# ARGUS: Hallucination and Omission Evaluation in Video-LLMs

Ruchit Rawal  Reza Shirkavand  Heng Huang  Gowthami Somepalli*  Tom Goldstein*

University of Maryland, College Park

## Abstract

*Video large language models have not yet been widely deployed, largely due to their tendency to hallucinate. Typical benchmarks for Video-LLMs rely simply on multiple choice questions. Unfortunately, VideoLLMs hallucinate far more aggressively on freeform text generation tasks like video captioning than they do on multiple choice verification tasks. To address this weakness, we propose ARGUS, a VideoLLM benchmark that measures freeform video captioning performance. By comparing VideoLLM outputs to human ground truth captions, ARGUS quantifies dual metrics. First, we measure the rate of hallucinations in the form of incorrect statements about video content or temporal relationships. Second, we measure the rate at which the model omits important descriptive details. Together, these metrics form a comprehensive view of video captioning.*

## 1. Introduction

Video Large Language Models (VideoLLMs) [11, 43, 52] have made significant strides in recent years, and improvements in their capabilities have been reflected in rising scores on benchmarks for both short video [5, 44] and long video [18, 39, 41, 42, 48, 57]. Despite these strides, VideoLLMs are not yet ready for widespread deployment. The main challenge is their tendency to hallucinate, with the frequency and severity of these hallucinations increasing along with the size and complexity of the input video.

Recent studies [3, 28] indicate that while language models are often effective as verifiers (providing yes/no or multiple choice answers), they are less reliable as freeform generators. This disparity highlights a crucial limitation: a VideoLLM's ability to verify the presence of an object or the occurrence of an event does not necessarily extend to open-ended tasks like dense video captioning. The latter is particularly important for models assisting users with perceptual disabilities. Hence, there is a pressing need for a dedicated benchmark that evaluates VideoLLMs on their freeform generative capabilities rather than solely on question-answering.

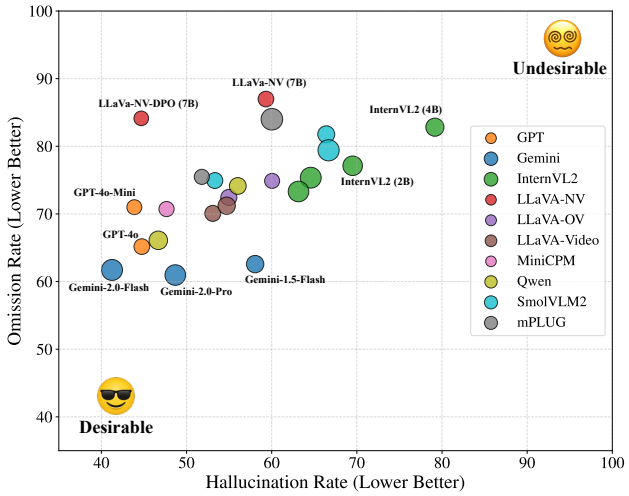


Figure 1. **Relationship between Hallucination and Omission**. The hallucination and omission cost metrics are correlated; however, most models exhibit more omissions than hallucinations. Marker size indicates the average caption length per model. Gemini-2.0-Flash achieves the best performance.

In this work, we develop ARGUS[1], a novel benchmark that evaluates the rate of hallucination in free-form video captions from VideoLLMs. Unfortunately, measuring hallucination alone is problematic — a model can avoid falsehoods simply by generating the empty string, and so it is meaningless to measure freeform accuracy without a dual metric of completeness. Hallucinations and omissions represent two sides of the same fundamental challenge — ensuring both the accuracy and completeness of video understanding. To the best of our knowledge, no benchmarks currently exist that systematically evaluate hallucinations and omissions in a freeform setting for VideoLLMs.

Specifically, to address this gap, the ARGUS framework compares the sentences from the VideoLLM generated caption to the human sentences, and an entailment analysis is used to quantify the rate hallucinations in the form of (i)

*Equal contribution. Correspondence: ruchitr@umd.edu

[1]Named after Argus, the hundred-eyed giant of Greek mythology renowned for his vigilance and ability to monitor every detail.

