ELV-Halluc: Benchmarking Semantic Aggregation Hallucinations in Video Understanding

Supplementary Material

7. Extended Analysis of ELV-Halluc

7.1. Rationale for Employing Adversarial Pairs

According to previous research[11, 17, 28], designing adversarial QA pairs can enhance the robustness of benchmarks in evaluation. Meanwhile, we aim to prevent models from answering questions correctly merely by guessing. For standard binary (yes/no) QA evaluations, models have a 50% chance of guessing the correct answer directly. However, under the evaluation mechanism using adversarial QA, if a model makes random guesses and incorrectly responds "No" to a hallucinated caption, adding a corresponding ground truth (GT) pair with the mandatory answer "Yes" can reduce the model's random guessing accuracy to 25%. (As most models cannot distinguish between GT and hallucinated questions, they tend to provide consistent answers, which further lowers the actual accuracy—with the minimum dropping to 1%.) Furthermore, we isolate and study SAH independently by designing an "in-out pair" method. The approach of subtracting the results of the two QA tasks can further improve the stability of the evaluation.

7.2. Necessity of the ELV-Halluc Metric

Many previous studies have been fraught with controversies regarding whether hallucinations can be distinguished from temporal localization, perception disability, and other issues, as well as how to separate hallucinations themselves from other disabilities. Therefore, we conducted a study by comparing the SAH-ratio and Accuracy from ELV-Halluc with those from four existing general multimodal large model benchmarks (Video-MME[7], which represents general video understanding capability; Charades-STA[9], which denotes temporal localization capability in video understanding; MLVU[42], which reflects models' long video understanding capability; and HallusionBench[11], which manifests models' perceptual hallucination on single images). We adopted models from the Qwen2.5-VL series, Qwen3-VL series, and InternVL3 series, calculated the Spearman correlation of these models across the six metrics.

As shown in figure 12, the three metrics most correlated with SAH-ratio are Charades-STA (0.74), HallusionBench (0.32), and ELV-Halluc Accuracy (0.18), which represent the model's temporal localization capability, the model's perceptual hallucination on single images, and the overall hallucination evaluation that includes partial SAH, respectively. This aligns perfectly with our hypothesis about SAH: it occurs when a model accurately perceives frame-level semantics but hallucinates during the process of semantic aggregation from frame-level to event-level. In contrast, the correlation coefficients between SAH-ratio and the remaining general video understanding metrics are all less than 0.2. This indicates that SAH-ratio reflects an aspect not captured by previous benchmarks, and a practical new metric is indeed needed to measure this specific issue.

Meanwhile, the overall accuracy reflects the model's general hallucination capability. Since it encompasses hallucinations caused by perceptual and comprehension errors such as perception disability and suboptimal frame sampling, accuracy is actually strongly correlated with general perceptual and comprehension abilities. Experimental results validate the above argument: ELV-Halluc Accuracy exhibits high correlations with general video understanding benchmarks—for instance, Video-MME (0.93), MLVU (0.86), and HallusionBench (0.83).

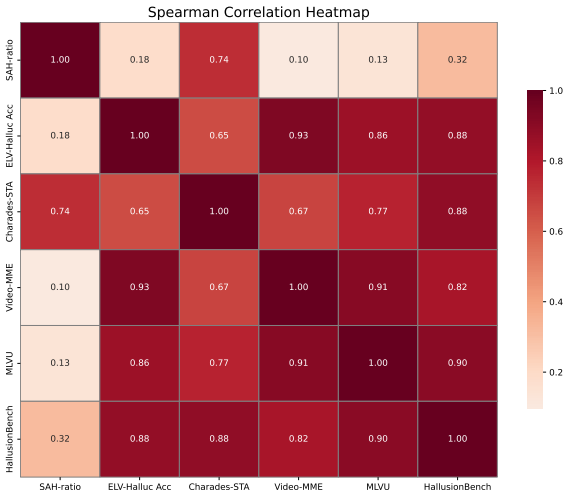


Figure 12. Spearman Correlation Between Six Multimodal Large Language Model (MLLM) Metrics. Note that SAH-ratio was multiplied by -1, as a lower value indicates better performance.

7.3. Stability Assessment Across Multiple Runs

As shown in Figure 13, to systematically evaluate the stability of the ELV-Halluc benchmark, we conducted triplicate experiments (three separate trial runs) using four representative models spanning different scales and model series. The experimental results consistently demonstrate that ELV-Halluc yields relatively stable overall accuracy and

SAH ratio across all trial runs, with no substantial fluctuations in these core metrics—confirming the benchmark's robustness to experimental variability.

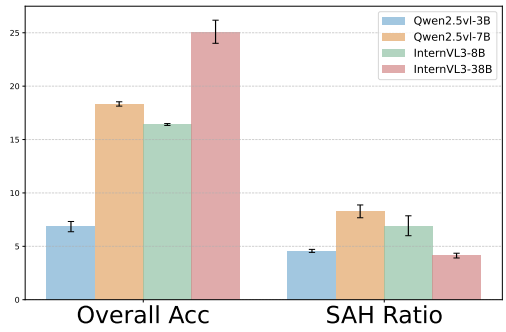


Figure 13. Stability evaluation of ELV-Halluc through repeated experiments. We report overall accuracy and SAH ratio across 3 runs for four representative models of different sizes and series, demonstrating that ELV-Halluc maintains consistent performance.

