

ARGUS: Hallucination and Omission Evaluation Framework for Video-LLMs

Appendix

A. The Limitations of Current QA Approaches - Extended

Current evaluation strategies predominantly rely on a question-answering (QA) paradigm, where the model is tested with two types of questions: one basic type that expects a “Yes” (confirming the presence of information in the video) and another “leading/hallucinated” type that expects a “No” (flagging potential hallucinations). Although this setup simplifies evaluation, it suffers from several critical limitations.

Lack of Dependency on Over Visual Understanding A key limitation of binary QA-based evaluation is that models can sometimes answer questions correctly using general world knowledge alone, without processing the visual input [34, 68]. To demonstrate that this issue persists in current video hallucination datasets, we conduct an experiment where an LLM is tested on QA metrics without access to visual context. High performance in this setting indicates that these evaluations do not reliably measure visual grounding. Specifically, we use GPT-4o to answer a randomly selected subset of questions from Video-HalluCer [58] using only the textual question. Surprisingly, the model correctly answers 32.52% of basic hallucination-related question pairs, with some subcategories (`external_nonfactual_instruct`) reaching 61.33%, despite a chance performance of just 25%.

Verification Ability Does Not Equate to Strong Generation Video LLMs are used for a range of tasks, including question-answering (QA) and open-ended generation, such as captioning and summarization. While QA evaluates a model’s ability to verify information, it does not necessarily reflect its capability to generate accurate and coherent descriptions. Recent work on hallucination evaluation in image captioning [28] suggests that proficiency in fact verification does not always correlate with strong open-ended generation and may even be inversely related. We observe a similar pattern in the qualitative example shown in Figure 3, where a Video-LLM correctly verifies that there is only one chameleon in a video clip but mistakenly identifies it as two distinct chameleons in free-form generation. This highlights the need for a dedicated benchmark to assess hallucinations specifically in open-ended generation. To further quantify this behavior, we conducted an experiment using 50 randomly selected videos from our dataset. For each video, we extracted each line (one at a time) from a model’s generated caption and reformulated it into a verification prompt, which was then presented to the same model along with the visual video input:

Evaluate the accuracy of this statement about the video content:

Statement: {`model-caption-line`}

(a) True

(b) False

Respond with only the letter of your evaluation: (a) or (b).

Using this setup, we measured the *verification accuracy* of the model. Since we already had ground-truth entailment labels from our captioning evaluation—indicating whether a given line was factual or hallucinated—we could compare them against the model’s verification responses. This allowed us to analyze the relationship between a model’s verification accuracy and its hallucination rate. Because both tasks were performed by the same model, a strong negative correlation between verification accuracy and hallucination cost would suggest that these abilities are intertwined. In contrast, a weak correlation would imply that a model may accurately verify information while still hallucinating in free-form generation, or vice versa. In our analysis, we observed a Pearson correlation of -0.48 (see Fig. 9), indicating a moderate negative correlation. These findings further highlight the need for a dedicated benchmark to assess hallucinations specifically in open-ended generation tasks.

Restricted Error Coverage Due to Predefined Scope Another limitation of the QA-based approach is its restricted scope; the predefined questions cover only a narrow range of possible errors. As a result, the full spectrum of hallucinations remains hidden until the model is allowed to generate free-form captions.

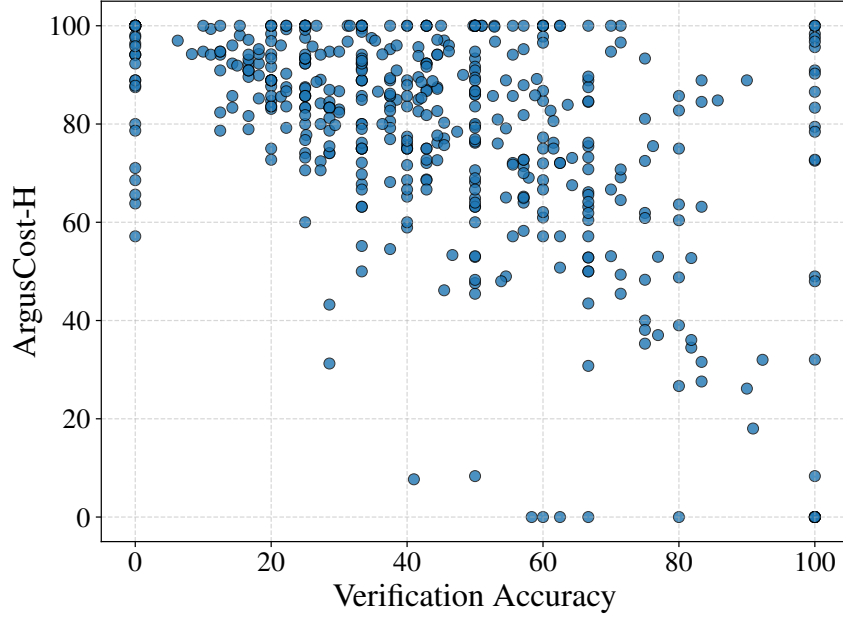


Figure 9. Relationship between verification accuracy and hallucination cost across 50 video samples.

Inability to Capture Multi-Event Hallucinations Many videos consist of multiple interrelated events, yet QA-based evaluation methods typically focus on isolated, short events, if at all. This narrow focus fails to account for the complex interactions and temporal dependencies between events, which can contribute to more severe hallucinations. For example, if a video contains four sequential events (event-a, event-b, event-c, and event-d), a binary QA approach would require separately probing the existence and order of each event. Moreover, with a 50% chance of guessing correctly in a binary setting, it raises questions about whether the model genuinely understands the content. In contrast, our free-form generation strategy allows the model to produce captions based on its interpretation of the video, enabling a more direct evaluation of its accuracy.

B. Extended Related Work

Natural Language Inference Natural Language Inference (NLI) is a classic natural language processing task aimed at determining whether a given hypothesis can logically be inferred from a provided premise [4, 6, 9, 13, 59, 60]. Formally, NLI involves classifying the relationship between a hypothesis and a premise into one of three categories: entailment (hypothesis logically follows from the premise), contradiction (hypothesis contradiction implies the hypothesis), or neutrality (no direct logical relation between the two). As one of the earliest works, Bowman et al. [4] released the SNLI corpus, a large-scale annotated datasets for NLI. SNLI consists of over 500k sentence pairs, each labeled to indicate entailment, contradiction, or semantic independence, enabling training of modern machine learning models. Williams et al. [59] introduced MNLI, a large scale NLI dataset extending the traditional natural language inference task to multiple-genres of spoken and written english such as reports, letters, fiction works, etc. Demszky et al. [13] proposed an automated way of generating NLI datasets based on existing question-answering datasets. More recently, NLI techniques have been applied to evaluating factuality in domains such as natural language summarization by analyzing the relationship between a generated summary and the source document [19, 50].

