
 39 contains repeated use of certain scenarios across tasks, indicated by similar question templates. To
 40 enhance caption diversity and minimize redundancy, we limit the number of examples for each
 41 unique scenario. The collected instances cover all five temporal aspects of VIDHAL.
- 42 • **AutoEval-Video** evaluates open-ended response generation in VLLMs through questions with
 43 detailed answers across nine skill dimensions. We focus on instances related to the *state transition*
 44 area, specifically assessing changes in object and entity attributes. For each instance, we retain the
 45 only answers to associated questions as they act as informative, long-form captions for the video.

46 **Incorrect Anchor Captions** A minority of
 47 videos contain anchor captions misaligned
 48 with their content, often due to noisy metadata.
 49 Such discrepancies subsequently lead to
 50 undesirable hallucinatory captions. To remove
 51 such instances, we use BLIP2 [3] to calculate
 52 frame-text matching scores across all video
 53 frames, selecting the maximum score as the
 54 representative video-text alignment score.
 55 Examples with incorrect anchor captions typically achieve low alignment scores, which are discarded
 56 as noisy instances.

57

58 **LLM-based Caption Generation** We utilize GPT-4o’s [6] text processing and generation capabili-
 59 ties to generate an anchor caption for each selected video, based on metadata from its original public
 60 dataset source. This metadata includes QA-based annotations for TempCompass, Perception Test, and
 61 MVBench, along with long-form answers for AutoEval-Video. The anchor caption is subsequently
 62 used as input for GPT-4o to generate corresponding hallucinatory captions.

63 To ensure the generated hallucinatory captions meet high-quality standards, we employ a detailed
 64 prompt adopting the following strategies to guide GPT-4o’s output:

Action: Action Sequence, Fine-Grained Action and Fine-Grained Pose
Direction: Moving Direction.
Object: Object Interaction, Object Existence.
Attribute: Moving Attribute, Moving Count.
Order: Action Sequence

Figure 5: Evaluation tasks in MVBench aligned with temporal aspects in VIDHAL, categorized by aspect.

- Aspect-specific definitions which outline the characteristics of each aspect to be varied, prompting GPT-4o to modify anchor captions accordingly.
 - Caption construction guidelines that define the structure, format, and hallucination levels required for the generated captions.
 - In-context examples to illustrate the desired form of each hallucinatory caption for each aspect.
- The prompts for generating anchor and hallucinatory captions are shown in Figures 6 to 8, respectively. Definitions for each aspect are provided in Figure 9, and aspect-specific in-context examples are detailed in Figures 10 to 14. Separate in-context examples are provided for each *Attribute* subaspect of *Shape*, *Size*, *Color*, *Count*, and *State Change* to account for their distinct natures.

You are given one or more questions targeted at content of a video and their corresponding answers. You are tasked with generating an appropriate and informative single line caption for the video using this information given to you. Ensure that you restrict yourself to only information present in the question-answer pairs provided. If the answers to the questions provide various types of information, concentrate on the color related to the subjects and objects in the video in your caption. Focus on providing clear and concise descriptions without using overly elaborate language.

<metadata>

Video description:

Figure 7: Prompts used for generating the anchor caption from QA-based annotations.

Caption Quality Scoring To identify video instances with the high quality generated captions, we utilize powerful LLMs to evaluate the quality of generated captions. The captions are assessed is based on three specific criteria:

- **Realism** determines whether generated scenarios are plausible.
- **Ordering Quality** evaluates whether the hallucination level ordering is appropriate.
- **Relevance** ensures that deviations from the anchor caption align with the designated aspect.

Binary questions are used to evaluate captions for each criterion, assigning a score of 1 for positive responses, *i.e.*, "yes", and 0 otherwise. The scores for each criterion are averaged across all models and prompts, and then summed across all criteria to produce a final quality assessment score for the generated captions of a video instance. We evaluate each set of captions using three LLMs: GPT-4o, Gemini-1.5 Flash [8], and LLaMA3 (70B) [2] along with three variants for each binary question.

You are given a long caption describing the content of a video. Your task is to provide a summarised and concise version of this caption. Ensure that you keep all essential detail in the original caption.

<metadata>

Video description:

Figure 6: Prompts used for generating the anchor caption from long-form captions.