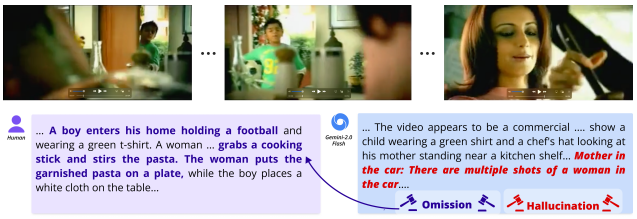


Figure 2. An example annotation (see here) of a video by the Gemini-2-Flash model. Using our framework ARGUS, we identify both hallucinations and omissions in this dense caption. We present the full human and model-captions in Appendix J.

inaccurate summarization, (ii) incorrect visual descriptions, and (iii) incorrect depiction of temporal relationships. At the same time, the VideoLLM caption is analyzed to ensure that it contains all the important statements extracted from the human caption, and the rate of omissions is quantified.

Our goal with ARGUS is to provide a fair platform for benchmarking and comparing models in a freeform setting, as depicted in Fig. 1. We posit that enhancing the accuracy and comprehensiveness of dense captioning will naturally bolster other downstream tasks such as VideoQA and reasoning, as a more holistic understanding of video content is inherently acquired. Dataset and artifacts are available at https://ruchitrawal.github.io/argus.

## 2. Related Work

**Hallucination Evaluations in Vision.** Many benchmarks such as CHAIR [49] and others [24, 27, 55] are object-centric, employing heuristics to determine the presence or absence of specific objects. Other approaches, like POPE [33], use Yes/No questions to assess how well a model understands a video, while others [36, 51] have integrated large language models into their pipelines for end-to-end annotations. These strategies often encounter challenges related to scalability, incomplete evaluations, or the introduction of additional errors stemming from the limitations of using LLMs as end-to-end evaluators. For a more comprehensive discussion see [37].

While understanding hallucinations is a bit mature research topic in Image captioning models, doing the same for videos is still very new. VideoHallucer [58], one of the first works, evaluates hallucinations in Video-LLMs using an adversarial binary question-answering approach, where a model is considered to be hallucinating if it answers either a basic or a carefully designed hallucinated question incorrectly. EventHallusion [22] follows a similar question-answering-based strategy but extends it to both binary and open-ended questions, focusing primarily on short, single-event videos averaging 11 seconds in length. VIDHAL [12] takes a different approach, proposing a caption reordering task to assess whether Video-LLMs can verify whether a caption contains more hallucinated content than another.

Existing benchmarks measure a model's ability to verify the presence of content, we focus on the more challenging task of generating open-ended captions. Additionally, unlike prior works, we measure both hallucination and omission. We compare our benchmark with other works in Tab. 1, demonstrating that we are the only dedicated benchmark for dense captioning. Moreover, we do not lose out on the number of videos or video length while providing more fine-grained evaluations than other benchmarks.

| Dataset | # Vid. | Avg. Sec | Eval. Strategy | Evals Video |
|---|---|---|---|---|
| HallusionBench [22] | 20 | $\leq$4 | QA | 8 |
| VideoHallucer [58] | 948 | 85.6 | QA | 1.89 |
| VIDHAL [12] | 1000 | 15.8 | Capt. Order | 1 |
| Event-Hallu. [65] | 397 | 11.2 | QA | 1.77 |
| **ARGUS (Ours)** | 500 | 26.3 | Dense-Cap | 19 |

Table 1. **Comparison of various video-hallucination benchmarks.** "Evals Video" refers to the number of evaluations done on average per video, task could be different depending on the benchmark. Refer to Sec. 2 for the exact tasks.

