

VIDHALLUC: Evaluating Temporal Hallucinations in Multimodal Large Language Models for Video Understanding

Supplementary Material

A. Additional Related Works

Different neural network architectures, trained on distinct datasets, develop unique inductive biases that influence their feature representation capabilities. For instance, the CLIP series [15, 17, 21], pretrained with text-image contrastive alignment, excels at capturing global semantic features. In contrast, the DINO series [14, 22], pretrained with vision-only contrastive learning, specializes in fine-grained perception and object-level details. Prior research shows that integrating features from complementary networks can lead to a more balanced and robust model behavior [6, 16].

Building on these strengths, COMM [8] demonstrates the effectiveness of integrating features from different layers of CLIP and DINOv2 to enhance the visual capabilities of multimodal large language models (MLLMs). By leveraging the complementary nature of these architectures, COMM achieves improved performance across various visual tasks. Similarly, CLIP-DINOiser [19] uses localization priors from DINO to refine CLIP’s global image features, resulting in smoother and more precise outputs for semantic segmentation tasks. Furthermore, Nguyen *et al.* [13] explore the multi-level features of DINO to fine-tune the final block of CLIP. This approach not only tackles the challenge posed by the limited scale of training datasets in deepfake detection but also enhances model interpretability through the use of attention mechanisms. These studies collectively highlight the potential of combining the strengths of diverse architectures to overcome individual limitations, achieve superior performance, and improve explainability in complex visual tasks. However, these methods often rely on specific architectures and require additional adjustments or training for integration, which limits their flexibility and scalability.

To address these limitations, our proposed DINO-HEAL offers a flexible and architecture-agnostic solution that can be applied at the inference stage without modifying the underlying model structures. Unlike previous approaches, DINO-HEAL does not require additional training parameters or fine-tuning, making it particularly suitable for mitigating hallucinations in resource-constrained scenarios.

B. Implementation Details

B.1. Data Collection on Existing Datasets

To apply our data collection pipeline to the selected datasets, we employ a structured process for segmenting and pairing videos. Specifically, for ActivityNet [7] and YouCook2 [24], videos are divided into multiple action-based segments. We

then compute similarities between all segment pairs within each video to identify pairs meeting the similarity criteria. For VALOR32K [3], videos are randomly paired, and their similarities are calculated to determine if they satisfy the conditions. The resulting filtered video pairs, characterized by high semantic similarity but low visual similarity, form the core of our dataset, enabling effective investigation into hallucination phenomena.

B.2. Prompts for Different Hallucination Tasks

The following subsections show the prompts we use to test models on each hallucination task: action hallucination (ACH), temporal sequence hallucination (TSH), and scene transitional hallucination (STH).

B.2.1. Prompt for Binary QA in ACH

```
<Video>
Is the primary action in the video
{Action}?
Only answer with "No" or "Yes".
```

The placeholder {Action} in the prompt is dynamically replaced with a specific action, such as ‘turning the steering wheel’. To further enhance the diversity of the benchmark, each Binary QA question is randomly assigned one of four templates: “Is the prominent action in the video {Action}?”, “Does the video primarily feature {Action}?”, “Is the key action shown in the video {Action}?”, or “Is the primary action in the video {Action}?” These variations introduce linguistic diversity while preserving the semantic meaning.

B.2.2. Prompt for MCQs in ACH

```
<Video>
"Question": "What is the prominent
action in the video?" Please select
the correct answer (one or more options),
only return the choice letter (i.e., A,
B, C, D) of your answer(s).

"Choices":
"A": "{Action A}"
"B": "{Action B}"
"C": "{Action C}"
"D": "{Action D}"
```

The placeholders {Action A}, {Action B}, {Action C}, and {Action D} are dynamically replaced with specific actions, such as “wakeboarding,” “changing gears,” “adjusting the

rearview mirror,” and “turning the steering wheel.” To introduce linguistic diversity and enhance the robustness of the benchmark, each MCQ is randomly assigned one of the following templates: “*What is the prominent action in the video?*”, “*What is the key action shown in the video?*”, “*What is the primary action in the video?*”, or “*What is the predominant action captured in the video?*”. These variations ensure that the benchmark reflects a range of natural question formulations while maintaining consistency in meaning.