You are a chatbot tasked with generating hallucinatory captions for a video given the input ground truth caption provided. Your objective is to modify the `<aspect>` present in the provided caption to generate 2 incorrect captions of different levels of hallucination. `<aspect_definition>`. The extent of hallucination of each caption is measured on a scale of 1 to 3 in increasing levels of hallucination, with 1 denoting no hallucinations present and 3 denoting a large extent of hallucination. A description of the extent of hallucination represented by each score is given as follows:

1. The caption contains no hallucination. The caption that representing this score is the ground truth caption.
2. The caption includes moderate hallucination, describing an event that is different from the ground truth, yet possible given the context of the video
3. The caption contains high hallucination, describing an event that is realistic, but typically unlikely to happen given context reflected by the original caption.

The generated hallucinated captions should follow the guidelines below.

Guidelines:

1. Focus only on modifying the temporal aspect provided in the instruction. Do not change any other temporal aspect associated with objects or subjects in the video.
2. Keep your modifications brief but coherent. Your generated captions should be of similar length to the original caption.
3. Ensure that your generated captions depict realistic and believable scenarios even as they deviate from the original context. For example, avoid creating fictitious scenarios such as "Person flying on a broomstick" and "Monkey painting a picture".
4. You may rephrase the provided caption to maintain consistent sentence structure across all captions. However, make sure the factual content of the ground truth caption remains unchanged.
5. Each generated hallucinatory caption should be of the form `<score> : <caption>`, `<score>` takes a value from the hallucination scale defined and `<caption>` represents your provided hallucinatory caption.
6. No two generated `<caption>` should share the same `<score>`, and each caption should take on a unique level of hallucination from 2 to 3.

Here are some examples of how hallucinatory captions are expected to be constructed.

`<in_context_examples>`

Now, generate hallucinatory captions for the following video description.

Original Caption:
`<anchor_caption>`
Hallucinated Captions:

Figure 8: Prompt for generating aspect-specific hallucinatory captions based on anchor captions and in-context examples.

This ensemble of both models and prompts enhances the robustness of our evaluation.. Figures 15 and 16 provide details of the criterion-specific quality assessment queries and the prompt templates employed for each LLM. We select the top 1,000 examples with the highest quality assessment scores to construct VIDHAL.

A.3 Additional Dataset Examples

We provide additional qualitative examples of video instances and their corresponding captions in Figure 17 for each of the five temporal aspects.

B Human Validation Details

B.1 Human Validation Process

Action: Actions refer to observable movements or activities performed by entities that may involve interaction with objects or the environment in the video.
Direction: Direction refers to the course or path along which objects or subjects move in the video.
Order: Order refers to the sequential arrangement of events that occur in the video.
Object: Objects refer to inanimate, physical entities or items present within the video.
State: State refers to the condition or status of an object or subject, indicating its current properties, position or the phase of action the subject is taking or phase of process the object is undergoing.
Count: Count refers to the frequency of an action being performed or an event occurring. It may also refer to the number of objects or subjects involved in an event or interaction.
Color: Color refers to the hue or shade of an object or subject.
Shape: Shape refers to the form or outline of an object or subject.
Size: Size refers to the dimensions or magnitude of an object or subject.

Figure 9: Definitions incorporated into the prompt for generating hallucinatory captions for each aspect, with separate definitions provided for each sub-aspect in the *Attribute* aspect.

<p>Original Caption: 1 : A boy inflates the balloon, which grows vertically. Hallucinated Captions: 2 : A boy inflates the balloon, which grows horizontally. 3 : A boy deflates the balloon, which shrinks horizontally.</p> <p>Original Caption: 1 : The bag expands in height as items are being placed inside. Hallucinated Captions: 2 : The bag expands in width as items are being placed inside. 3 : The bag shrinks in height as items are being placed inside.</p> <p>Original Caption: 1 : The size of the puddle of water is increasing. Hallucinated Captions: 2 : The size of the puddle of water is decreasing. 3 : The size of the puddle of water remains unchanged.</p>	<p>Original Caption: 1 : A circle shaped block is placed in a wooden box. Hallucinated Captions: 2 : A square shaped block is placed in a wooden box. 3 : A star shaped block is placed in a wooden box.</p> <p>Original Caption: 1 : Cubes are transforming into cylinders. Hallucinated Captions: 2 : Cubes are transforming into cones. 3 : Cubes are transforming into spheres.</p> <p>Original Caption: 1 : The clouds form a fluffy circle in the sky. Hallucinated Captions: 2 : The clouds form a fluffy square in the sky. 3 : The clouds form a fluffy triangle in the sky.</p>
---	---

Figure 10: In-context examples for the *Size* (Left) and *Shape* (Right) sub-aspects.