**Limitations of Current QA Approaches.** Existing benchmarks 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 entities) and another that expects a "No" (flagging potential hallucinations). Although this setup simplifies evaluation, it suffers from the following limitations - **(1) Lack of Dependence on Visual Understanding:** Many binary QA-based evaluations [34, 68] contain questions that can be answered without processing the visual input. To illustrate this, in our experiment with GPT-4o on a subset of VideoHalluCer [58], the text-only model correctly answered 32.52% of basic hallucination-related question pairs (with 61.33% accuracy in the `external_nonfactual_instruct` subcategory), despite a 25% chance performance. **(2) Verification Ability Does Not Equate to Strong Generation:** Video LLMs can verify facts in a QA setting yet fail in open-ended generation; for instance, while LLaVa-OV-7B [30] verified that there is one chameleon in a clip, it mistakenly generated a caption describing two [28] (see Fig. 3 for a qualitative example and Appendix A for quantitative results). **(3) Restricted Error Coverage Due to Predefined Scope:** The QA-based approach is limited by its predefined set of questions, covering only a narrow range of errors and leaving many hallucinations undetected until free-form captions are generated. **(4) Inability to Capture Multi-Event Hallucinations:** QA-based methods, often focusing on isolated short events, do not account for complex interactions and temporal dependencies in videos with multiple interrelated events (e.g., four sequential events with a 50% chance of guessing correctly in a binary setting), unlike free-form generation strategies. Please refer to Appendix A for expanded

discussion on this topic and Appendix B for related work on Natural Language Inference and Dense Video Captioning.

## 3. ARGUS: Freeform Captioning Benchmark

We propose ARGUS, a novel evaluation framework to quantify both hallucinations and omissions in VideoLLM captions that addresses the drawbacks of previous benchmarks. We introduce a metric ArgusCost-H that quantifies a model's average rate of hallucination, and a metric ArgusCost-O that reflects the rate of omissions in VideoLLM. In the remainder of this section, we detail the computation of the hallucination metric, ArgusCost-H. We use a similar methodology to compute ArgusCost-O.

We analyze VideoLLMs captions by matching selected sentences to a corresponding sentence in a ground-truth human caption. Using this matching, we identify two types of hallucinations. First, we consider hallucinations where the model incorrectly recognizes *false content* in the video. We identify such errors by using Natural Language Entailment (with an LLM as a judge) between the VideoLLM sentence and its matched ground truth. Second, we penalize hallucinations in which the sentences are correct on their own, but the model misrepresents the *order of events*. It is important to score these errors, as time is the key difference between image and video understanding. We check whether the order of sentences that make temporal claims in the VideoLLM caption matches the order of their corresponding sentences in the human caption, and assign a penalty to each sentence that is proportional to its level of anachronism.

If the model fabricates an event that did not happen, this should be considered a content error and not a temporal error, and therefore left out of the temporal matching. However, if such a fabrication gets mistakenly matched to a similar but faraway sentences in the ground truth, this will disrupt the matching between otherwise correct temporal statements and cause artificially large error scores. To prevent hallucinated events from disrupting the temporal matching, we use a dynamic program that assigns each event to either the content or temporal error category in order to minimize the overall temporal matching penalty.

### 3.1. Sentence-Level Entailment

For a given video, we assume access to a set of $m$ source sentences annotated by human annotators, i.e., $S = \{s_1, s_2, \ldots, s_m\}$, describing the video in detail. Similarly, we also assume access to $n$ target sentences generated by a VideoLLM, i.e., $T = \{t_1, t_2, \ldots, t_n\}$. Our goal is to determine whether any of these target sentences contain hallucinated content that has no grounding in $S$. If a target sentence $t_i$ is entailed by the source set $S$, then it is considered valid; otherwise, it is hallucinated. LLMs and reasoning models now match human performance on complex logical reasoning tasks [23, 38] and are increasingly used as "judges" in dynamic evaluation pipelines [31, 66]. We, hence use a strong model, GPT-4o, as the entailment judge. We tested alternatives like DeepSeek-V3 and the reasoner DeepSeek-R1 in Sec. 4.3, finding similar results. We input both $S$ and $T$ to an LLM-judge, which evaluates each target sentence $t_i \in T$ along three key dimensions:

- A type $\theta_i \in \{\text{SUM}, \text{VD}, \text{DA}\}$, categorizing $t_i$ as either a summary (SUM), a visual description (VD), or a dynamic action (DA).
- A verdict $v_i \in \{\text{EN}, \text{CON}, \text{UD}\}$, indicating whether $t_i$ is entailed (EN), contradictory (CON), or undetermined (UD) with respect to $S$.
- An evidence line $e_i \in \{1, 2, \ldots, m\} \cup \{\text{null}\}$, corresponding to the location of a supporting sentence in $S$ or marked as null if no supporting evidence exists.