In this work, we adapt the broader NLI-framework to assess hallucination and omission in free-form generation by Video-LLMs. Specifically, we frame this evaluation as a textual NLI problem, treating the ground truth human caption as the premise (and hypothesis in omission evaluation) and the generated model caption as the hypothesis (and the premise in omission evaluation).

Dense Video Captioning. Video Captioning is one of the traditional tasks since the dawn of deep learning. Dense video captioning is a fine-detail oriented subtask in which the model has to annotate all the important visual and temporal details in the video. Some of the common applications include surveillance monitoring, video content retrieval, and educational

accessibility. Additionally, dense video captioning can generate synthetic data for training video-generative models since dense captioning by humans at scale can be quite expensive. Most of the early works [14, 25, 26, 29, 32, 40, 45, 56, 61, 67] handled this via combination of multiple video encoders and LSTM-based or transformer language decoders and have shown good results on highly specific academic datasets. These pioneering approaches laid the groundwork for the task by demonstrating that integrating various temporal and spatial features was key to producing accurate and meaningful captions. However, the reliance on narrowly focused datasets often limited the generalizability of these systems in more diverse, real-world scenarios. Recent breakthroughs have led to the emergence of generalist dense video captioning capabilities, driven by large-scale multimodal models such as Gemini [52] and GPT-4v [2] and others. These modern approaches rely on end-to-end training or fine-tuning expansive transformer-based decoder models, augmented with specialized video encoders like SigLIP [64], CLIP [46], Video-MAE [54], and Eva-CLIP [16] and on top of internet-scale multi-modal interleaved datasets.

C. Additional Evaluation Details

We use two NVIDIA A40 GPUs, each with 48GB of memory, and two NVIDIA A100 GPUs, each with 82GB of memory, for experiments with open-source models. All models and their checkpoints are listed in Table 3. The open-source models are available via [HuggingFace](#), while the proprietary models are accessible through their respective providers’ APIs. All evaluations for the proprietary models were conducted in February 2025.

D. Additional Sensitivity Analysis

Sensitivity to Intra-Prompt Variation By intra-prompt variation, we refer to the inherent stochasticity in Video-LLM responses due to decoding at a default temperature. To assess whether this variability significantly impacts our results, we sample multiple captions for each model using our default prompt (“Describe the video in great detail.”) and analyze the variation in hallucination costs. In Figure 10, we plot the average hallucination cost across three runs along with the

Model Name	Checkpoint
Gemini-1.5-Flash	gemini-1.5-flash
Gemini-2.0-Pro	gemini-2.0-pro-exp-02-05
Gemini-2.0-Flash	gemini-2.0-flash
GPT-4o	gpt-4o
GPT-4o-Mini	gpt-4o-mini
InternVL2 (1B)	OpenGVLab/InternVL2-1B
InternVL2 (2B)	OpenGVLab/InternVL2-2B
InternVL2 (4B)	OpenGVLab/InternVL2-4B
InternVL2 (8B)	OpenGVLab/InternVL2-8B
LLaVA-OV (7B)	lmms-lab/llava-onevision-qwen2-7b-ov
LLaVA-OV (500M)	lmms-lab/llava-onevision-qwen2-0.5b-ov
LLaVA-Video (7B)	lmms-lab/LLaVA-Video-7B-Qwen2
LLaVA-Video-Only (7B)	lmms-lab/LLaVA-Video-7B-Qwen2-Video-Only
LLaVA-NV-DPO (7B)	lmms-lab/LLaVA-NeXT-Video-7B-DPO
LLaVA-NV (7B)	lmms-lab/LLaVA-NeXT-Video-7B
MiniCPM-V-2.6	openbmb/MiniCPM-V-2.6
mPLUG-Owl3 (1B)	mPLUG/mPLUG-Owl3-1B-241014
mPLUG-Owl3 (7B)	mPLUG/mPLUG-Owl3-7B-241101
Qwen2.5-VL (7B)	Qwen/Qwen2.5-VL-7B-Instruct
Qwen2.5-VL (3B)	Qwen/Qwen2.5-VL-3B-Instruct
SmolVLM2 (256M)	HuggingFaceTB/SmolVLM2-256M-Video-Instruct
SmolVLM2 (500M)	HuggingFaceTB/SmolVLM2-500M-Video-Instruct
SmolVLM2 (2B)	HuggingFaceTB/SmolVLM2-2.2B-Instruct

Table 3. Details on model names and corresponding checkpoints. All open-source models are available via [HuggingFace](#), and proprietary models are available via respective providers.

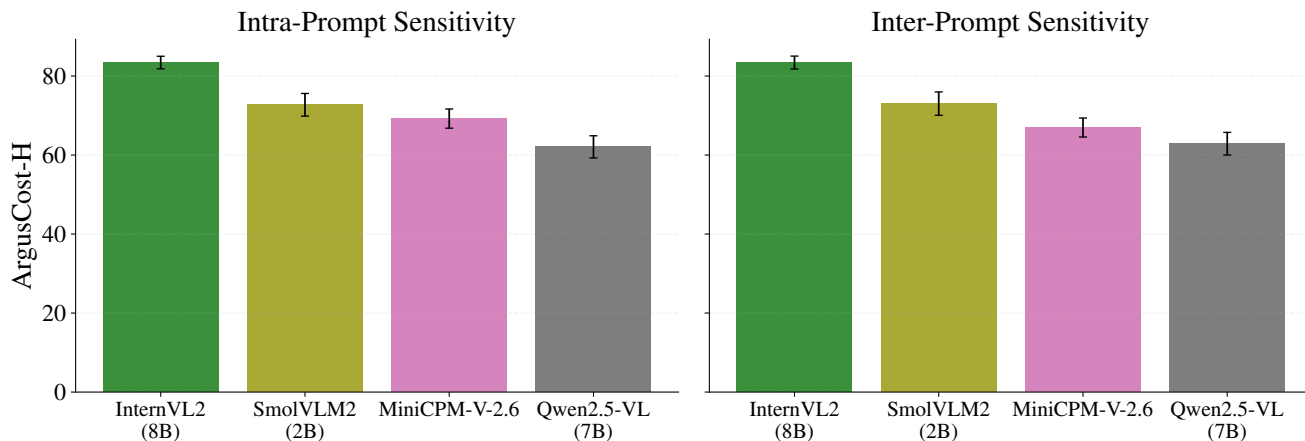


Figure 10. **Sensitivity to intra- and inter-prompt variations.** Low standard error both intra-prompt (same prompt, default temperature) and inter-prompt (across prompts) suggests consistent hallucination costs across models.

standard error. We find that the standard error is minimal, indicating little variation in results. For instance, for InternVL2, we observe a standard error of 1.63, and for Qwen, the standard error is 2.86. This suggests that despite stochasticity in decoding, intra-prompt variation does not substantially affect our conclusions.