<p>Original Caption: 1 : A leaf with holes turns green to red. Hallucinated Captions: 2 : A leaf with holes turns from green to orange. 3 : A leaf with holes turns from yellow to orange.</p> <p>Original Caption: 1 : A yellow ball bounces on the ground, and lands in the pool. Hallucinated Captions: 2 : A red ball bounces on the ground, and lands in the pool. 3 : A blue ball bounces on the ground, and lands in the pool.</p> <p>Original Caption: 1 : A stationary purple cup appears at the beginning of the video. Hallucinated Captions: 2 : A stationary blue cup appears at the beginning of the video. 3 : A stationary green cup appears at the beginning of the video.</p>	<p>Original Caption: 1 : The man wearing a jacket performed three backflips. Hallucinated Captions: 2 : The man wearing a jacket performed four backflips. 3 : The man wearing a jacket performed five backflips.</p> <p>Original Caption: 1 : Four birds perched on the wire. Hallucinated Captions: 2 : Five birds perched on the wire. 3 : Six birds perched on the wire.</p> <p>Original Caption: 1 : One car drove down the road. Hallucinated Captions: 2 : Two cars drove down the road. 3 : Three cars drove down the road.</p>
---	---

Figure 11: In-context examples for the *Color* (Left) and *Count* (Right) sub-aspects.

<p>Original Caption: 1 : A red bucket of liquid goes from empty to half full. Hallucinated Captions: 2 : A red bucket of liquid goes from empty to completely full. 3 : A red bucket of liquid goes from completely full to empty.</p> <p>Original Caption: 1 : The light in the room is slowly dimming. Hallucinated Captions: 2 : The light in the room slowly dims, then brightens again. 3 : The light in the room is slowly getting brighter.</p> <p>Original Caption: 1 : The sky changes from clear to partly cloudy. Hallucinated Captions: 2 : The sky changes from clear to completely overcast. 3 : The sky changes from partly cloudy to clear.</p>

Figure 12: In-context examples for the *State* sub-aspect.

<p>Original Caption: 1 : The man hits another object with a bat. Hallucinated Captions: 2 : The man hits another object with a racket. 3 : The man hits another object with a broom.</p> <p>Original Caption: 1 : The ball bounces down the slanted plane. Hallucinated Captions: 2 : The ball rolls down the slanted plane. 3 : The ball zigzags down the slanted plane.</p> <p>Original Caption: 1 : A person puts two rectangles and one circle into the bag. Hallucinated Captions: 2 : A person puts a rectangle, a square and a circle into the bag. 3 : A person puts two squares and a circle into the bag.</p>	<p>Original Caption: 1 : A person puts a bottle in the bag. Then, he puts a book in the bag. Lastly, he puts a pencil case into the bag. Hallucinated Captions: 2 : A person puts a book in the bag. Then, he puts a bottle in the bag. Lastly, he puts a pencil case into the bag. 3 : A person puts a pencil case in the bag. Then, he puts a book in the bag. Lastly, he puts a bottle into the bag.</p> <p>Original Caption: 1 : A man writes letters in the following order: A, V, T, Y. Hallucinated Captions: 2 : A man writes letters in the following order: A, Y, T, V. 3 : A man writes letters in the following order: Y, T, V, A.</p> <p>Original Caption: 1 : A woman with white coat places a book on the table. She takes two vials of liquid and mixes them together. Hallucinated Captions: 2 : A woman with white coat places a book on the table. She takes off her coat. Then, she takes two vials of liquid and mixes them together. 3 : A woman with white coat takes two vials of liquid and mixes them together. She then places a book on the table.</p>
---	--

Figure 13: In-context examples for the *Object* (Left) and *Event-Order* (Right) aspects.