**Why do we need type and verdict categorization?** ARGUS scores sentences differently depending on their type. Dynamic actions have an inherent temporal structure, requiring a penalty when their order described in the target caption deviates from the human caption. In contrast, summarization and visual-descriptions do not follow a strict temporal order, but should still be checked for entailment.

### 3.2. Calculating Total Cost

We first separately define for two key components of the hallucination cost: (1) the base cost, which captures the penalty for sentences that violate entailment, and (2) the order penalty, which accounts for temporal misalignment in dynamic actions. We then unify these costs using a dynamic programming formulation that computes the optimal alignment between the source $S$ and generated caption $T$.

#### 3.2.1. Base Cost Matrix

We define a cost matrix $C \in \mathbb{R}^{n \times m}$, where each entry represents the hallucination cost for a target sentence $t_i$ given its evidence in the source sentence $s_j$:

$$C_{ij} = \begin{cases} 1, & \text{if } v_i \neq \text{EN}, \\ 0, & \text{if } v_i = \text{EN and } \theta_i \in \{\text{SUM}, \text{VD}\}, \\ \begin{cases} 0, & \text{if } j = e_i, \\ 1, & \text{otherwise}, \end{cases} & \text{if } v_i = \text{EN and } \theta_i = \text{DA}. \end{cases} \quad (1)$$

Visualized in Fig. 4, the Base Cost Matrix primarily captures the cost associated with hallucinations based on entailment labels. If $t_i$ is not entailed ($v_i \neq$ EN), we assign a maximum penalty of 1. Conversely, if $t_i$ is entailed and belongs to the summary (SUM) or visual description (VD) category, we assign a cost of 0. This is because summary sentences often do not correspond to a specific source sentence but rather capture the overall essence of the video, while visual descriptions provide general scene details that do not adhere to a strict temporal structure.

For dynamic actions (DA), however, temporal alignment matters. If $t_i$ is entailed, we check whether its supporting