Sensitivity to Inter-Prompt Variation Another source of variation is the specific prompt used to generate captions. By default, we prompt all models with “Describe the video in great detail.” while excluding any special tags required by certain models (e.g., `<|user|>`, etc.). To assess the impact of prompt variation, we run experiments using three additional prompts: “Explain the content of this clip thoroughly.”, “Can you summarize all the key elements and events of this video?”, and “Walk me through this video scene by scene.” Figure 10 reports the hallucination cost averaged over these runs, along with the standard error. As with intra-prompt variation, we find minimal variation across different prompts, indicating that the trends remain stable.

Sensitivity to the LLM-as-judge Another potential source of variation is the choice of LLM used as the judge for generating NLI judgments, as the entire evaluation process depends on it. By default, we use GPT-4o. However, since GPT-4o is also one of the models being evaluated, there is a possibility of self-bias. To investigate this, we conducted experiments using four additional LLM-as-judge models from different families and sizes, all of which are strong in NLI evaluation: DeepSeek-R1 [23], DeepSeek-V3 [35], LLaMa-3.3 [20], and Qwen-2.5 [62]. Figure 8 presents the Pearson ranking correlations r between rankings produced by different judge models. We find that these correlations are very high ($r \geq 0.92$), indicating strong agreement across judge models. Notably, GPT-4o (our default judge) has a ranking correlation of $r = 0.96$ with DeepSeek-R1, $r = 0.97$ with DeepSeek-V3, $r = 0.93$ with LLaMa-3.3, and $r = 0.92$ with Qwen-2.5, suggesting that our evaluation remains robust when a frontier judge model is used.

Sensitivity to Decoding-Temperature In Figure 11, we visualize the effect of increasing decoding temperature on hallucination and omission costs. We do not see a clear trend across models here, as the hallucination costs oscillates for MiniCPM-V-2.6 and (to a slightly smaller degree) for Qwen2.5-VL. Whereas, for SmolVLM2, we see a consistent decrease in hallucination and omission cost as the temperature increases.

E. Human Study Details

While we collect ground-truth human captions, several components of our evaluation pipeline are automated. For example, the NLI-based evaluation uses a large language model (LLM) that compares human and model-generated captions. To ensure reliability, we follow best practices for LLM-as-judge approaches and provide in-context examples. However, differences in detail can still lead to occasional mismatches. A model might include a minor element omitted by the human—either deemed unimportant or overlooked. If that detail isn’t commonly inferred knowledge, the LLM may flag it as a hallucination. For instance, “he is running” versus “he is running fast” is generally acceptable by the evaluator-LLM, as

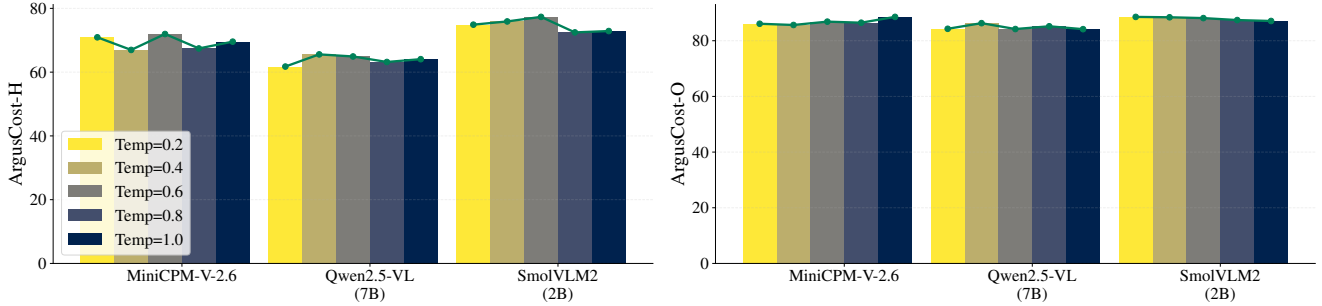


Figure 11. **Sensitivity across decoding temperatures.** Hallucination costs show little variation across temperatures, indicating low sensitivity to decoding changes

speed is implied. In contrast, mentioning a background flower not noted by the human may be marked as a hallucination. While human captions emphasize salient temporal events, occasional false positives in hallucination detection may still occur.

To assess the reliability of the LLM-based evaluation alongside human annotations, we conducted a human study. Participants were shown the original video along with the NLI evaluator LLM’s line-by-line outputs, including whether each segment of the model caption was labeled a “hallucination.” They were then asked to indicate whether they agreed with each verdict. Agreement suggests alignment between the LLM’s judgment and human perception, reinforcing the evaluation’s reliability. Disagreement may indicate either limitations in the human caption (e.g., under-specification) or errors in the LLM’s reasoning, potentially affecting the final evaluation score. Please see Fig. 12 for details of the registration webpage and the task webpage from an example video in the survey.

This study was conducted with 26 graduate student volunteers. The human study protocol was reviewed and granted exemption by our institution’s Institutional Review Board (IRB). All participants provided informed consent prior to viewing the videos and completing the survey. No personally identifiable information was collected. Each human participant reviewed a minimum of three video clips, with the option to evaluate more; on average, participants assessed 3.54 clips. We measured the average agreement rate *per clip*—defined as the proportion of lines within a clip for which human reviewers agreed with the LLM evaluator’s verdict—and found it to be 91.26%. A detailed breakdown of disagreement sources is presented in Fig. 13. The highest rate of disagreement occurred for sentences categorized as “Visual Description” that were labeled as “Hallucination” by the LLM evaluator but judged as “Entailment” by human reviewers. This pattern is expected, as video-based LLMs could mention fine-grained visual details from both the foreground and background, many of which could be difficult to capture exhaustively in the initial human annotation. Nevertheless, we believe that this is not a major concern, as the overall disagreement rate remains low, and even the strongest models exhibit substantially higher hallucination rates—sometimes reaching up to 40% disagreement. Moreover, the fact that most disagreements are skewed toward false positives for hallucination is, in fact, advantageous. It indicates that the evaluation is conservative: lines labeled as “Entailment” are highly likely to be genuinely grounded in the video content. While this may lead to slight over-penalization for accurate but omitted visual details, such conservatism is well-suited to high-stakes evaluation settings, where ensuring the reliability of entailment judgments is crucial.

F. Additional Benchmark Details

Additional Benchmark Curation Details. Our dataset’s diversity also stems from its varied video sources. As discussed earlier, we construct the dataset through manual curation and annotation, supplemented by existing video-understanding and text-to-video generation benchmarks such as VDC [7] and TC-Bench [17], which themselves integrate diverse video collections. Specifically, 12% of our dataset comes from Ego4D [21], which focuses on egocentric perspective videos, while 25% is sourced from Panda-70M [8], a dataset of high-resolution YouTube clips spanning various domains, including TV shows, cooking, gaming, and sports. The remaining 63% is drawn from free stock platforms like Mixkit ², Pexels ³, and Pixabay ⁴, providing high-resolution footage of scenic views and human activities. Additionally, about 4% is curated from

²<https://mixkit.co/>

³<https://www.pexels.com/>

⁴<https://pixabay.com/>

Registration Webpage

Caption Evaluation Study – Consent Form

Purpose of the Study

You are invited to participate in a research study that aims to understand and improve how large language models (LLMs) evaluate video captions. Your feedback will help researchers refine automatic caption-quality assessments.