<p>Original Caption: 1 : The people are cooking in the video. Hallucinated Captions: 2 : The people are chopping in the video. 3 : The people are washing in the video.</p> <p>Original Caption: 1 : A car is driving down the road. Hallucinated Captions: 2 : A car is reversing down the road. 3 : A car is being repaired along the road.</p> <p>Original Caption: 1 : A dog is digging a hole near the tree. Hallucinated Captions: 2 : A dog is scratching the tree. 3 : A dog is barking at the tree</p>	<p>Original Caption: 1 : An eagle is flying from left to right diagonally upwards. Hallucinated Captions: 2 : An eagle is flying from left to right horizontally. 3 : An eagle is flying from left to right diagonally downwards.</p> <p>Original Caption: 1 : The car drives forward and makes a right turn. Hallucinated Captions: 2 : The car drives forward and continues driving straight. 3 : The car drives forward and makes a left turn.</p> <p>Original Caption: 1 : The ball on the table rolls away from the camera. Hallucinated Captions: 2 : The ball on the table rolls from left to right. 3 : The ball on the table rolls towards the camera.</p>
---	---

Figure 14: In-context examples for the *Action* (Left) and *Direction* (Right) aspects.

As varying hallucination levels are a distinctive feature of our benchmark, we prioritize validating the robustness of caption ordering produced by our annotation pipeline. Each anchor caption is derived from the original video metadata, making it the most accurate reflection of the video content. Our primary objective is to ensure that the ordering of hallucinatory captions aligns with human judgment. To achieve this, human annotators are shown the video instance along with both hallucinatory captions and are tasked with selecting the caption that better aligns with the video content, as illustrated in Figure 18. Each video instance is reviewed by multiple annotators, with the final human-aligned order determined through a majority vote and compared with our automatically generated order.

B.2 Misaligned Instances

Table 1 lists video instances that fail to meet the majority agreement threshold established by our annotation process. We additionally provide the corresponding human agreement scores for each instance.

Video ID	Agreement Score
action_55	0.429
action_88	0
action_90	0.308
action_118	0.200
action_153	0.250
order_60	0.500
order_109	0.154
attribute_90	0.400
attribute_180	0.071
attribute_192	0.188
object_25	0.375
object_170	0
direction_188	0.400

Table 1: Video examples with generated caption orders misaligned with the human-preferred order during the quality verification process. The agreement score indicates the proportion of human respondents who select the ordering generated in our annotations.

GPT-4o & Gemini-1.5 Flash:
You are provided with a ground truth description of a video, and 2 other captions that contain hallucinations in the aspect of `<aspect>`. The hallucinated captions are displayed in increasing order of hallucination, where the first caption contains the least amount of hallucinated elements and the last caption having significant hallucination. You are tasked with answering a question regarding the quality of the hallucinated captions. Provide your answer as detailed in the question, without further explanation of your answer.

Ground truth caption:
`<anchor_caption>`

Hallucinated captions:
`<hallucinatory_captions>`

Question:
`<quality_assessment_question>`

Answer:

LLaMA3 (70B):
`<|begin_of_text|><|start_header_id|>system<|end_header_id|>`
You are provided with a ground truth description of a video, and 2 other captions that contain hallucinations in the aspect of `<aspect>`. The hallucinated captions are displayed in increasing order of hallucination, where the first caption contains the least amount of hallucinated elements and the last caption having significant hallucination. You are tasked with answering a question regarding the quality of the hallucinated captions. Provide your answer as detailed in the question, without further explanation of your answer.
`<|eot_id|>`
`<|start_header_id|>user<|end_header_id|>`
Ground truth caption:
`<anchor_caption>`

Hallucinated captions:
`<hallucinatory_captions>`

Question:
`<quality_assessment_question>`

Answer:
`<|eot_id|>`
`<|start_header_id|>assistant<|end_header_id|>`

Figure 15: Prompt template for evaluating the quality of generated captions for the GPT-4o, Gemini-1.5 Flash, and LLaMA3 (70B) models.

Realism:

1. Is the scenario presented in caption `<option>` realistic? Provide your answer only as a single "yes" or "no".
2. Is the event in caption `<option>` believable? Provide your answer only as a single "yes" or "no".
3. Is the setting present in caption `<option>` plausible? Provide your answer only as a single "yes" or "no".