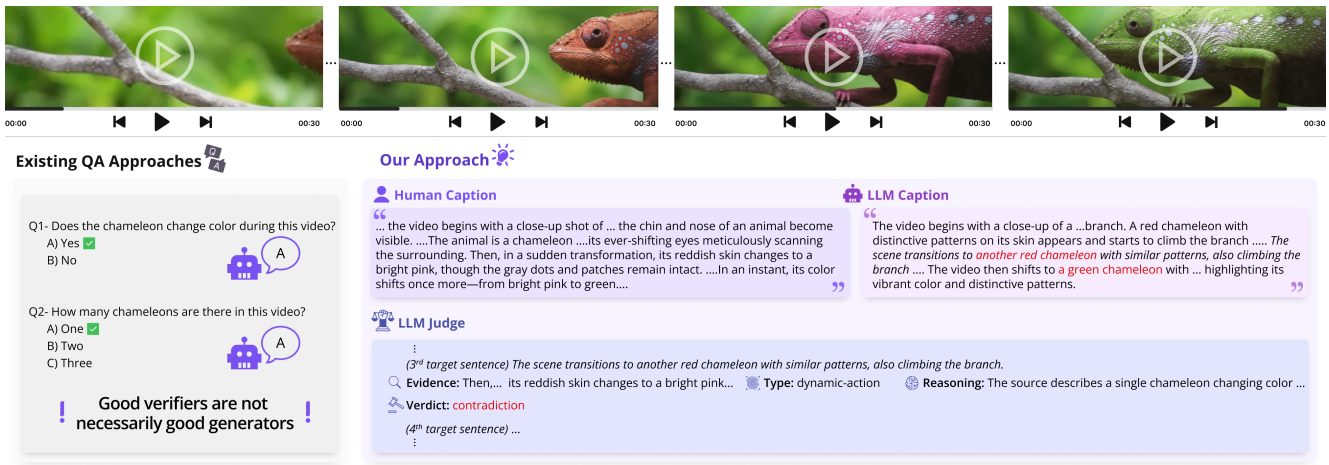


Figure 3. A Video LLM can correctly answer targeted questions about a video (left; see here) but still generate hallucinated content when describing the video (right), highlighting the disconnect between verification and open-ended generation. Our approach evaluates hallucinations at the sentence level from the generated caption, providing a more precise and comprehensive assessment. Refer to Sec. 3.1 to see how we ground each sentence using an LLM judge to compute a hallucination score. Example LLaVa-OV-7B caption and response on a video from ArgusBench. We present the full human and model-captions in Appendix J

evidence $e_i$ is the same as the source index $j$. If $j = e_i$, we assign a cost of 0, indicating a correct match. Otherwise, we impose a penalty of 1, reflecting a misalignment in the order in which events are described.

### 3.2.2. Dynamic Programming Formulation

**DP State:** We define a DP table $D \in \mathbb{R}^{(n+1)\times m}$ where each entry $D(i, j)$ represents the minimum cost of aligning the first $i$ target sentences with the source sentences, such that the $i$-th target is matched with source sentence $s_j$. In addition to tracking the cost, we also need to maintain information about the matching structure—specifically, which source sentences were matched with previous target sentences. This alignment history is necessary for computing ordering penalties accurately.

**Alignment History:** Let $A_{i,j}$ denote the sequence of source indices $(a_1, a_2, \ldots, a_{i-1}, j)$ that represents the optimal alignment of the first $i$ target sentences, with the $i$-th target aligned to source $s_j$. This alignment history enables us to evaluate temporal consistency in the sequence of described actions.

**Ordering Penalty:** The ordering penalty function quantifies temporal inconsistencies in the alignment. Given an alignment history $A_{i,k}$, we define:

$$\pi(A_{i,k}, j) = \lambda \cdot \sum_{r \le i} \mathbf{1}\Big[a_r > j\Big] \quad \text{s.t. } \theta_r = \theta_j = \text{DA} \quad \text{and } v_r = v_j = \text{EN}, \tag{2}$$

where $0 \le \lambda \le 1$ is a penalty factor, and $\mathbf{1}[\cdot]$ is the indicator function. This function counts ordering violations between pairs of dynamic actions. A violation occurs when an earlier action in the target sequence is aligned with a later action in the source sequence, relative to another action pair. The severity of these violations increases with the number of pairs that appear out of order, making sequences with multiple ordering inversions more heavily penalized than those with fewer inversions.

Intuitively, as $\lambda$ increases, ARGUS becomes less tolerant of temporal inconsistencies; sentences placed significantly out of their correct order may effectively be treated as hallucinations during the DP-recurrence minimization. In our setup, we use $\lambda = 0.1$.

**Recurrence Relation:** Intuitively, $D(i, j)$ equals the total cost of aligning the previous $i - 1$ sentences ($D(i - 1, .)$) plus the base cost of aligning the $i^{th}$ with the $j^{th}$ source sentence, and an ordering penalty if misalignment occurs. Formally, the recurrence relation for each $i \in 1, 2, \ldots, n$ and $j \in 1, 2, \ldots, m$:

$$D(i, j) = C_{i,j} + \min_{k \in \{1,\ldots,m\}} \{D(i - 1, k) + \pi(A_{i-1,k}, j)\} \tag{3}$$

When computing $D(i, j)$, we consider all possible alignments for the previous target sentence and choose the one that minimizes the sum of the previous cost, the base cost for the current alignment, and the ordering penalty based on the full alignment history. Note, that the base-case here is $D(0, j) = 0$ for all $j \in \{1, 2, \ldots, m\}$.

**Total Cost:** The minimal total hallucination cost is given by $\min_{j \in \{1,\ldots,m\}} D(n, j)$, and the optimal matching between the source and target sentences is denoted by $A_{n,j^*}$, where $j^*$ is the source index that achieves the minimum cumulative hallucination cost.