What You Will Do (Procedures)

In this study, you will review verdicts generated by an Evaluator LLM for individual caption lines of short video clips. Specifically, you will:

- View content:** Watch a short video clip and read its corresponding caption.
- Review verdicts:** For each caption line, you will see the Evaluator LLM’s verdict—either *Entailment* (supported by the video) or *Hallucinated* (not fully supported/contradicted).
- Provide your judgment:** Indicate whether you agree or disagree with each verdict, based on the video and caption provided.

This process is expected to take ~10 minutes.

Risks & Benefits

Risks: The study involves minimal risk. The primary risk is a potential breach of confidentiality. To mitigate this, all data are stored on a secure, password-protected server accessible only to the research team. No personal identifiers are collected.

Benefits: While you may not receive direct personal benefits, your participation will advance research on automated video caption evaluation, which could ultimately improve accessibility tools.

Confidentiality

No personally identifiable information will be collected. You will be assigned a random participant ID for this session only. All data will be stored on encrypted, password-protected servers and reported in aggregate form so individual responses cannot be traced back to you.

Your Rights

Participation is entirely voluntary. You may decline to answer any question or withdraw from the study at any time without penalty or loss of benefits to which you are otherwise entitled.

☐ **Statement of Consent:** I have read and understood the information above (Purpose, Procedures, Risks/Benefits, Confidentiality, and Participant Rights). By checking this box, I confirm that I am 18 years of age or older and voluntarily agree to participate in this research study.


Consent to Start Evaluation

Evaluation Webpage

Evaluating Video Captions

Clip: floating_cup

Video Clip:



Model Caption:

The video showcases a glass filled with a dark beverage, likely a dark liquid, being poured from a white pitcher into it. The glass is placed on a wooden surface, and the background features a gray wall. The scene transitions to a close-up of the glass, highlighting the texture and color of the drink. The video then shifts to a different setting, where a white mug is placed on a wooden surface, and a glass of dark liquid is poured from the mug into the glass. The mug is positioned on the left side of the frame, and the glass is on the right side. The video concludes with a close-up of the mug, emphasizing its texture and the liquid inside.

Caption Evaluation

Please watch the video and evaluate each line of the caption. For each line, decide whether it is:

- Entailed by Video:** The caption line is supported by what you see in the video
- Hallucinated Caption:** The caption line contains information not supported by the video

You must provide feedback for all caption lines before submitting.

[Click to expand task rules](#)

Line 1: The video showcases a glass filled with a dark beverage, likely a dark liquid, being poured from a white pitcher into it.

Verdict: hallucinated

Reasoning: The source video describes cream being poured into the coffee from a white creamer, not a dark liquid being poured from a white pitcher. The action described in the target is different from the source.

I Agree with this Verdict

I Disagree with this Verdict

Line 2: The glass is placed on a wooden surface, and the background features a gray wall.

Verdict: entailment

Reasoning: The source video confirms the glass is on a wooden surface and the presence of a gray wall in the background.

I Agree with this Verdict

I Disagree with this Verdict

Figure 12. Registration Webpage (Left) and the task webpage (right) from an example video in the human study survey.

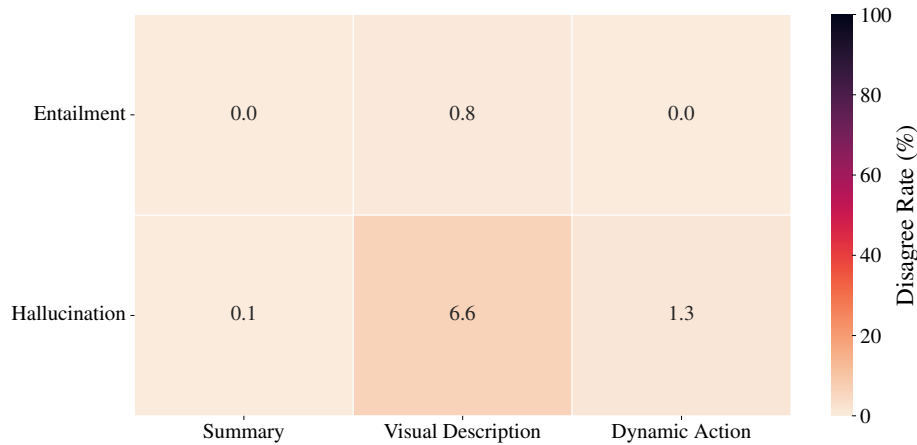


Figure 13. Human-study disagreement rate decomposed into “LLM-verdict-type” and the “sentence-type”. Overall, disagreement stands at ~ 9%, with the majority contribution being “Visual-Description” sentences that were classified as “Hallucination” by the NLI-evaluation LLM.

miscellaneous sources, further augmenting the benchmark with a variety of temporal events. We note that for VDC [7] the first round of captioning is performed by GPT-4o, which is later manually verified and corrected by the human authors. Since, we use GPT-4o as a judge for our entailment task, there is a possibility of GPT-4o’s scores inflating because of