B.2.3. Prompt for Sorting Questions in TSH

```
<Video>
Below are two actions in the video:
Action A. {Action A}
Action B. {Action B}

Sort these two actions in the order they
occur in the video, and return which
action happens before which one. For
example, "Action A before Action B" or
"Action B before Action A". If you only
detect one action of these two in the
video, return that action.
```

The placeholders {Action A}, {Action B} are replaced with specific actions. For instance, with the actions of skiing and driving a car, the prompt will look as the following example:

“Below are two actions in the video: Action A. skiing, Action B. driving a car. Sort these two actions in the order they occur in the video, and return which action happens before which one. For example, ‘Action A before Action B’ or ‘Action B before Action A’. If you only detect one action of these two in the video, return that action.”

B.2.4. Prompt for Open-ended Questions in STH

<Video>

A scene change is defined as a significant transition in the overall environment or location within the video. This means a change from one distinct setting to another, such as moving from a kitchen to a living room, or from indoors to outdoors. Watch the given video and determine if a scene change occurs. If there is a scene change, respond in the format: "Scene change: Yes, Locations: from [location] to [location2]." If no change occurs, respond: "Scene change: No, Locations: None".

B.3. Additional Dataset Statistics

Figure 1 displays a word cloud of our benchmark, providing a more intuitive presentation of VIDHALLUC. As shown



Figure 1. The wordcloud of VIDHALLUC.

in the figure, the questions in our benchmark are diverse and prominently feature action-related terms, such as “playing,” “walking,” “cutting,” “cleaning,” and “jumping.” These terms highlight the dynamic nature of the video content in our benchmark, emphasizing the focus on activities and interactions within videos. The word cloud reflects the breadth of actions and events covered, including activities like “mixing ingredients,” “pouring water,” “riding,” and “driving.” This diversity aligns with our goal of capturing the challenges associated with action recognition, temporal coherence, and scene understanding in videos. We believe our benchmark can effectively reveal potential hallucination issues in MLLMs, especially those related to understanding complex actions in dynamic video content.

C. More Quantitative Results on VIDHALLUC

We provide additional metrics and detailed scores of state-of-the-art MLLM performance on VIDHALLUC. For open source models, we include Video-ChatGPT [12], VideoLLaVA [10], ShareGPT4Video [2], Chat-UniVi [9], LLaVA-NeXT-Video [23], PLLaVA [20], VideoLLaMA2 [4], and VILA 1.5 [11]. For proprietary models, we select Gemini-1.5-Pro [18], and GPT-4o [1].

Quantitative results on ACH. Table 1 presents two distinct versions of the quantitative results for these models on the ACH task. The binary QA and MCQ scores are computed by dividing the number of correct answers by the total number of questions, following the metric defined as:

$$\text{Accuracy} = \frac{N_{\text{correct}}}{N_{\text{total}}}, \quad (1)$$

where N_{correct} and N_{total} denote non-negative counts of correct answers and total answers. In contrast, binary QA Pair and MCQ Pair scores are based on a stricter criterion requiring both questions in a pair to be answered correctly. This stricter evaluation ensures that the model fully understands both semantically similar but visually different videos.

As mentioned in the main paper, an interesting observation is that the accuracy for MCQ is higher than that for

Models	Binary QA	Binary QA Pair	MCQ	MCQ Pair
Video-ChatGPT [12]	9.50	0.19	24.58	5.56
Video-LLaVA [10]	26.84	9.87	64.45	40.34
ShareGPT4Video [2]	29.96	10.65	44.78	19.18
Chat-UniVi [9]	23.77	6.39	54.79	29.37
LLaVA-NeXT-Video [23]	26.60	12.00	77.57	60.03
PLLaVA [20]	35.30	16.26	76.96	59.94
VideoLLaMA2 [4]	50.04	29.09	<u>83.84</u>	<u>69.85</u>
VILA1.5 [11]	58.46	37.77	81.88	67.95
Gemini-1.5-Pro [18]	<u>75.27</u>	<u>59.10</u>	79.25	63.36
GPT-4o [1]	81.15	66.79	90.95	83.00

Table 1. Performance comparison of existing models on action hallucination (ACH). The numbers in the table represent accuracy percentages (%). **Bold** numbers denote the best performance, and underlined numbers indicate the second-best performance.

binary questions. This result defies common intuition, as one might expect binary questions, with only two possible answers, to be inherently simpler and, therefore, yield higher accuracy compared to MCQs, which involve selecting from multiple options. This discrepancy suggests that models may leverage contextual or comparative cues more effectively in MCQ scenarios, while binary questions might require more precise reasoning or direct understanding, exposing potential weaknesses in model comprehension.