7.4. In-depth analysis of SAH mechanism

We analyze SAH on Qwen2.5-VL-7B via visual-token attention gaps between misled in-video QA and GT cases to isolate injected in-video semantics.

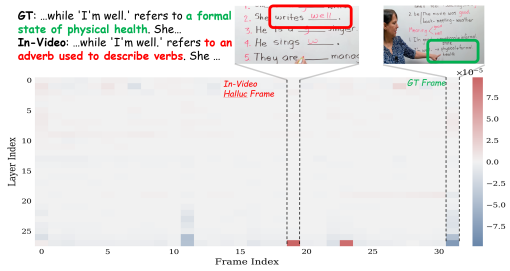


Figure 14. Difference heatmap (in-video - GT) in an SAH case.

Attention heatmaps show that model assigns higher attention to distracting in-video semantics at both the second and the final layers. Fig. 14 illustrates a case where delta attention emerges at the last layer. Such result indicates that SAH affects both early and deep stages in the LLM.

To quantify semantic aggregation, we define the temporal visual attention centroid $C^{(l)}$ for each layer l as: $C^{(l)} = \sum_{i=0}^{N-1} i\, a_i^{(l)} \,/\, \sum_{j=0}^{N-1} a_j^{(l)}$, where $a_i^{(l)}$ is the attn weight of i-th frame. We calculate the in-video offset $\Delta C^{(l)}$:

$$\Delta C^{(l)} = |C^{(l)}\text{In-V} - Tgt| - |C^{(l)}\text{GT} - Tgt| \quad (2)$$

Positive $\Delta C^{(l)}$ indicates attention centroid drift from T_{gt}.

Tab. 5 shows consistently larger ΔC for in-video cases, indicating stronger attention drift, especially for declarative content(higher-level semantics induce stronger semantic shifts). Layer-wise, shifts are negligible early but increase with depth, peaking at the deep stage.

Table 5. Analysis of Layer-wise Attention Centroid Shifts.

Perspective	Detail / Stage	In-Video ΔC (↑)	Out-Video ΔC (↑)
Visual Details	Average	0.0256	0.0066
Object	Average	0.0324	0.0113
Action	Average	0.0234	0.0231
Declarative Content	Average	**0.0792**	**0.0467**
Early Stage	Layers 1–9	0.0011	0.0016
Mid Stage	Layers 10–18	0.0039	0.0061
Deep Stage	Layers 19-28	**0.1154**	**0.0581**

We assume that SAH arises from progressive semantic capture: early-layer distractions by in-video semantics are compounded across layers, leading large attn centroid shifts in deep layers. Suggesting that SAH reflects high-level semantic integration failure instead of perception errors.

Our strategies follow above hypothesis. Stronger RoPE variants better capture temporal dependencies [37], reducing progressive semantic drift across layers. Since attention shifts are driven by semantic bias, DPO suppresses in-video distractors, reducing deep-layer attention shifts and alleviating SAH.

7.5. Implementation Details of the Model

Model Checkpoints For all evaluated open-source models, we adopt the official weights provided on HuggingFace. For the experiments involving RoPE, we directly utilize the Qwen2-VL-based implementation and weights from VideoRoPE. The specific HuggingFace repository paths and versions of the closed-source models, are listed in Table 6.

Inference Parameters We provide a detailed description of the parameter configurations used during evaluation. For the Qwen2.5-VL series, we set the maximum video resolution to $128 \times 28 \times 28$ pixels and the maximum number of frames to 256, while all other settings follow the official defaults. For the InternVL3 series, we limit the number of frames to 64. For GPT-4o, we performed uniform sampling 50 frames from each video and provided the model with following instruction:

"Here are {nframes} frames sampled from a {video_length}-duration video."

For all other models, we use the default generation hyperparameters specified in their respective official checkpoints to maintain consistency across experimental setups.

DPO Experiment Setting All experiments are conducted on NVIDIA H100 GPUs. During the DPO training process, we train for one epoch with a learning rate of 1×10^{-6} and a batch size of 16. We use Qwen2.5-VL-7B as the base model, with the sequence length set to 32,768 during training. All original parameters (e.g., `fps`, `fps max frames`, `video max pixels`) are preserved. During