However, different VideoLLMs can generate captions of varying lengths for the same video, making direct compar-

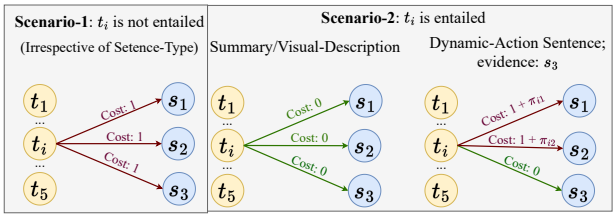


Figure 4. **Proposed dynamic programming formulation for sentence alignment.** Given sentence-level entailment scores from the LLM-as-a-judge, we compute the matching cost between a target sentence ($t_i$) and a source sentence ($s_j$) by combining the entailment score with a temporal order penalty ($\pi_{ij}$). The overall score follows the recurrence relationship defined in Sec. 3.2.

isons of cumulative hallucination cost unfair, as longer captions face higher penalties. To address this, we normalize the cost by computing the maximum possible hallucination cost for a given video, VideoLLM pair. This maximum cost is given by $(n-d) + \lambda \cdot \frac{d(d-1)}{2}$, where $d$ represents the number of dynamic-action targets with entailment. The first term accounts for the worst-case base cost when all non-entailed or non-dynamic targets receive the maximum penalty. The second term represents the highest possible ordering penalty, which occurs when all dynamic actions are perfectly inverted. We then define the normalized hallucination metric, i.e., **ArgusCost-H**, defined as the ratio of the observed cost to the maximum possible cost, as follows:

$$\text{ArgusCost-H: } \left( \frac{\min_{j \in \{1,\ldots,m\}} D(n,j)}{(n-d) + \lambda \cdot \frac{d(d-1)}{2}} \right) \cdot 100. \quad (4)$$

**Omission Cost:** Omission refers to information present in the human-annotated caption that is missing from the model-generated caption. **The normalized omission cost, ArgusCost-O, is assigned just by reversing the roles of the source ($S$) and target ($T$). We then assess whether each line in the human caption can be entailed by the LLM-generated caption.** If a human-annotated sentence is not entailed, it indicates that the model-generated caption has omitted that information.

# 4. Benchmarking Video-LLMs

We detail how we curated our evaluation dataset and describe evaluations conducted on several VLMs, along with insights derived from the ArgusCost trends.

## 4.1. ArgusBench: Curation and Statistics

We collect 500 videos along with their corresponding dense-caption pairs from three sources. First, we utilize existing video understanding datasets [7] that already contain captions. These videos are manually verified by human authors, and received well in the community. Second, we incorporate text-to-video generation datasets [17], which include reference videos and short prompts. Since these prompts are insufficient for dense captioning, we manually annotate 10 such videos. Lastly, the authors curate additional videos from publicly available sources, such as YouTube, under Creative Commons licenses. We curate 30 such videos, and also manually annotated, with cross-validation among the authors.

In Fig. 5 we provide a statistical analysis of our dataset, including our video length distribution, where nearly a quarter of the videos between 15 and 20 seconds, while 12% exceed 60 seconds. Fig. 5-middle shows the distribution of word counts (x-axis) in human captions from ArgusBench, highlighting the rich density of our textual descriptions, averaging 477 words per video and 24.4 words per second. Additionally, Fig. 5-right visualizes the distribution of sentence lengths (x-axis) in human captions from ArgusBench, illustrating how different sentence types contribute to each length category. For instance, we observe approximately 300 sentences with a length of around 20 words, which appear to be fairly evenly distributed across the three sentence categories. Unlike some existing datasets that over-represent one category, our benchmark maintains a balanced distribution, with approximately equal parts falling into each sentence type. We provide additional details on the types of video clips in our benchmark in Appendix F. Our sentence-level evaluation provides a much finer-grained analysis, averaging ≈19 evaluations per video compared to 1-2 questions per video in the baseline benchmarks.