Order Quality:

1. Which caption better matches the ground truth description: Caption `<option_A>` or `<option_B>`? Provide your answer only as a single number (`<option_A>` or `<option_B>`)
2. Which caption aligns more closely with the ground truth description: Caption `<option_A>` or `<option_B>`? Provide your answer only as a single number (`<option_A>` or `<option_B>`)
3. Which caption is more faithful to the ground truth description: Caption `<option_A>` or `<option_B>`? Provide your answer only as a single number (`<option_A>` or `<option_B>`)

Relevance:

1. Does hallucinated caption `<option>` differ from the ground truth caption only in the `<aspect>`? Provide your answer only as a single "yes" or "no".
2. Is the only difference between hallucinated caption `<option>` and the ground truth caption the `<aspect>`? Provide your answer only as a single "yes" or "no".
3. Did hallucinated caption `<option>` change the ground truth caption only with respect to the `<aspect>`? Provide your answer only as a single "yes" or "no".

Figure 16: Question prompts for evaluating caption quality based on the three assessment criteria. Prompts with the placeholder `<option>` are applied individually to the anchor and hallucinatory captions. For question associated with *order quality*, `<option_A>` and `<option_B>` are replaced with the corresponding hallucinatory caption options shown to the LLMs.

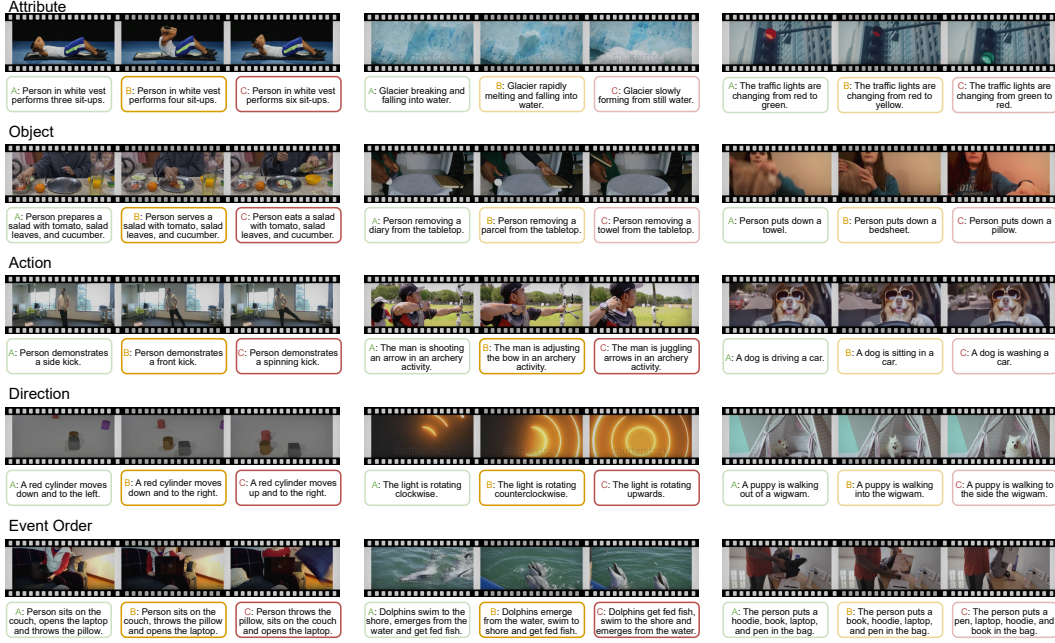


Figure 17: Qualitative examples of video instances and their corresponding generated captions in the VIDHAL Benchmark, across the five temporal aspects.

121 C Evaluation Pipeline Details

122 C.1 Evaluation Task Prompts

123 Figures 19 and 20 present the prompts used for the MCQA and naive caption ordering tasks,
 124 respectively. The same prompt used for both the MCQA task and the paired questions in the
 125 relative caption ordering task. Our manual inspection of these instances reveals that these videos
 126 often feature visually complex content, making them challenging even for human annotators.