Quantitative results on STH In STH, we benchmark MLLMs with the new criterion that evaluates both the classification of the scene and whether the model describes it in the correct sequence. For the classification scores, we use the Matthews correlation coefficient (MCC) to evaluate the model predictions against the ground truth labels.

$$\frac{n_{11} \times n_{10} - n_{01} \times n_{00}}{\sqrt{(n_{11} + n_{01})(n_{11} + n_{00})(n_{10} + n_{01})(n_{10} + n_{00})}}, \quad (2)$$

where $A \in \{0 \text{ (False)}, 1 \text{ (True)}\}$ represents the actual condition, $P \in \{0 \text{ (Negative)}, 1 \text{ (Positive)}\}$ represents the predicted condition, and n_{AP} denotes non-negative counts. To adjust MCC to range between 0 and 1 and to further penalize models that consistently answer only “Yes” or only “No”, we apply the transformation in order to adjust MCC to range between 0 and 1, obtaining the classification score $\text{Score}_{\text{cls}}$:

$$\text{Score}_{\text{cls}} = \left(\frac{\text{MCC} + 1}{2} \right)^2. \quad (3)$$

The description task measures the model’s ability to accurately identify and articulate the information of the scene. To evaluate this, scene descriptions are extracted from both the model’s output and the ground truth, structured for direct comparison. We then calculate the cosine similarity S between the SimCSE [5] embeddings of the corresponding scenes. Based on this similarity measure, each scene description score is calculated as:

Models	Score _{cls}	Score _{desc}	Score _{overall}
Video-ChatGPT [12]	5.07	11.65	7.70
Video-LLaVA [10]	25.00	36.50	29.60
ShareGPT4Video [2]	29.55	0.26	17.83
Chat-UniVi [9]	30.12	29.50	29.87
LLaVA-NeXT-Video [23]	55.91	27.14	44.40
PLLaVA [20]	29.86	36.32	32.44
VideoLLaMA2 [4]	87.43	31.67	<u>65.12</u>
VILA1.5 [11]	25.00	50.07	35.03
Gemini-1.5-Pro [18]	71.88	<u>52.08</u>	63.96
GPT-4o [1]	<u>80.17</u>	58.69	71.58

Table 2. Performance comparison of existing models on scene transition hallucination (STH). We assign a weight of 0.6 to the classification task and 0.4 to the description task. The numbers in the table represent accuracy percentages (%). **Bold** numbers denote the best performance, and underlined numbers indicate the second-best performance.

$$\text{Score}_{\text{desc}} = \begin{cases} 0, & \text{if } S \leq \text{THR}_{\text{low}} \\ \frac{\sigma(S) - \sigma(\text{THR}_{\text{low}})}{\sigma(1) - \sigma(\text{THR}_{\text{low}})}, & \text{if } S > \text{THR}_{\text{low}} \end{cases}, \quad (4)$$

where S denotes the cosine similarity between the SimCSE [5] embeddings of the corresponding scenes, THR_{low} represents the minimum threshold for assigning a score, and σ is the Sigmoid function. The overall description score is calculated as the average of the score for the “from” and “to” scenes. Finally, the overall evaluation score is computed as a weighted sum of the classification score and the normalized description score:

$$\text{Score}_{\text{overall}} = \alpha \times \text{Score}_{\text{cls}} + (1 - \alpha) \times \text{Score}_{\text{desc}}. \quad (5)$$

Table 2 summarizes the decomposed scores of classification and description scores for the STH category. An interesting observation is that both Video-LLaVA and VILA 1.5 achieve a classification score of 25% by consistently answering “Yes” to all questions, irrespective of their actual ability to recognize transitions between locations. This pattern, highlighted in the 10th row of Table 8, exposes a critical limitation in both the MCC metric and the model themselves. Their reliance on default affirmative response reveals a superficial understanding of spatial transition and suggests a lack of deeper reasoning.

Future work should focus on developing mechanisms to penalize such oversights and promote consistency in model behavior, ensuring that metrics better reflect genuine understanding and performance. This includes designing metrics or loss functions that discourage uniform responses and promote adaptive reasoning based on context.