Model	Checkpoint
Qwen2.5-VL-3B	Qwen/Qwen2.5-VL-3B-Instruct
Qwen2.5-VL-7B	Qwen/Qwen2.5-VL-7B-Instruct
Qwen2.5-VL-32B	Qwen/Qwen2.5-VL-32B-Instruct
Qwen2.5-VL-72B	Qwen/Qwen2.5-VL-72B-Instruct
Qwen3-VL-2B-Instruct	Qwen/Qwen3-VL-2B-Instruct
Qwen3-VL-4B-Instruct	Qwen/Qwen3-VL-4B-Instruct
Qwen3-VL-8B-Instruct	Qwen/Qwen3-VL-8B-Instruct
Qwen3-VL-32B-Instruct	Qwen/Qwen3-VL-32B-Instruct
Qwen3-VL-2B-Thinking	Qwen/Qwen3-VL-2B-Thinking
Qwen3-VL-4B-Thinking	Qwen/Qwen3-VL-4B-Thinking
Qwen3-VL-8B-Thinking	Qwen/Qwen3-VL-8B-Thinking
Qwen3-VL-32B-Thinking	Qwen/Qwen3-VL-32B-Thinking
InternVL3-1B	OpenGVLab/InternVL3-1B
InternVL3-2B	OpenGVLab/InternVL3-2B
InternVL3-8B	OpenGVLab/InternVL3-8B
InternVL3-14B	OpenGVLab/InternVL3-14B
InternVL3-38B	OpenGVLab/InternVL3-38B
InternVL3-78B	OpenGVLab/InternVL3-78B
SmolVLM2-2.2B-Instruct	HuggingFaceTB/SmolVLM2-2.2B-Instruct
Video-ChatGPT-7B	MBZUAI/Video-ChatGPT-7B
LLaVA-OneVision-Qwen2-7B	llava-hf/llava-onevision-qwen2-7b-ov-hf
GPT-4o(for modification)	gpt-4o-2024-08-06
GPT-4o(for evaluation)	gpt-4o-2024-11-20
Gemini2.5-flash	gemini-2.5-flash
Qwen2-VL-vanilla-rope	Wiselnn/Qwen2-VL-vanilla_rope-128frames-8k-context-330k-llava-video
Qwen2-VL-tad-rope	Wiselnn/Qwen2-VL-tad_rope-128frames-8k-context-330k-llava-video
Qwen2-VL-m-rope	Wiselnn/Qwen2-VL-m_rope-128frames-8k-context-330k-llava-video
Qwen2-VL-video-rope	Wiselnn/Qwen2-VL-videorope-128frames-8k-context-330k-llava-video

Table 6. Summary of evaluated models and their corresponding official HuggingFace checkpoints.

inference, we fix `nframes` to 64 for both **ELV-Halluc** and **VideoMME** benchmarks. Additional ELV-Halluc Case

8.

Studies As shown in Figure 15, we present the word cloud of ELV-Halluc captions to illustrate the diversity of captions.

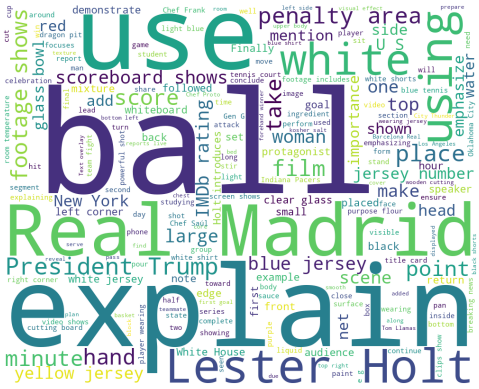


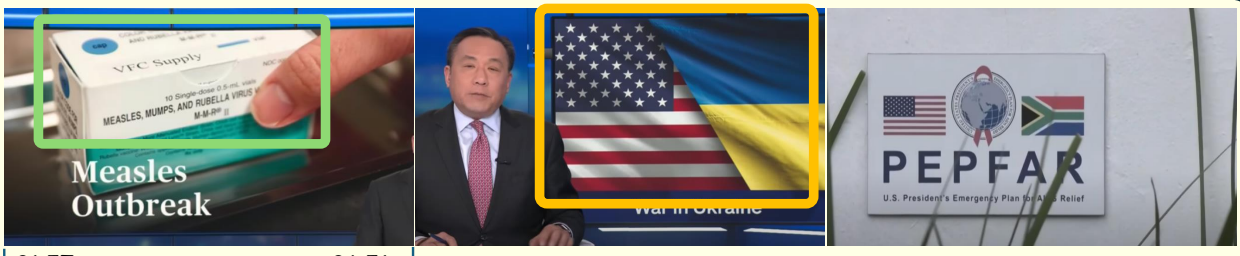
Figure 15. Word cloud of all captions in ELV-Halluc.

8.1. Illustrative ELV-Halluc Question-Answer Examples

In the following section, we present detailed examples of ELV-Halluc QA triplets covering four aspects: visual details (Figure 16), objects (Figure 17), actions (Figure 18), and declarative content (Figure 19). Each triplet includes three questions (ground truth, in-video hallucination, out-video hallucination), highlighted in green, orange, and red respectively; corresponding colored boxes mark these parts in the video.

8.2. Representative ELV-Halluc SAH Cases

In the following section, we present cases illustrating SAH occurrence in ELV-Halluc. The typical SAH scenario is: ground truth = "yes", In-Video response = "yes", Out-Video response = "no", indicating the model correctly identifies Out-Video hallucination but is misled by in-video semantics. As shown in Figures 20–23, these examples confirm ELV-Halluc effectively captures and evaluates SAH, with incorrect model answers highlighted in red.