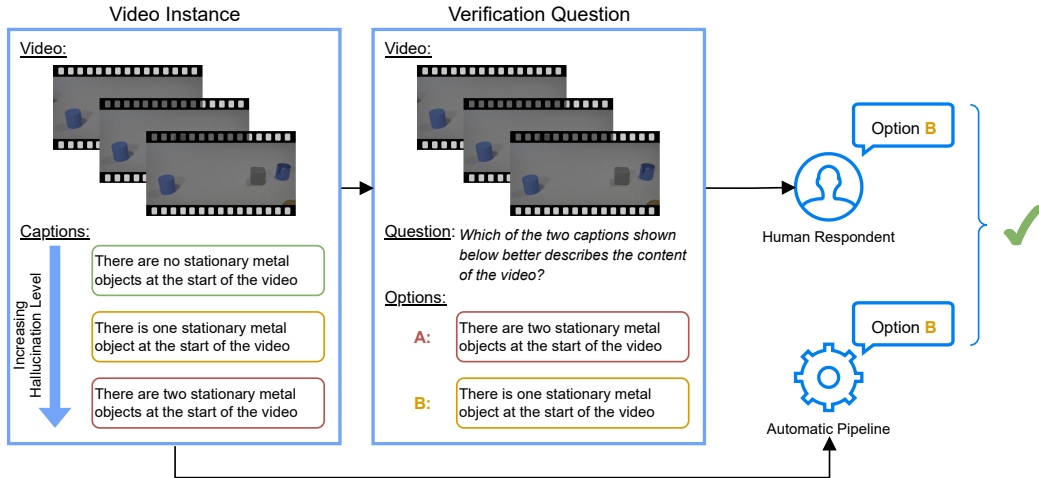


Figure 18: Pipeline for validating the quality of generated caption orders in VidHal. For each instance, human annotators are provided with the video and its associated hallucinatory captions. The annotators then select the caption that best aligns with the video content. The selected response is subsequently checked for consistency with the caption with lower hallucination according to our annotation process.

You are provided with a video and a set of several captions. Your task is to watch the video provided carefully, and select the caption that best describes the video. Provide your answer only as a single letter representing the option whose caption that best describes the video, without any explanation.

Watch the video provided, and choose the option whose caption describes the video most accurately.

A. <caption_A>

B. <caption_B>

Figure 19: Prompt template for the MCQA and relative caption ordering evaluation tasks.

Watch the video provided, and rank the captions below in order from the most accurate to the least accurate in describing the video. Provide your response only as a sequence of comma separated option letters matching the corresponding captions. Do not give any additional explanation for your answer.

For example, if option B contains the caption that best describes the video, option A contains the caption that describes the video second best and option C contains the caption that describes the video least accurately, provide your response as: B, A, C.

A. <caption_A>

B. <caption_B>

C. <caption_C>

Figure 20: Prompt template for the naive caption ordering evaluation task.

127 C.2 Relative Order Parsing

128 Prompting the VLLM to predict the order of captions based on their hallucinatory level in the relative
 129 caption ordering task involves asking a series of paired questions derived from different caption
 130 combinations. However, providing the model with all possible pairs at once may result in cyclic and
 131 non-transitive orderings. To address this, we present each caption pair to the VLLM in a systematically
 132 selected sequence, beginning with two paired questions. The final paired question is presented to the
 133 model to resolve inconsistencies if the multiple possible orderings can be derived from the responses
 134 to the first two paired questions. The responses across all paired questions presented to the VLLM is
 then parsed according to the workflow illustrated in Figure 21.