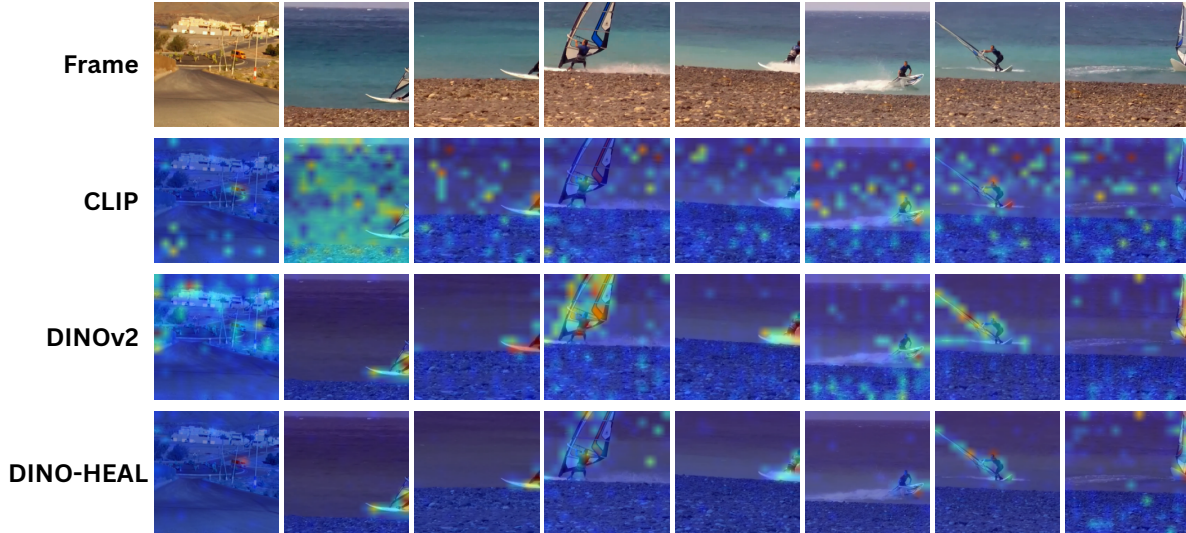


Figure 2. The visualization of saliency maps.

D. DINO-HEAL

D.1. Saliency Analysis

Figure 2 illustrates the input frames, saliency maps generated by CLIP and DINOv2, and the adjusted saliency map produced by DINO-HEAL. The saliency maps generated by CLIP often show significant noise, which can be attributed to its inductive bias toward capturing global contextual frame information. This characteristic, while beneficial for understanding broader scene-level features, can result in a lack of focus on specific, spatially important regions within the frame. In contrast, DINOv2 demonstrates a stronger capacity to localize and emphasize key objects within the scene, leveraging its vision-only contrastive learning to identify fine-grained details efficiently. The adjusted saliency maps created by DINO-HEAL reflect an integration of these two approaches, balancing the strengths of both CLIP and DINOv2. By mitigating the noisiness of CLIP’s global focus and incorporating the precise localization capabilities of DINOv2, DINO-HEAL effectively emphasizes spatially significant features. This result strongly supports our hypothesis that DINO-HEAL serves as a complementary mechanism to CLIP, enhancing its ability to prioritize critical regions and improving overall spatial feature representation.

D.2. Additional Results on VIDHALLUC

Tables 3 and 4 show the results of the ACH and STH tasks when baseline models are augmented with our hallucination mitigation method, DINO-HEAL. A particularly noteworthy observation is the significant improvement in scores for Video-LLaVA and VILA 1.5 on the STH task. Previously, these models consistently defaulted to answering “Yes” regardless of the correct location-based response. With the

integration of DINO-HEAL, however, they demonstrate an improved ability to discern and appropriately respond with “No” when necessary, as elaborated in Section C. This indicates a meaningful enhancement in their spatial reasoning and decision-making capabilities. This improvement underscores the potential of DINO-HEAL to refine spatial reasoning and address their shortcomings effectively.

Tables 5 and 6 further detail the results for Video-ChatGPT, Video-LLaVA, and VideoLLaMA2 on the ACH and TSH tasks, comparing performance with and without DINO-HEAL. For the ACH task, we use the binary QA accuracy metric. Without DINO-HEAL, none of the models correctly predict all question pairs, though VideoLLaMA2 is able to infer the second question pair accurately. When DINO-HEAL is applied, however, all models can predict both pairs accurately, showcasing the method’s effectiveness in mitigating hallucinations. On the TSH task, we observe further improvements. Initially, Video-ChatGPT recognizes that two actions occur simultaneously, while the ground truth sequence is “starting a fire” followed by “gutting a fish”. Video-LLaVA and VideoLLaMA2 only identify one action, “starting a fire”. After integrating DINO-HEAL, all three models correctly identify and sequence both actions, underscoring the method’s ability to enhance temporal understanding in complex tasks.