01:57 ——— 06:51

Ground Truth: Is the following caption totally correct? Reply with "Yes" or "No" only.
01:57-06:51 - PBS News Weekend anchor John Yang sits at the desk reporting on the measles outbreak. A screen behind him displays **a box labeled 'Measles, Mumps, and Rubella Virus Vaccine.'** He then shows a map of the United States with 12 states marked in purple to indicate measles cases. The top displays the number '198,' representing the number of cases. He introduces infectious disease epidemiologist Jessica Malaty Rivera from the Debomont Foundation, who appears via video call to discuss the nature of measles, the effectiveness of the MMR vaccine, and the impact of vaccine refusal, noting that populations become vulnerable when vaccination rates drop below 95%. Images of healthcare workers handling vaccine bottles and vaccinating infants are shown. Yang inquires about the Trump administration's preparedness, and Rivera expresses concerns about 'widespread attacks on public health institutions' and misinformation about measles treatment.
Ref Answer: Yes

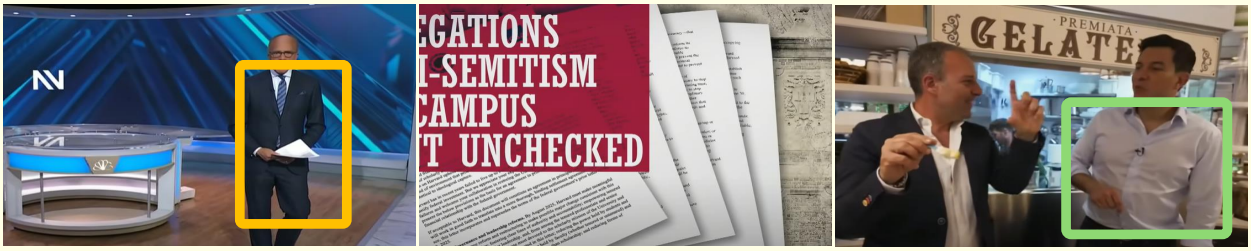
In-Video: 01:57-06:51 - PBS News Weekend anchor John Yang sits at the desk reporting on the measles outbreak. A screen behind him displays **split flags of the United States and Ukraine.** He then shows a map of the United States with 12 states marked in purple to indicate measles cases. ...
Ref Answer: No

Out-Video: 01:57-06:51 - PBS News Weekend anchor John Yang sits at the desk reporting on the measles outbreak. A screen behind him displays **a bottle labeled 'COVID-19 Rapid Antigen Test.'** He then shows a map of the United States with 12 states marked in purple to indicate measles cases. ...
Ref Answer: No

Figure 16. Example of *visual details* QA triplets from ELV-Halluc.

9. Prompt Templates and Examples

We provide all prompts used to construct ELV-Halluc. Figure 24 presents the prompt employed to guide Gemini-2.5-Flash in generating scratch captions for each event-by-event video. Figures 25, 26, 27, 28 display our prompts for injecting hallucinations of visual details, objects, actions, and declarative content, respectively. Figure 29 illustrates the prompt design for instructing GPT-4o to conduct rechecks of hallucinated captions.



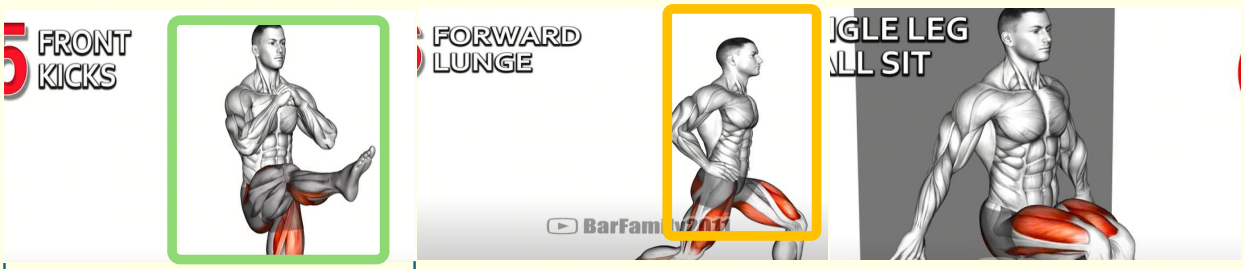
18:30 — 19:28

Ground Truth: Is the following caption totally correct? Reply with "Yes" or "No" only.
18:30-19:28 - Lester Holt introduces a whimsical piece about the Pope having significant meaning related to Pope Francis, who frequents a nearby Vatican City bakery named for the ivy growing on its walls. The footage shows a reporter in the shop talking to the owner, who is dressed in a black suit with a white shirt underneath and hands a spoonful of white ice cream to the reporter in **a white shirt**, who takes a taste. The report mentions Pope Francis' favorite desserts being lemon and mango ice cream, and he has requested hundreds of cakes to share with the homeless.
Ref Answer: Yes

In-Video: 18:30-19:28 - ...who is dressed in a black suit with a white shirt underneath and hands a spoonful of white ice cream to the reporter in **a dark suit and striped tie**, who takes a taste. The report mentions Pope Francis' favorite desserts being lemon and mango ice cream, and he has requested hundreds of cakes to share with the homeless.
Ref Answer: No

Out-Video: 18:30-19:28 - ...who is dressed in a black suit with a white shirt underneath and hands a spoonful of white ice cream to the reporter in **a green jacket**, who takes a taste. The report mentions Pope Francis' favorite desserts being lemon and mango ice cream, and he has requested hundreds of cakes to share with the homeless.
Ref Answer: No

Figure 17. Example of *object* QA triplets from ELV-Halluc.