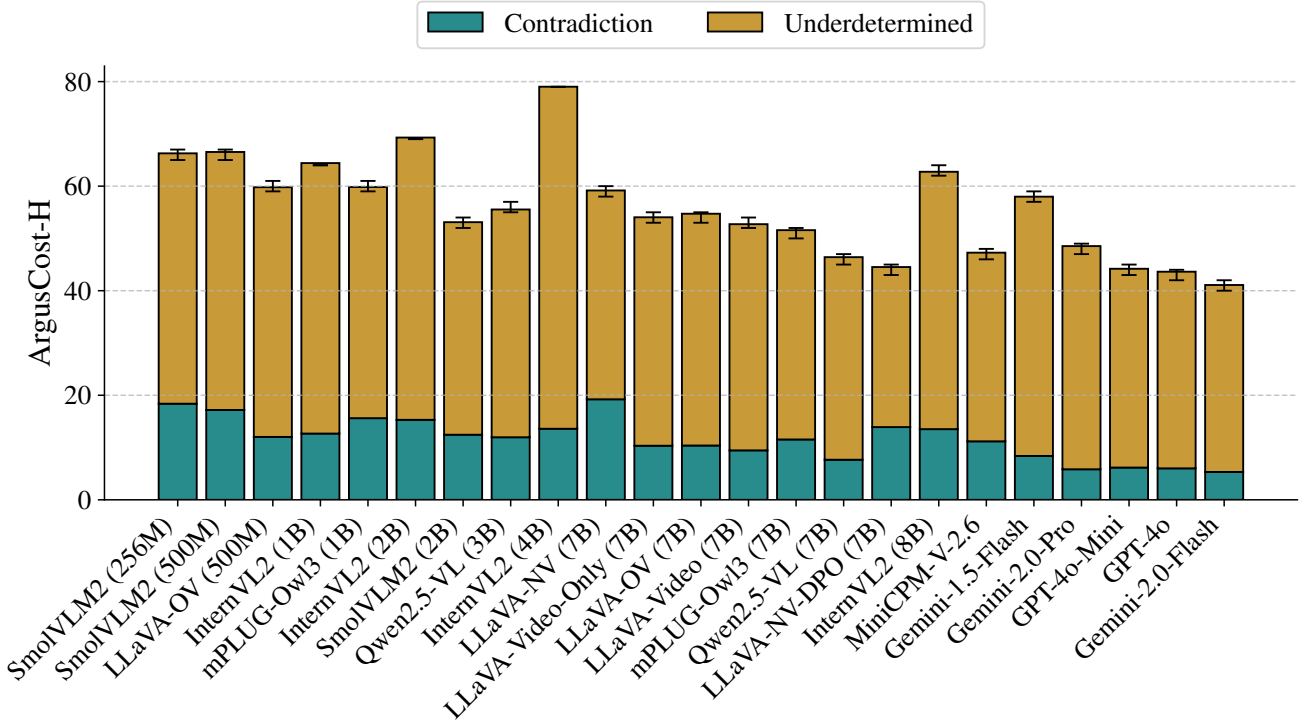


Figure 14. Breakdown of hallucination errors into “contradiction” and “undetermined”. Undetermined errors dominate the majority of errors, across model families and sizes.

self-bias. To ensure, that is not the case, we performed an ablation with many different state-of-the-art LLMs-as-judges in Section 4.3, and found the Pearson ranking correlation to be extremely high.

G. Additional Evaluation Ablations

Cost-Breakdown by Hallucination-Type In Section 4.2, we analyzed the ArgusCost-H breakdown by sentence type (summary, visual description, dynamic action). Another way to examine ArgusCost-H is by categorizing hallucination errors as either “contradiction” or “undetermined.” Figure 14 visualizes this breakdown, showing that approximately two-thirds of errors fall into the “undetermined” category, while the remaining one-third are contradictions. This trend holds across all models and sizes, indicating that models generate more fabricated content rather than directly contradicting the input. Additionally, as models improve (moving right on the x-axis), the proportion of undetermined errors increases, suggesting that improvements primarily come from reducing contradiction errors, leading to a better overall understanding of the video. This observation aligns with the correlation between ArgusCost-H and ArgusCost-O, as models with fewer contradictions also tend to omit less information.

Cost-Breakdown by Cost-Type Another way we break down ArgusCost-H is by examining the Base Costs and Temporal Penalties defined in Section 3. In Figure 15, we visualize the temporal penalties incurred by different models. Since these penalties are much smaller than the base cost, making up only about 1% to 4% of the total cost, we present the temporal penalty costs separately. We observe that for most models, temporal penalties are close to zero, with a maximum of 4% for LLaVa-Video and InternVL2 (8B). Our qualitative analysis suggests that weaker models often fabricate information or produce contradictions, meaning they fail to generate dynamic actions in the first place, resulting in minimal temporal penalties. In contrast, stronger models still make some errors but tend to preserve the correct event order for the actions they get right, leading to lower temporal penalties overall.

Effect of Clip Duration In Section 4.2, we discussed the mild positive correlation between ArgusCost-H, ArgusCost-O, and clip duration. Figure 16 provides a visualization to further analyze this effect. The figure plots ArgusCost-H and

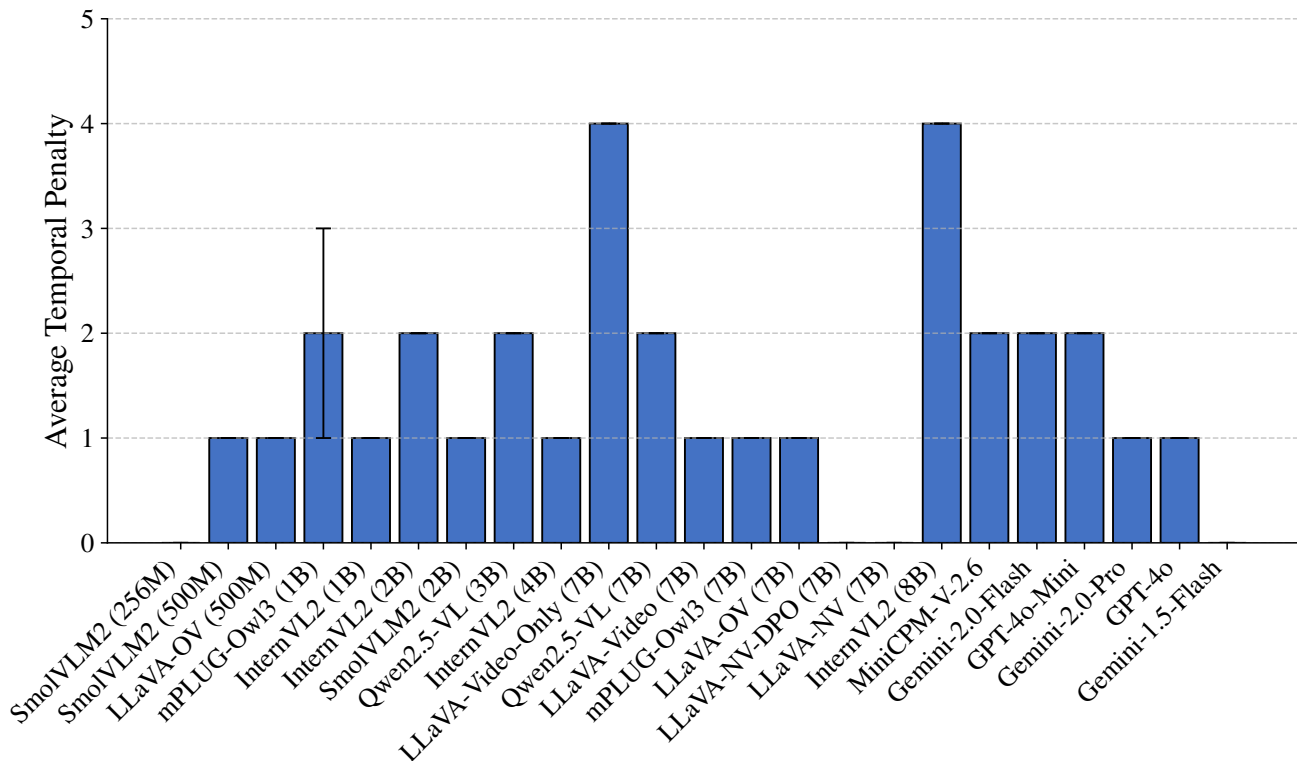


Figure 15. Contribution of Temporal Penalties in the overall cost. Temporal penalties make a minor proportion of the total costs.