E. More Qualitative Results on VIDHALLUC

Tables 7, 8, and 9 present randomly selected examples showcasing multiple model responses for the ACH, TSH, and STH tasks, respectively. For the ACH task, the Binary QA Pair metric is used, which applies a stricter evaluation criterion requiring both questions in a pair to be answered correctly. In particular, the adversarially crafted pairs in

Models	Binary QA	Binary QA Pair	MCQ	MCQ Pair
Video-ChatGPT	9.50	0.19	24.58	5.56
+DINO-HEAL	13.96 _{+4.46}	0.42 _{+0.23}	28.81 _{+4.23}	6.76 _{+1.20}
Video-LLaVA	26.84	9.87	64.45	40.34
+DINO-HEAL	33.80 _{+6.96}	14.17 _{+4.3}	66.25 _{+1.8}	41.64 _{+1.3}
ShareGPT4Video	29.96	10.65	44.78	19.18
+DINO-HEAL	30.41 _{+0.45}	9.73 _{-0.92}	44.43 _{-0.35}	18.62 _{-0.56}
VILA1.5	58.46	37.77	81.88	67.95
+DINO-HEAL	60.63 _{+2.17}	40.34 _{+2.57}	81.85 _{-0.03}	67.90 _{-0.05}
VideoLLaMA2	50.04	29.09	83.84	69.85
+DINO-HEAL	50.01 _{-0.03}	29.07 _{-0.02}	83.84 _{+0.0}	69.85 _{+0.0}

Table 3. Performance comparison of models on action hallucination (ACH), with and without DINO-HEAL. Improvements from DINO-HEAL are shown as subscripts. **Bold** numbers denote the best performance after applying DINO-HEAL.

Models	Score _{cls}	Score _{desc}	Score _{overall}
Video-ChatGPT	5.07	11.65	7.70
+DINO-HEAL	5.56 _{+0.49}	12.15 _{+0.5}	8.20 _{+0.5}
Video-LLaVA	25.00	36.50	29.60
+DINO-HEAL	27.89 _{+2.89}	35.18 _{-2.61}	30.81 _{+0.69}
ShareGPT4Video	29.55	0.26	17.83
+DINO-HEAL	28.80 _{-0.75}	2.78 _{+2.52}	18.39 _{+0.56}
VILA1.5	25.00	50.07	35.03
+DINO-HEAL	26.66 _{+1.66}	50.38 _{+0.21}	36.15 _{+1.12}
VideoLLaMA2	87.43	31.67	65.12
+DINO-HEAL	89.19 _{+1.76}	31.63 _{-0.04}	66.17 _{+1.05}

Table 4. Performance comparison of models on scene transition hallucination (STH), with and without DINO-HEAL. We assign a weight of 0.6 to the classification task and 0.4 to the description task. Improvements from DINO-HEAL are shown as subscripts. **Bold** numbers denote the best performance applying DINO-HEAL.

the ACH category require MLLMs to discern between semantically similar actions, posing a significant challenge for accurate interpretation. None of the models could predict both pairs accurately, highlighting the complexity of the ACH task. Moreover, Table 8 details the description scores for each example. These examples demonstrate that locations highly relevant to the ground truth achieve high scores. For instance, the phrase “in a swimming pool” exhibits strong semantic description similarity with “pool” or “a swimming pool,” while “in a bathtub” aligns well with “the bathtub”. Conversely, locations with differing semantic descriptions or inaccurate interpretations receive lower description scores. For example, “in a swimming pool” has limited similarity to descriptions such as “indoor bathroom” or “indoor pool,” and “in a bathtub” similarly diverges from “indoor bathroom”. These results highlight the importance of semantic alignment in achieving accurate description scores.

DINO-HEAL Example, Action Hallucination

Clip 1



Clip 2



Question 1 Is the primary action in the video mixing the ingredients?
 Question 2 Is the primary action in the video watching the oven?

Ground Truth
 Answer for Question 1 with Clip 1: **Yes**.
 Answer for Question 1 with Clip 2: **No**.
 Answer for Question 2 with Clip 1: **No**.
 Answer for Question 2 with Clip 2: **Yes**.

Video-ChatGPT
 Answer for Question 1 with Clip 1: **Yes**, the primary action in the video is mixing the ingredients.
 Answer for Question 1 with Clip 2: **Yes**, the primary action in the video is mixing the ingredients.
 Answer for Question 2 with Clip 1: **No**. The man and the child are making cookies in the kitchen.
 Answer for Question 2 with Clip 2: **No**.