02:55 — 03:33

Ground Truth: Is the following caption totally correct? Reply with "Yes" or "No" only.
02:55-03:33 - The male character in the animation performs a **front kick**. After standing upright, he alternately raises one knee to his chest and extends the leg forward in a controlled kicking motion, using hip flexors and quadriceps. He maintains balance on the standing leg.
Ref Answer: Yes

In-Video: 02:55-03:33 - The male character in the animation performs a **forward lunge**. After standing upright, he alternately raises one knee to his chest and extends the leg forward in a controlled kicking motion, using hip flexors and quadriceps. He maintains balance on the standing leg.
Ref Answer: No

Out-Video: 02:55-03:33 - The male character in the animation performs a **cartwheel**. After standing upright, he alternately raises one knee to his chest and extends the leg forward in a controlled kicking motion, using hip flexors and quadriceps. He maintains balance on the standing leg.
Ref Answer: No

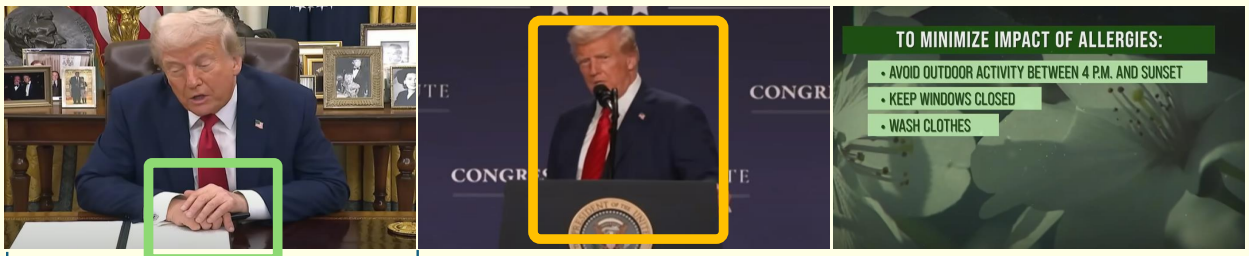
Figure 18. Example of *action* QA triplets from ELV-Halluc.

Ground Truth: 15:13-15:48 - Marco Asensio, wearing the No.20 green Real Madrid jersey, receives an accurate pass and sprints down the left flank. He dribbles past Barcelona's No.3 Gerard Piqué and powerfully shoots with his left foot, sending the ball spinning into the top left corner of the goal, with Barcelona goalkeeper Marc-André ter Stegen (yellow jersey, No.1) unable to stop it. Asensio celebrates the goal by lifting his jersey and pointing to the crowd, joined in celebration by teammates including No.17 Lucas Vázquez and No.14 Casemiro. **Real Madrid fans cheer fervently as the final score shows FCB 1-3 RMA.**
Ref Answer: Yes

In-Video: 15:13-15:48 - Marco Asensio, wearing the No.20 green Real Madrid jersey, receives an accurate pass and sprints down the left flank. He dribbles past Barcelona's ...**The referee issues a yellow card to Asensio for the shirt removal action, and the scoreboard reads FCB 1-3 RMA.**
Ref Answer: No

Out-Video: 15:13-15:48 - Marco Asensio, wearing the No.20 green Real Madrid jersey, receives an accurate pass and sprints down the left flank. He dribbles past Barcelona's ...**Barcelona manager looks visibly upset as Real Madrid fans cheer fervently, while the final score shows FCB 1-3 RMA.**
Ref Answer: No

Figure 19. Example of *declarative content* QA triplets from ELV-Halluc.



05:16 — 07:49

Ground Truth: Is the following caption totally correct? Reply with "Yes" or "No" only.
05:16-07:49 - Lester Holt turned to President Trump's pledged dual tariffs on the auto industry. Visuals included President Trump **signing documents,** ports filled with containers, and assembly lines at car factories. Garrett Haake interviewed Ford dealer owner Jim Sevet and autoworker James Benson in Dearborn, Michigan. …
Ref Answer: Yes
Qwen2.5-VL-3B:Yes Qwen2.5-VL-7B :Yes Qwen2.5-VL-72B :Yes InternVL3-14B:Yes InternVL3-78B:Yes

In-Video:
05:16-07:49 - …Visuals included Trump's **inauguration and interactions with the media**, ports filled with containers, and assembly lines at car factories. Garrett Haake interviewed Ford dealer owner Jim Sevet and autoworker James Benson in Dearborn, Michigan. …
Ref Answer: No
Qwen2.5-VL-3B:**Yes** Qwen2.5-VL-7B :**Yes** Qwen2.5-VL-72B :**Yes** InternVL3-14B:**Yes** InternVL3-78B:No

Out-Video:
05:16-07:49 - …Visuals included President Trump signing documents, **expansive fields of sunflowers**, and assembly lines at car factories. Garrett Haake interviewed Ford dealer owner Jim Sevet and autoworker James Benson in Dearborn, Michigan. …
Ref Answer: No
Qwen2.5-VL-3B:No Qwen2.5-VL-7B :**Yes** Qwen2.5-VL-72B :No InternVL3-14B:No InternVL3-78B:No