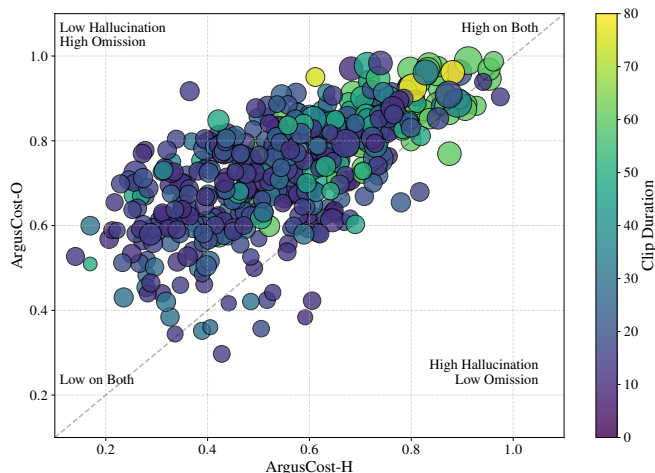


Figure 16. Visualizing the correlation between clip duration, ArgusCost-H and ArgusCost-O, and caption density

ArgusCost-O, with clip duration represented by colorbar hue and model-generated caption size indicated by marker size. The correlation is evident as brighter markers (longer videos) are clustered in high hallucination and omission areas, while darker markers are positioned elsewhere. Additionally, marker size is skewed toward the top-right, indicating that longer videos tend to generate denser captions, as expected.

Effect of Frames on Omission We discussed the effect of increasing the total number of frames provided to model from 2 to 64 on ArgusCost-H in the main paper. Here, in Figure 17 we discuss the effect on ArgusCost-O. We note that omission

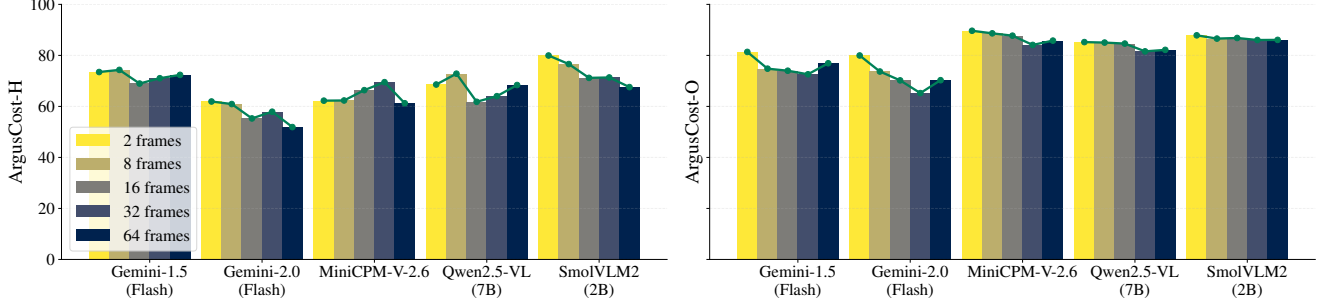


Figure 17. Effect of total number of frames on ArgusCost-H and ArgusCost-O. Gemini models and SmolVLM2 show consistent reduction in hallucinations with more frames, while MiniCPM-V and Qwen2.5-VL exhibit fluctuating levels of hallucinations (as measured by ArgusCost-H). All models show consistent improvement in ArgusCost-O results as we increase the number of frames.

rate consistently decreases as the number of frames increase for all the models. Specifically, when moving from 2 to 64 frames, Gemini-1.5-Flash improves by 4.45%, Gemini-2.0-Flash improves by 9.69%, MiniCPM-V-2.6 improves by 3.86%, Qwen2.5-VL improves by 3.04%, SmolVLM2 improves by 1.78%.

H. Relationship With Video Characteristics

In the main paper, we discussed the correlation between various video characteristics—such as video duration—with both ArgusCost-H and ArgusCost-O. In this section, we expand on that analysis by examining a broader range of video characteristics and their specific relationship with ArgusCost-H, aiming to provide a more comprehensive understanding of the influence of video content.

1. **LAION Aesthetic Score [53]:** The LAION aesthetic score is commonly used to assess the aesthetic quality of an image (ranging between 1-10). In our setting, we adapt this score by computing the average aesthetic score across all frames in a video. While this approximation has limitations, it serves as a reasonable starting point. One hypothesis is that certain models—depending on their training—may perform worse on videos with lower aesthetic quality, or conversely, may be biased towards more aesthetically pleasing content.
2. **Lighting Unique Count:** We utilize a shot detection model i.e. `diffusers/shot-categorizer-v0 [15]` to identify the lighting type in each frame of a video. We then compute the number of unique lighting types (e.g., Daylight, Sunny) and use this count as a metric. The underlying assumption is that videos with rapid or frequent lighting changes may pose challenges for models, potentially increasing the likelihood of hallucinations.
3. **Subject Presence:** The number of distinct entities present in a video may also influence model performance, with more entities potentially leading to increased confusion for VideoLLMs. To quantify this, we employ the Segment Anything Model (SAM) [47] to generate segmentation masks and count the number of unique entities per video. We then analyze the distribution of these counts across the dataset.

We visualize these results in Fig. 18. For the metrics examined, we do not observe a strong or consistent correlation with ArgusCost-H. It is important to note that this is an initial analysis and not intended to be exhaustive. In future work, we plan to explore more comprehensive features, including those that capture motion dynamics and other relevant characteristics. In Fig. 19, we visualize the correlation heatmaps separately for each model. Among the filters, we observe that the models exhibit the strongest (negative) correlation with the LAION aesthetic score.

I. Additional LLM-as-a-judge Details

I.1. Extended Discussion on Type and Verdict Categorization

Type Categorization. We observed that Video-LLMs primarily generate three types of sentences: (1) summary sentences that provide a high-level overview, often at the start of a caption; (2) visual descriptions that highlight key attributes of entities in the video; and (3) dynamic action descriptions that capture important events. Categorizing sentences by type allows us to analyze which kinds of hallucinations these models tend to produce. Additionally, dynamic actions have an inherent temporal structure, meaning the order in which they appear is crucial. To account for this, we introduce an order penalty when the model-generated sequence deviates from the source descriptions. However, summary and visual

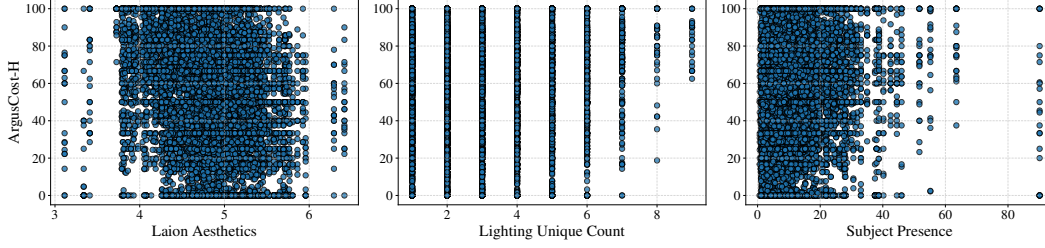


Figure 18. Correlation between various video-characteristics and ArgusCost-H.