Video-ChatGPT + DINO-HEAL
 Answer for Question 1 with Clip 1: **Yes**.
 Answer for Question 1 with Clip 2: **No**.
 Answer for Question 2 with Clip 1: **No**.
 Answer for Question 2 with Clip 2: **Yes**.

Video-LLaVA
 Answer for Question 1 with Clip 1: **Yes**.
 Answer for Question 1 with Clip 2: **Yes**.
 Answer for Question 2 with Clip 1: **Yes**.
 Answer for Question 2 with Clip 2: **Yes**.

Video-LLaVA + DINO-HEAL
 Answer for Question 1 with Clip 1: **Yes**.
 Answer for Question 1 with Clip 2: **No**.
 Answer for Question 2 with Clip 1: **No**.
 Answer for Question 2 with Clip 2: **Yes**.

VideoLLaMA2
 Answer for Question 1 with Clip 1: **Yes**.
 Answer for Question 1 with Clip 2: **Yes**.
 Answer for Question 2 with Clip 1: **No**.
 Answer for Question 2 with Clip 2: **Yes**.

VideoLLaMA2 + DINO-HEAL
 Answer for Question 1 with Clip 1: **Yes**.
 Answer for Question 1 with Clip 2: **No**.
 Answer for Question 2 with Clip 1: **No**.
 Answer for Question 2 with Clip 2: **Yes**.

Table 5. An example of the action hallucination category of VIDHALLUC dataset with and without DINO-HEAL. **Green** text indicates correct answers, and **red** text indicates incorrect answers.


DINO-HEAL Example, Temporal Sequence Hallucination	
	
Question	<p>Below are two actions in the video:</p> <p>Action A. gutting a fish</p> <p>Action B. starting a fire</p> <p>Sort these two actions in the order they occur in the video, and return which action happen before which one. For example, ‘Action A before Action B’ or ‘Action B before Action A’. If you only detect one action of these two in the video, return that action.</p>
Ground Truth	BA.
Video-ChatGPT	In the video, the man is seen gutting a fish and starting a fire. The order of these two actions is not clear, as they both occur simultaneously .
Video-ChatGPT + DINO-HEAL	In the video, the man starts a fire by striking a match on a piece of wood. After that , he goes to the beach and starts gutting a fish .
Video-LLaVA	Action B . starting a fire.
Video-LLaVA + DINO-HEAL	Action B happens before Action A .
VideoLLaMA2	I only detect one action in the video, which is Action B : starting a fire.
VideoLLaMA2 + DINO-HEAL	Action B . starting a fire happens before Action A . gutting a fish.
ShareGPT4Video	Answer: ‘Action B happens before Action A ’
ShareGPT4Video + DINO-HEAL	Answer: ‘Action B happens before Action A ’
VILA1.5	Action B . starting a fire happens before Action A . gutting a fish.
VILA1.5 + DINO-HEAL	Action B . starting a fire happens before Action A . gutting a fish.

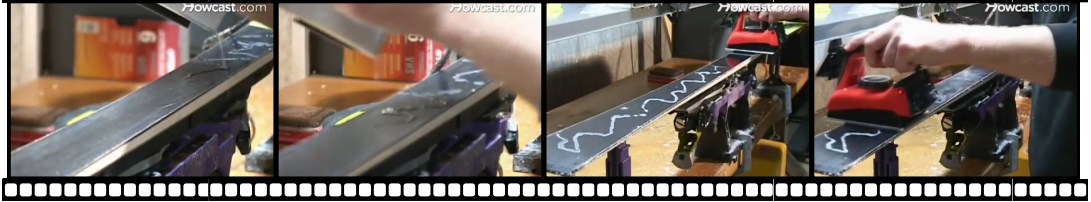
Table 6. An example from the temporal sequence hallucination category of VIDHALLUC dataset with and without DINO-HEAL. **Green** text indicates correct answers, and **red** text indicates incorrect answers.

VIDHALLUC Example, Action Hallucination

Clip 1



Clip 2



Question 1 Is ironing wax on the ski the main activity in the video?
 Question 2 Is the primary action in the video skiing?

Ground Truth
 Answer for Question 1 with Clip 1: **No**.
 Answer for Question 1 with Clip 2: **Yes**.
 Answer for Question 2 with Clip 1: **Yes**.
 Answer for Question 2 with Clip 2: **No**.