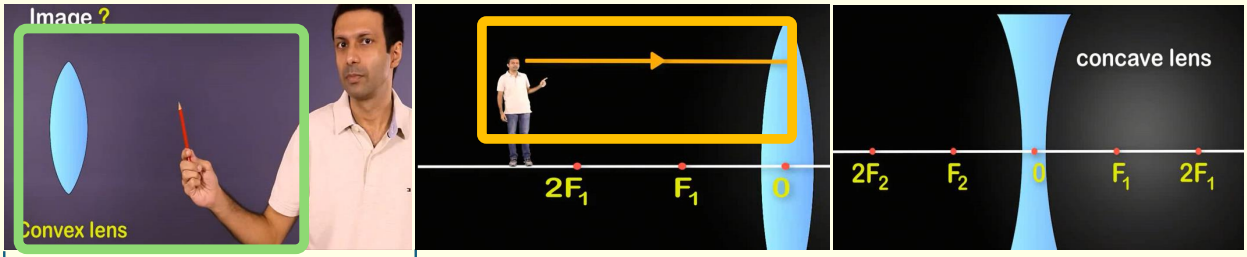
Figure 20. SAH case in ELV-Halluc.

Ground Truth: Is the following caption totally correct? Reply with "Yes" or "No" only.
02:37-02:59 - River Plate's No. 29 player Marcos Acuña, wearing a white jersey, takes a corner kick from the left side of the field. The ball curves into the crowded penalty area, and River Plate's No. 8 player **Maxi Meza** leaps and powerfully heads the ball into the top left corner of the goal, over the yellow goalkeeper's reach. ...
Ref Answer: Yes
Qwen2.5-VL-3B:Yes Qwen2.5-VL-7B :Yes Qwen2.5-VL-72B :Yes InternVL3-14B:Yes InternVL3-78B:Yes

In-Video:
02:37-02:59 - River Plate's No. 29 player Marcos Acuña, wearing a white jersey, takes a corner kick from the left side of the field. The ball curves into the crowded penalty area, and River Plate's No. 8 player **Driussi** ...
Ref Answer: No
Qwen2.5-VL-3B:**Yes** Qwen2.5-VL-7B :**Yes** Qwen2.5-VL-72B :**Yes** InternVL3-14B:No InternVL3-78B:**Yes**

Out-Video:
02:37-02:59 - River Plate's No. 29 player Marcos Acuña, wearing a white jersey, takes a corner kick from the left side of the field. The ball curves into the crowded penalty area, and River Plate's No. 8 player **Gabriel Batistuta**...
Ref Answer: No
Qwen2.5-VL-3B:**Yes** Qwen2.5-VL-7B :**Yes** Qwen2.5-VL-72B :No InternVL3-14B:**Yes** InternVL3-78B:No

Figure 21. SAH case in ELV-Halluc.



00:00 02:19

Ground Truth: Is the following caption totally correct? Reply with "Yes" or "No" only.

00:00-02:19 - A man introduces the imaging of spherical lenses, holding a pencil to **show convex and concave lenses.** He explains that he will be the object and demonstrates how to accurately draw the principal axis, optical center (O), and focal points (F1, 2F1, F2, 2F2) of a convex lens using a ruler, emphasizing drawing thin lenses.

Ref Answer: Yes

Qwen2.5-VL-3B:Yes Qwen2.5-VL-7B :**No** Qwen2.5-VL-72B :Yes InternVL3-14B:Yes InternVL3-78B:**No**

In-Video:

00:00-02:19 - A man introduces the imaging of spherical lenses, holding a pencil to **trace rays from the top of his head onto convex and concave lenses**. He explains ...

Ref Answer: No

Qwen2.5-VL-3B:**Yes** Qwen2.5-VL-7B :No Qwen2.5-VL-72B :**Yes** InternVL3-14B:**Yes** InternVL3-78B:**Yes**

Out-Video:

00:00-02:19 - A man introduces the imaging of spherical lenses, holding a pencil to **clean the lenses thoroughly before drawing convex and concave lenses.** He explains ...

Ref Answer: No

Qwen2.5-VL-3B:No Qwen2.5-VL-7B :**Yes** Qwen2.5-VL-72B :No InternVL3-14B:No InternVL3-78B:No

Figure 22. SAH case in ELV-Halluc.

Ground Truth: Is the following caption totally correct? Reply with "Yes" or "No" only.
01:46-02:19 - On the blue tennis court, Novak Djokovic, wearing a shirt with blue and white patterns, rallied with Stefanos Tsitsipas, who was dressed in a white shirt and colorful shorts. Tsitsipas attempted a drop shot, but Djokovic quickly rushed forward, extended himself, and hit a sharp-angled backhand winner that landed just inside the line, preventing Tsitsipas from converting a set point and bringing the match to deuce.
Ref Answer: Yes
Qwen2.5-VL-3B:Yes Qwen2.5-VL-7B :Yes Qwen2.5-VL-72B :No InternVL3-14B:Yes InternVL3-78B:Yes

In-Video:
01:46-02:19 - On the blue tennis court, Novak Djokovic, wearing a shirt with blue and white patterns, rallied with Roberto Carballes Baena, who was dressed in a white shirt and colorful shorts. Carballes Baena attempted a drop shot, but…Ref Answer: No
Qwen2.5-VL-3B:Yes Qwen2.5-VL-7B :Yes Qwen2.5-VL-72B :Yes InternVL3-14B:Yes InternVL3-78B:Yes

Out-Video:
01:46-02:19 - On the blue tennis court, Novak Djokovic, wearing a shirt with blue and white patterns, rallied with Rafael Nadal, who was dressed in a white shirt and colorful shorts. Nadal attempted a drop shot, but …
Ref Answer: No
Qwen2.5-VL-3B:No Qwen2.5-VL-7B :Yes Qwen2.5-VL-72B :No InternVL3-14B:No InternVL3-78B:No

Figure 23. SAH case in ELV-Halluc.