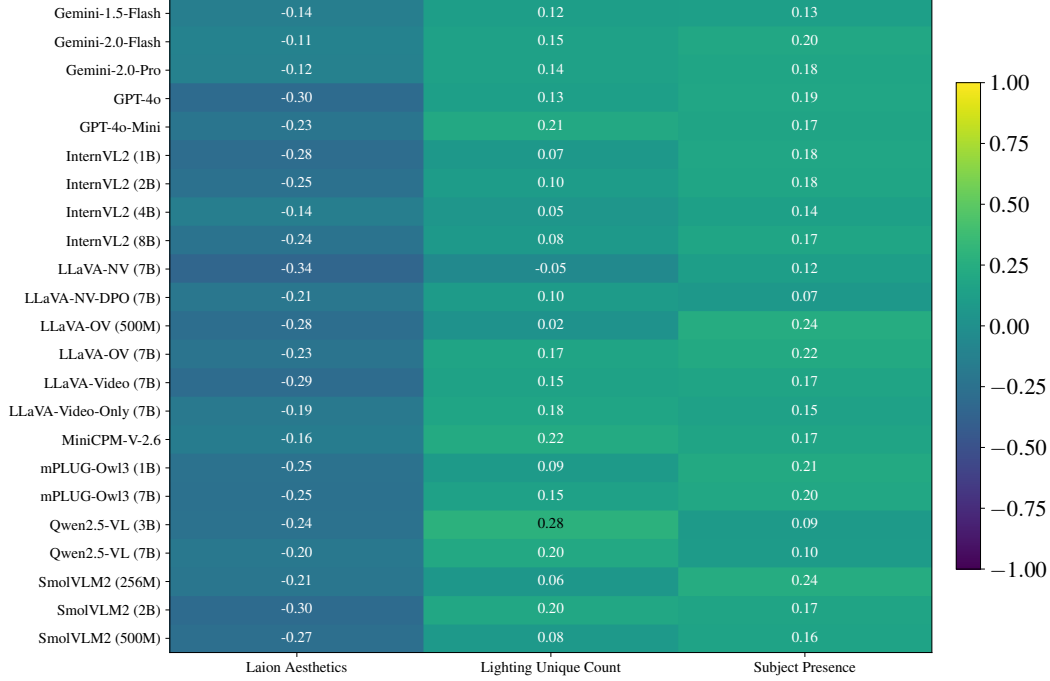


Figure 19. Model-wise correlation heatmaps between video filter metrics and ArgusCost-H.

description sentences (assuming the visual attributes remain unchanged) do not follow a strict temporal order, making their placement in the generated captions more flexible. By labeling each sentence type, we ensure that only dynamic actions are evaluated for order discrepancies. Intuitively, our temporal penalty operates on the principle that dynamic-action sentences in the VideoLLM caption should follow the same order as their evidence lines in the human caption. Thus, by aligning generated dynamic actions with their corresponding source events, we can systematically compare event sequences and quantify any inconsistencies in temporal order.

Verdict Categorization. Our classification approach is grounded in the natural language inference (NLI) framework, which defines three key relationships between a generated sentence and the source descriptions: entailment (clear supporting evidence), contradiction (clear evidence against), and undetermined (insufficient evidence to confirm or refute). Contradictions represent a specific type of hallucination where the model generates content that directly conflicts with the source. For example, if a video shows a man cooking in a kitchen but the model, due to pretraining biases, describes a woman cooking, this constitutes a contradiction. On the other hand, undetermined cases arise when the model includes details that cannot be confirmed or denied based on the available evidence. For instance, if a blurred shot in the video shows a person cooking and the human-annotated caption states “a person is cooking,” but the model-generated caption specifies “a woman is cooking,” the gender detail is undetermined. While both contradiction and undetermined cases are forms of hallucination, they differ in nature. Distinguishing between them allows us to categorize errors more precisely and measure

their effects separately, leading to a more nuanced evaluation of model reliability.

I.2. Computational Costs

Monetary Costs for NLI Evaluation using GPT-4o:

- Input Tokens: 843 (Base Prompt Instructions) + 477 (Average Human Caption Tokens) + 234 (Average Model-Generated Tokens) + 910 (In-Context Examples) = 2464
- Cost per Input Token: \$2.50 per 1M tokens
- Input Cost per Video: \$0.00616
- Average Output Tokens: 640
- Cost Per Output Tokens: \$10.0 per 1M tokens
- Output Cost per Video: 0.006
- Cost Per Video: \$0.01216

This is an upper-bound on the total cost, since the base instruction tokens can be cached. Additionally, as we see that SOTA open source also correlate highly with GPT-4o's judgment, so one could potentially use them as well.

I.3. Prompt

You are given two inputs: a `source_caption` (serving as the premise) and a `target_caption` (serving as a series of hypothesis lines). For each line in the `target_caption`, evaluate its relationship to the `source_caption` according to the following guidelines:

- ****Entailment:**** The target line is fully supported by the `source_caption`. This includes cases where the target uses alternative phrasing or synonymous descriptions, and minor variations in attributes (for example, "pink" (or "red") and "reddish pink" are treated as equivalent). Additionally, If the target line asserts that an aspect (such as lighting, background consistency, or non-occurrence of an event) remained unchanged or did not happen, and the `**source_caption**` is silent on that aspect, treat it as entailment since if there were a change the `**source_caption**` would have mentioned it. Similarly, also include natural or highly probable attributes/states that would only be mentioned in the source if they deviated from the norm. For example: If there's mention of an airplane, we can assume it's large, as most airplanes naturally are. However, this only applies to immediately obvious and universally expected attributes. Any substantial elaboration or specific details beyond the most basic expectations should still be treated as undetermined.

- ****Contradiction:**** The target line contains a direct conflict with the information in the `source_caption`. This means that the target asserts a fact or detail that explicitly opposes a fact stated in the `source_caption` (for example, attributing an action to Person-X in the source versus Person-Z in the target).

- ****Undetermined:**** The target line introduces additional details or makes occurrence or attribute claims that are semantically independent from the information provided in the `source_caption`. In these cases, the `source_caption` neither provides strong evidence to support the extra details nor directly contradicts them. This category also covers instances where coreference between events or entities is ambiguous|for example,

when it is unclear whether a new event or entity mentioned in the target corresponds to one in the source. In such cases, because the evidence is insufficient to clearly support or refute the claim, the relationship should be classified as undetermined.