Chat-Univi [9] Answer for Question 1 with Clip 1: **No**, the main activity in the video is skiing down a snowy slope.
 Answer for Question 1 with Clip 2: **No**, the main activity in the video is not ironing wax on the ski.
 The video showcases a person using a variety of tools and techniques to create snowboards.
 Answer for Question 2 with Clip 1: **Yes**, the primary action in the video is skiing.
 Answer for Question 2 with Clip 2: **Yes**, the primary action in the video is skiing.

PLLaVA [20] Answer for Question 1 with Clip 1: **Yes**.
 Answer for Question 1 with Clip 2: **Yes**.
 Answer for Question 2 with Clip 1: **Yes**.
 Answer for Question 2 with Clip 2: **Yes**.

Video-ChatGPT [12] Answer for Question 1 with Clip 1: **Yes**, the main activity in the video is ironing wax on the ski.
 Answer for Question 1 with Clip 2: **No**, the main activity in the video is not ironing wax on the ski.
 The video shows a man using a machine to scrape wood.
 Answer for Question 2 with Clip 1: **No**, the primary action in the video is not skiing.
 Answer for Question 2 with Clip 2: **No**, the primary action in the video is not skiing. The video shows a man working on a piece of wood with a tool.

LLaVA-NeXT-Video [23] Answer for Question 1 with Clip 1: **Yes**.
 Answer for Question 1 with Clip 2: **Yes**.
 Answer for Question 2 with Clip 1: **Yes**.
 Answer for Question 2 with Clip 2: **Yes**.

Table 7. An example from the action hallucination category of VIDHALLUC dataset. **Green** text indicates correct answers, and **red** text indicates incorrect answers.

VIDHALLUC Example, Temporal Sequence Hallucination



Question

Below are two actions in the video:

Action A. unloading shingles onto the roof

Action B. removing old shingles

Sort these two actions in the order they occur in the video, and return which action happens before which one. For example, 'Action A before Action B' or 'Action B before Action A'. If you detect only one of the two actions in the video, return that action.

Ground Truth

AB.

LLaVA-NeXT-Video [23]

Action **B**. removing old shingles.

Chat-UniVi [9]

Based on the video, Action **B** (removing old shingles) occurs **before** Action **A** (unloading shingles onto the roof).

Video-LLaVA [10]

Action **B**. removing old shingles happens **before** Action **A**. unloading shingles onto the roof.

PLLaVA [20]

Action **B**. removing old shingles.

Gemini-1.5-Pro [18]

The order of actions is **BA**:

1. ****B. removing old shingles:**** A person is shown using a shovel to remove old shingles from the roof.
2. ****A. unloading shingles onto the roof:**** A crane is used to lift new shingles onto the roof.

GPT-4o [1]

BA.

Video-ChatGPT [12]

In the video, the person is first seen unloading shingles onto the roof, and then they are seen removing old shingles. The order in which these actions occur is **unloading shingles onto the roof** first, **followed by removing old shingles**.

ShareGPT4Video [2]

Answer: 'Action **A** happens **before** Action **B**'.

VideoLLaMA2 [4]

Action **A**. unloading shingles onto the roof happens **before** Action **B**. removing old shingles.

VILA1.5 [11]

Action **B**. removing old shingles happens **before** Action **A**. unloading shingles onto the roof.

Table 8. An example from the temporal sequence hallucination category of the VIDHALLUC dataset. **Green** text indicates correct answers, and **red** text indicates incorrect answers.

VIDHALLUC Example, Scene Transition Hallucination



Question

A scene change is defined as a significant transition in the overall environment or location within the video. This means a change from one distinct setting to another, such as moving from a kitchen to a living room or from indoors to outdoors. Watch the given video and determine if a scene change occurs. If there is a scene change, respond in the format: ‘Scene change: Yes, Locations: from [location1] to [location2].’ If no change occurs, respond: ‘Scene change: No, Locations: None’.

Ground Truth Scene change: **Yes**; Locations: from **in a swimming pool** to **in a bathtub**.