Prompt for gemini-2.5-flash to generate scratch captions.

You are an expert in understanding videos.
Task: You will be provided with a {} video. Your task consists of three main steps: 1. Comprehensively understand the video using both audio and visual information. 2. Segment the video into several parts, while each part contains a isolate and complete **{}**. 3. Generate detailed captions for each part according to the following guidelines.
Note: These parts are parallel, which is **a list of distinct but equally important elements**. (e.g., several complete news articles, multiple goals, different cooking steps, different movie or TV shows, different parts in vlog, or isolated topics)
Guidelines for Video Caption Generation: Each caption **must include the following aspects**:
1. **Time Range** - Provide the precise time interval of the segment (e.g., "02:15–02:42").
2. **Visual Details** – Describe specific visual elements in the segment with fine granularity: - Objects: Include object shape, color, texture, written text (OCR), logos, jersey numbers, etc. - Scene Layout: Describe background elements. - Spatial Relationships: Explain relative positions.
3. **Subject** – Clearly identify the main subject(s) in the segment.
4. **Action** – Describe physical activities with verbs.
5. **Event** – Describe what happens as a complete narrative.
Output Instructions: 1. Captions should be coherent and complete sentences. 2. Ignore redundant and uninformative content such as static frames, repetitive scenes, intros/outros, idle moments, and summary frames at the beginning or end of the video. 3. Ignore transitional or filler content between each major segment. 4. Each video must be conclude into less than 10 parts.
Output Format: You must strictly follow the output format below and output nothing beyond the specified structure. "'json {{ "part_1": "mm:ss-mm:ss - ...", "part_2": "mm:ss-mm:ss - ...", ... }} "'

Figure 24. Prompt for gemini-2.5-flash to generate scratch captions.

Prompt for GPT-4o to inject visual detail hallucinations.

You are a professional caption generation assistant. You will be provided with captions for different parts of the same video. Your task is to generate hallucinated captions following the instructions below.
Current Part Caption (Original Caption): {current_part}
Other Parts Captions: {other_parts}
Instructions: 1. Break down the current part caption into the following components: - **Visual details**: attributes such as color, shape, size, patterns, spatial relationships, or on-screen text (OCR). - **Objects**: refers to humans or physical objects involved in the caption. - **Actions**: refers to the key activity or motion being performed. - **Declarative Content**: refers to high-level descriptive or propositional statements that summarize a situation, assert an outcome, or convey a belief, stance, or result. These are not concrete actions or events. - **Time range**: the timestamp segment covered by this caption.
2. Identify a **visual detail** in the current part caption.
3. Replace that visual detail using: - **in_video hallucination**: Replace it with a visual detail that appears in *other parts captions* of the same video. - **out_video hallucination**: Replace it with a visual detail that *never appears in any part captions* of this video.
4. The hallucinated captions must remain **plausible and reasonable**, so that a model cannot reliably judge which is correct (among original, in_video, and out_video) without actually seeing the video.
IMPORTANT: - You must strictly preserve all the rest of the current part caption (including its time range and sentence structure). Only the selected **visual detail** is to be replaced. - The replacement must be logically consistent and naturally integrated, with no contradictions or cues that obviously reveal the modification. - The replacement must change the semantics of the original caption, and cannot use synonyms.
Output Format: "'json {{ "in_video": "in_video caption here", "out_video": "out_video caption here" }} "'

Figure 25. Prompt for GPT-4o to inject visual detail hallucinations.

Prompt for GPT-4o to inject object hallucinations.

You are a professional caption generation assistant. You will be provided with captions for different parts of the same video. Your task is to generate hallucinated captions following the instructions below.
Current Part Caption (Original Caption): {current_part}
Other Parts Captions: {other_parts}
Instructions: 1. Break down the current part caption into the following components: - **Visual details**: attributes such as color, shape, size, patterns, spatial relationships, or on-screen text (OCR). - **Objects**: refers to humans or physical objects involved in the caption. - **Actions**: refers to the key activity or motion being performed. - **Declarative Content**: refers to high-level descriptive or propositional statements that summarize a situation, assert an outcome, or convey a belief, stance, or result. These are not concrete actions or events. - **Time range**: the timestamp segment covered by this caption.
2. Identify a **object** in the current part caption.
3. Replace that object using: - **in_video hallucination**: Replace it with an object that appears in *other parts captions* of the same video. - **out_video hallucination**: Replace it with an object that *never appears in any part captions* of this video.
4. The hallucinated captions must remain **plausible and reasonable**, so that a model cannot reliably judge which is correct (among original, in_video, and out_video) without actually seeing the video.
IMPORTANT: - You must strictly preserve all the rest of the current part caption (including its time range and sentence structure). Only the selected **object** is to be replaced. - The replacement must be logically consistent and naturally integrated, with no contradictions or cues that obviously reveal the modification. - The replacement must change the semantics of the original caption, and cannot use synonyms.
Output Format: "'json {{ "in_video": "in_video caption here", "out_video": "out_video caption here" }} "'

Figure 26. Prompt for GPT-4o to inject object hallucinations.