## 4.2. Evaluations & Insights

In this section, we evaluate a range of closed-source and open-source Video-LLMs using the ARGUS framework, and examine performance trends along with key challenges faced by different models. In total, we evaluate 23 video models, including 18 open-source models and 5 state-of-the-art proprietary ones. We sample a total of 16 frames per video (unless specified otherwise) to evaluate the models. If a model has a predefined frame sampling strategy, we adapt it else we apply uniform sampling. Further details on the evaluation setup, including model checkpoints, computational and monetary costs, evaluation prompts, etc, are provided in Appendix C.

**Video-LLMs have a hallucination problem.** We present the ArgusCost-H–the normalized hallucination costs–for different models in Fig. 6, where we note that even the best-in-the-market Video-LLMs generate hallucinated content frequently, with significant cost variation across models and training strategies. While Gemini-2.0-Flash achieves the lowest ArgusCost-H at 41%, outperforming both GPT-4o and GPT-4o-mini (each at approximately 44%), its larger counterpart, Gemini-2.0-Pro, performs worse at 48%. In contrast, the difference in scale between GPT-4o and GPT-

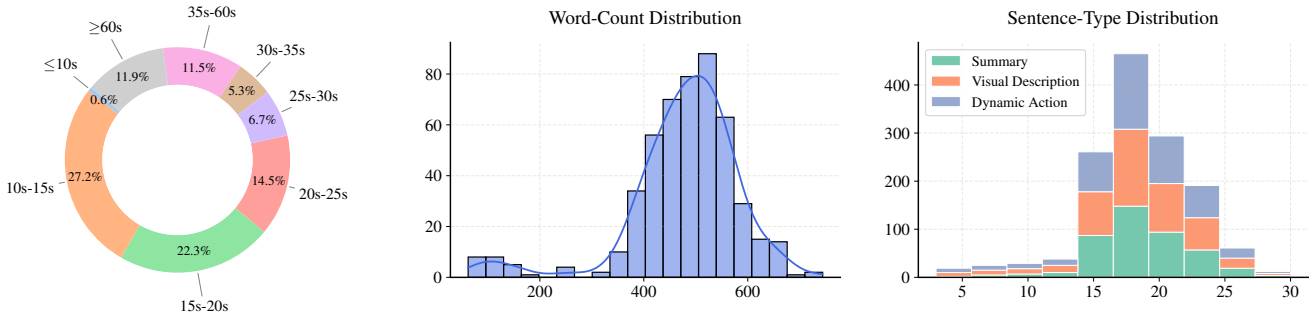


Figure 5. **ArgusBench Statistics:** video lengths (left), word-counts (middle), and sentence-length (x-axis) distribution by sentence-type (right). Our dataset has a balanced representation across durations and sentence types, and a high word-count density.