For each line in the **target_caption**, first output the information category, i.e., either it's a summary sentence (summary), or describing a static visual detail of the video like color of an entity (visual-description), or a dynamic action that includes events, attribute and relationship changes, etc (dynamic-action). Next output an evidence line or phrase from the **source_caption** that serves as the basis for your verdict. If no evidence exists, use an empty string. Then, provide reasoning for your verdict based on the evidence, followed by the final classification: "entailment," "contradiction," or "undetermined." Do not include any additional commentary or summary.

****Output Format:****

Line {num}: {Line text}

- Type: {summary/visual-description/dynamic-action}
- Evidence: {quote a specific line/phrase from the source upon which to base your verdict; use an empty string if no evidence exists.}
- Reasoning: {explanation for the verdict based on the evidence}
- Verdict: {entailment/contradiction/undetermined}

****Final-Instructions:****

1. Treat the **source_caption** as the premise and each line of the **target_caption** as a hypothesis.
2. For each hypothesis line, decide if it is:
 - ****Entailment****: Completely supported by the source_caption.
 - ****Contradiction****: Contains information that directly conflicts and is unsupported by the source_caption.
 - ****Undetermined****: Contains additional details or claims that the source_caption neither strongly supports nor directly contradicts. Mostly includes cases of ambiguous coreference between entities or events.
3. Follow the output format exactly as specified, without any extra summaries or notes.

In-Context Examples:

{IN_CONTEXT_EXAMPLES}

```
**source_caption:**  
"""  
{source_caption}  
"""
```

```
**target_caption:**  
"""  
{target_caption}  
"""
```

J. Details Regarding Qualitative Examples

J.1. Figure-1 Details.

Human Caption:

The black title screen reads \SUNFEAST PASTA TREAT, cooking, 45 sec., 14th May 2008" in white font. A boy enters his home holding a football and wearing a green t-shirt. A woman, wearing a pink top, calls him over to the kitchen and asks him to explain how \Sunfeast Pasta" is cooked. The boy stands on a stool and puts on a white chef's hat. He jumps off the stool and seems to signal to the woman to get something. Now, the boy is holding a pair of scissors and cuts the pasta packet where the brand name "Sunfeast" is clearly visible. The woman shows the boy some teacups. We see a shot of boiling water. The woman passes the pasta to the boy, followed by a shot of the pasta being added to the boiling water. We see spices being added to the pasta. The boy grabs a cooking stick and stirs the pasta. The woman puts the garnished pasta on a plate, while the boy places a white cloth on the table. The woman appears impressed and places the pasta on the table. Then we see a paper with the pasta recipe: 2 tea cups of water, pasta, sauce mix, and 8 minutes. The woman eats the pasta and looks impressed. She salutes the boy, who bows. At the end, we see the company's brand name, logo, their tagline "Quick and easy".

Gemini-Flash-2.0 Caption:

Here's a detailed description of the video, based on the images provided:

****General Impression:****

The video appears to be a commercial for Sunfeast Pasta Treat, a quick-and-easy pasta product. It features a mother and son preparing the pasta together. The overall tone is upbeat and suggests convenience.

****Key Visual Elements:****

1. ****Introductory text on a black screen:**** The video starts with text stating "SUNFEAST PASTA TREAT," followed by "*cooking*," "45 sec." and the date "14th May, 2008."
2. ****Family Scene:**** The initial scenes show a child wearing a green shirt and a chef's hat looking at his mother standing near a kitchen shelf
3. ****Product Focus:**** The video clearly highlights the packaging of the "Sunfeast Pasta Treat" product, with the bag appearing close up to the camera. A pair of hands is shown cutting open the product's packaging. The contents of the pasta treat are displayed.
4. ****Cooking Process:**** We see shots of the pasta being cooked, likely to emphasize the simplicity of the preparation.

5. ****Appearance of the finished product:**** The finished pasta is served on a white plate, decorated with herbs.
6. ****Mother in the car:**** There are multiple shots of a woman in the car.
7. ****Child's "Recipe Card":**** Briefly, there is a close-up of a handwritten note that contains instructions.
8. ****"Quick and Easy" Text Overlay:**** Text emphasizing the convenience of the product appears on screen, accompanying a display of the pasta varieties.

****Overall Narrative:****

The commercial likely follows this narrative:

- * Introduction: Setting the scene, indicating it's a cooking segment/advertisement.
- * Family Cooking: It looks to feature the child in the kitchen with his mother.
- * Showcasing product: Emphasizing the convenience and ease of use.
- * End: the "Quick and Easy" tagline to reinforce the product's core benefit.

J.2. Figure-2 Details.

Human Caption:

In a natural daylight setting, the video begins with a close-up shot of a thin, grayish tree branch against a blurred background of green foliage. The branch splits into two sub-branches in the middle and extends slightly upward toward the left side of the screen.

On the rightmost edge of the frame, the chin and nose of an animal become visible. It appears reddish-brown, with a sharply triangular-shaped snout, while the rest of its body remains unseen. As the video progresses, the animal moves from right to left, gradually revealing its large, expressive eyes that move slowly and deliberately. Around its nose and eyes, intricate dot-like patterns emerge, varying in color|some blending with its skin while others, particularly between the eye and stomach, take on a grayish hue. These grayish patterns extend beyond just dots, forming irregular patches along the face and stomach area. Below its chin, faint grill-like structures can also be seen.

The animal is a chameleon. As it moves forward, it lifts its left limb and grips the portion of the branch ahead. After this initial movement, it briefly pauses, its ever-shifting eyes meticulously scanning the surroundings. Then, in a sudden transformation, its reddish skin changes to a bright pink, though the gray dots and patches remain intact. After the color transformation, the chameleon continues forward, now using its right limb to grasp the portion of the branch ahead, momentarily flicking out its tongue while its eyes persist in their careful assessment of the environment.

Following another brief pause, the chameleon reaches the point where the branch splits, its mouth now close to the division. In an instant, its color shifts once more|from bright pink to green|while the gray dots remain unchanged. During this transition, an intermediate blend of orange and yellow briefly appears. The chameleon then holds its position, its eyes still in constant motion, meticulously observing its surroundings. The background remains consistently blurred throughout.

LLaVa-OV-7B:

The video begins with a close-up of a thin, gray branch set against a blurred green background, likely foliage. A red chameleon with distinctive patterns on its skin appears and starts to climb the branch. The chameleon's eyes are large and expressive, and it moves slowly and deliberately,

showcasing its unique features. The scene transitions to another red chameleon with similar patterns, also climbing the branch. This chameleon is more vibrant in color and has a slightly different pattern compared to the first one. The background remains blurred green foliage, emphasizing the chameleon's movement and interaction with the branch. The video then shifts to a green chameleon with white spots on its skin, also climbing the same thin, gray branch. The green chameleon's eyes are large and expressive, and it moves slowly and deliberately, showcasing its unique features. The background remains blurred green foliage, emphasizing the chameleon's movement and interaction with the branch. The video concludes with the green chameleon continuing to climb the branch, highlighting its vibrant color and distinctive patterns.