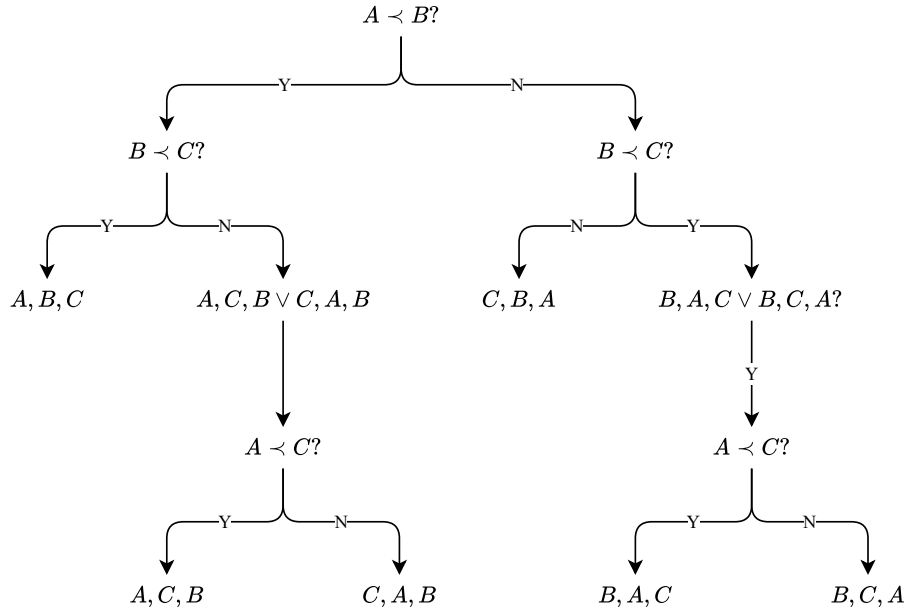


Figure 21: Decision tree for determining the final caption order based on VLLM responses to paired questions in the relative caption ordering evaluation task.

D Additional Experiments

D.1 Input Order Sensitivity

To assess the robustness of VLLM responses to the order of displayed captions, we conducted additional experiments by evaluating three VLLMs using a fixed static display order across all instances. We repeated this process across all different permutations of input caption order, presenting the results of these models in Figure 22. We observe that the performance of these VLLMs is highly sensitive to the order in which captions are displayed, reflected by their varying results across different order permutations. This instability intensifies with smaller model sizes, with VideoLLaMA2 (7B) showing the highest variance in evaluation results and VideoLLaMA2 (72B) the lowest. Our findings suggest that VLLMs may be particularly vulnerable to input caption order, potentially confounding their performance.

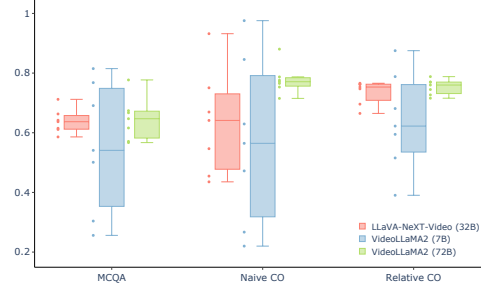


Figure 22: Distribution of results of VLLMs across varied input caption orders for the three evaluation tasks.

D.2 Model Consensus

We further assess the reliability of VLLM responses by measuring their *consensus* on the MCQA task, defined as the proportion of shared correct predictions across correctly answered instances between model pairs. As shown in Figure 23, advanced VLLMs typically show higher consensus with other models sharing similar performance. In contrast, smaller models display less consistency in reasoning, with VideoChat2 demonstrating especially low response agreement among the smaller VLLMs.

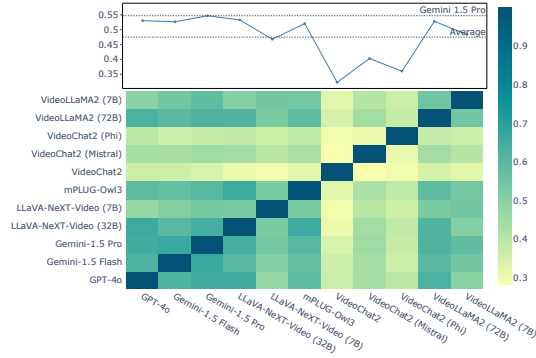


Figure 23: (Top) Averaged consensus score of each respective model, and (Bottom) consensus between each model pairs. *Average* represents the average consensus score across all VLLMs, with Gemini-1.5 Pro achieving the highest.

References

- [1] Xiuyuan Chen, Yuan Lin, Yuchen Zhang, and Weiran Huang. Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering. *CoRR*, abs/2311.14906, 2023.
- [2] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024.
- [3] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR, 2023.
- [4] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Lou, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206. IEEE, 2024.
- [5] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics*, pages 8731–8772. Association for Computational Linguistics, 2024.
- [6] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.
- [7] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adrià Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchet, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems*, 2023.
- [8] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, and et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530, 2024.