Prompt for GPT-4o to inject action hallucinations.

You are a professional caption generation assistant. You will be provided with captions for different parts of the same video. Your task is to generate hallucinated captions following the instructions below.
Current Part Caption (Original Caption): {current_part}
Other Parts Captions: {other_parts}
Instructions: 1. Break down the current part caption into the following components: - **Visual details**: attributes such as color, shape, size, patterns, spatial relationships, or on-screen text (OCR). - **Objects**: refers to humans or physical objects involved in the caption. - **Actions**: refers to the key activity or motion being performed. - **Declarative Content**: refers to high-level descriptive or propositional statements that summarize a situation, assert an outcome, or convey a belief, stance, or result. These are not concrete actions or events. - **Time range**: the timestamp segment covered by this caption.
2. Identify an **action** in the current part caption.
3. Replace that action using: - **in_video hallucination**: Replace it with an action that appears in *other parts captions* of the same video. - **out_video hallucination**: Replace it with an action that *never appears in any part captions* of this video.
4. The hallucinated captions must remain **plausible and reasonable**, so that a model cannot reliably judge which is correct (among original, in_video, and out_video) without actually seeing the video.
IMPORTANT: - You must strictly preserve all the rest of the current part caption (including its time range and sentence structure). Only the selected **action** is to be replaced. - The replacement must be logically consistent and naturally integrated, with no contradictions or cues that obviously reveal the modification. - The replacement must change the semantics of the original caption, and cannot use synonyms.
Output Format: "'json {{ "in_video": "in_video caption here", "out_video": "out_video caption here" }} "'

Figure 27. Prompt for GPT-4o to inject action hallucinations.

Prompt for GPT-4o to inject declarative content hallucinations.

You are a professional caption generation assistant. You will be provided with captions for different parts of the same video. Your task is to generate hallucinated captions following the instructions below.
Current Part Caption (Original Caption): {current_part}
Other Parts Captions: {other_parts}
Instructions: 1. Break down the current part caption into the following components: - **Visual details**: attributes such as color, shape, size, patterns, spatial relationships, or on-screen text (OCR). - **Objects**: refers to humans or physical objects involved in the caption. - **Actions**: refers to the key activity or motion being performed. - **Declarative Content**: refers to high-level descriptive or propositional statements that summarize a situation, assert an outcome, or convey a belief, stance, or result. These are not concrete actions or events. - **Time range**: the timestamp segment covered by this caption.
2. Identify a **declarative content** in the current part caption.
3. Replace that declarative content using: - **in_video hallucination**: Replace it with a declarative content that appears in *other parts captions* of the same video. - **out_video hallucination**: Replace it with a declarative content that *never appears in any part captions* of this video.
4. The hallucinated captions must remain **plausible and reasonable**, so that a model cannot reliably judge which is correct (among original, in_video, and out_video) without actually seeing the video.
IMPORTANT: - You must strictly preserve all the rest of the current part caption (including its time range and sentence structure). Only the selected **declarative content** is to be replaced. - The replacement must be logically consistent and naturally integrated, with no contradictions or cues that obviously reveal the modification. - The replacement must change the semantics of the original caption, and cannot use synonyms.
Output Format: "'json {{ "in_video": "in_video caption here", "out_video": "out_video caption here" }} "'

Figure 28. Prompt for GPT-4o to inject declarative content hallucinations.

Prompt for GPT-4o to check hallucinated captions.

You are a strict video caption checker.
You will be given multiple captions from different parts in a same video. Each part has a ground-truth caption and two hallucinated captions: ‘in_video‘ (containing a modification that should come from another part of the same video) and ‘out_video‘ (containing a plausible but fabricated modification that does not appear in any part of the video). The hallucination aspect of this part is **aspect**.
First, identify what information has been added or changed in the hallucinated captions compared to the ground_truth. Second, verify whether: 1. The **in_video** caption introduces a/an **aspect** modification that **semanticly appears** in at least one **other part's** ‘ground_truth‘ (excluding the current part). 2. The **out_video** caption introduces a/an **aspect** modification that **semanticly does NOT appear** in any of the ground_truth captions. Focus on the differences between the hallucinated captions and the original ground_truth. Please return your judgment strictly follows the below JSON format, don't provide any reasons, explanations, verifications: “‘json {{ ”in_video_valid”: true/false, ”out_video_valid”: true/false }} “‘ Ground Truth: {ground_truth}
In-Video Hallucination: {in_video}
Out-Video Hallucination: {out_video}
Other Parts' Ground Truths: {other_parts_gt}

Figure 29. Prompt for GPT-4o to check hallucinated captions.