Video-ChatGPT [12] Scene change: **No**; Locations: **None**.
 $Score_{desc1} = 0, Score_{desc2} = 0$

LLaVA-NeXT-Video [23] Scene change: **Yes**; Locations: from **pool** to **bathtub**.
 $Score_{desc1} = 0.727, Score_{desc2} = 0.892$

Chat-UniVi [9] Scene change: **Yes**; Locations: from **indoor pool** to **outdoor pool**.
 $Score_{desc1} = 0.514, Score_{desc2} = 0$

Video-LLaVA [10] Scene change: **Yes**; Locations: from **the pool** to **the bathtub**.
 $Score_{desc1} = 0.862, Score_{desc2} = 0.961$

PLLaVA [20] Scene change: **Yes**; Locations: from **indoor bathroom** to **indoor bathroom**.
 $Score_{desc1} = 0, Score_{desc2} = 0.266$

ShareGPT4Video [2] Scene change: **No**; Locations: **None**.
 $Score_{desc1} = 0, Score_{desc2} = 0$

VideoLLaMA2 [4] Scene change: **Yes**; Locations: from **pool** to **bathtub**.
 $Score_{desc1} = 0.727, Score_{desc2} = 0.892$

VILA1.5 [11] Scene change: **Yes**; Locations: from **[location1]** to **[location2]**.
 $Score_{desc1} = 0, Score_{desc2} = 0$

Gemini-1.5-Pro [18] Scene change: **Yes**; Locations: from **pool** to **bathtub**.
 $Score_{desc1} = 0.727, Score_{desc2} = 0.892$

GPT-4o [1] Scene change: **Yes**; Locations: from **a swimming pool** to **a bathtub**.
 $Score_{desc1} = 0.941, Score_{desc2} = 0.946$

Table 9. An example from the scene transition hallucination category of the VIDHALLUC dataset. **Green** text indicates correct answers, and **red** text indicates incorrect answers. Each model’s description performance is evaluated using two scores: $Score_{desc1}$ and $Score_{desc2}$, derived from Equation 4. These scores correspond to the model’s ability to describe the two distinct scenes in the video accurately. The model’s overall $Score_{desc2}$ is computed as the average of these two scores.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmerschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv 2303.08774*, 2023. 2, 3, 9, 10
- [2] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. Sharegpt4video: Improving video understanding and generation with better captions. In *NeurIPS*, 2024. 2, 3, 9, 10
- [3] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weinong Wang, Jinhui Tang, and Jing Liu. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv 2304.08345*, 2023. 1
- [4] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv 2406.07476*, 2024. 2, 3, 9, 10
- [5] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *EMNLP*, 2021. 3
- [6] Gregor Geigle, Chen Cecilia Liu, Jonas Pfeiffer, and Iryna Gurevych. One does not fit all! on the complementarity of vision encoders for vision and language tasks. In *RepL4NLP*, 2023. 1
- [7] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 1
- [8] Dongsheng Jiang, Yuchen Liu, Songlin Liu, Jin’e Zhao, Hao Zhang, Zhen Gao, Xiaopeng Zhang, Jin Li, and Hongkai Xiong. From clip to dino: Visual encoders shout in multi-modal large language models. *arXiv preprint arXiv:2310.08825*, 2023. 1
- [9] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *CVPR*, 2024. 2, 3, 8, 9, 10
- [10] Bin Lin, Yang Ye, Bin Zhu, Jiayi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *EMNLP*, 2024. 2, 3, 9, 10
- [11] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *CVPR*, 2024. 2, 3, 9, 10
- [12] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *ACL*, 2024. 2, 3, 8, 9, 10
- [13] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Exploring self-supervised vision transformers for deepfake detection: A comparative analysis. *arXiv preprint arXiv:2405.00355*, 2024. 1
- [14] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024. 1
- [15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *PMLR*, 2021. 1
- [16] Cristian Rodriguez-Opazo, Edison Marrese-Taylor, Ehsan Abbasnejad, Hamed Damirchi, Ignacio M. Jara, Felipe Bravo-Marquez, and Anton van den Hengel. Unveiling backbone effects in clip: Exploring representational synergies and variances. *arXiv preprint arXiv:2312.14400*, 2023. 1
- [17] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv 2303.15389*, 2023. 1
- [18] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv 2403.05530*, 2024. 2, 3, 9, 10
- [19] Monika Wyszczanska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzcinski, and Patrick Pérez. Clip-dinoiser: Teaching clip a few dino tricks for open-vocabulary semantic segmentation. In *ECCV*, 2024. 1
- [20] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava : Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv 2404.16994*, 2024. 2, 3, 8, 9, 10
- [21] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 1
- [22] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *ICLR*, 2023. 1
- [23] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. LLaVA-NeXT: A strong zero-shot video understanding model, 2024. 2, 3, 8, 9, 10
- [24] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018. 1