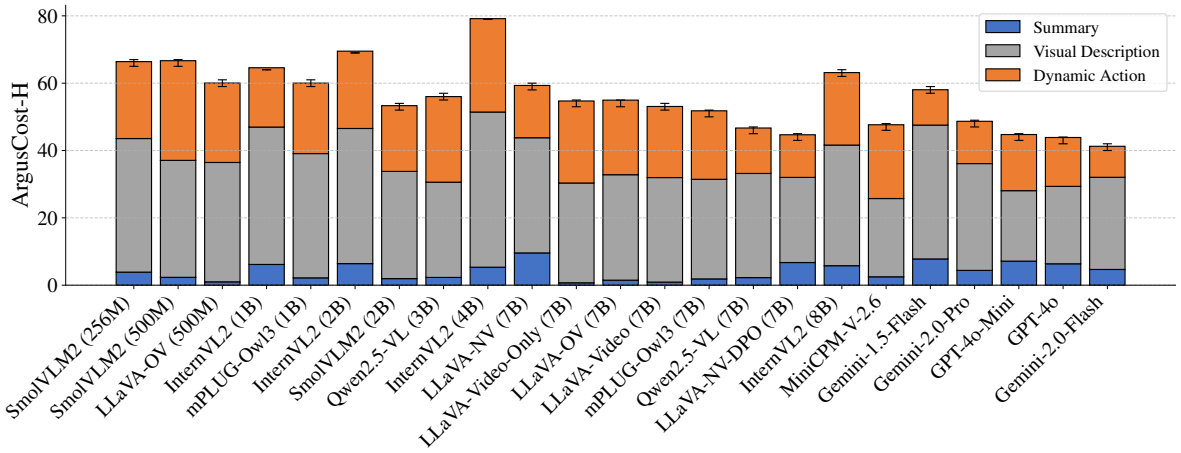


Figure 6. **ArgusCost-H (Hallucination Cost) across Video-LLMs.** Even top performers like Gemini-2.0-Flash produce up to 40% hallucinated content. Although summary errors are low, stronger models still fabricate visual details despite improved dynamic action descriptions. Open-source models are ordered by size along the x-axis. Lower ArgusCost-H is better.

4o-mini has little impact on ArgusCost-H. The Gemini-2.0 series also shows a clear improvement over the 1.5 series, with Gemini-1.5-Flash trailing behind at 58%.

Among open-source models, LLaVA-Next-Video (DPO) achieves the best result with a ArgusCost-H of 45%, closing the gap with proprietary models. However, its non-DPO version performs significantly worse at 59%, highlighting the potential role of reinforcement learning-based post-training strategies in reducing hallucination. Other strong performers in the open-source category include Qwen 2.5-VL (7B) and MiniCPM-V-2.6, both with ArgusCost-H below 50%. On the other hand, the worst-performing model is InternVL2 (4B), with a ArgusCost-H of 80%, indicating that smaller models are not inherently more prone to hallucination. In fact, some smaller models, such as SmolVLM2 (2B) and Qwen 2.5-VL (3B), perform (ArgusCost-H around 55%) substantially better than even some larger models.

Fig. 6 also breaks down ArgusCost-H by sentence type, showing that "summary" sentences contribute minimally–likely because each caption contains only a few, which models generally handle well. However, we observe that proprietary models like Gemini-2.0-Flash and GPT-4o exhibit slightly higher summary sentence errors than some weaker open-source models. We qualitatively examined these errors and found that these models tend to generate abstract interpretations of motives, atmosphere, and mood, especially when such information is subjective and cannot be reliably grounded. Errors in visual descriptions and dynamic actions are fairly evenly distributed among poorly performing models. However, as we go from models with high to low ArgusCost-H, errors in dynamic actions decrease more rapidly than in visual descriptions, suggesting that stronger video models are more reliable at describing temporal events than at avoiding fabricated visual details. We present additional results on ArgusCost-H breakdown by verdict-type (contradiction v/s underdetermined), and cost-type (base-cost v/s order penalty) in Appendix G.

**Do hallucination rates correlate with omission rates?** Two competing factors govern the relationship between hallucinations and omissions. First, overall stronger models may have lower hallucination rates and also reduced ArgusCost-O. Alternatively, weaker models may tradeoff hallucination and omission by being more (or less) verbose, thus causes an inverse relationship between the two. We

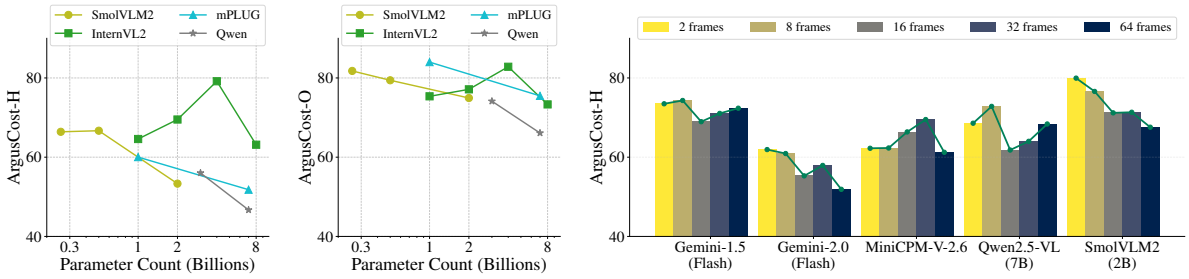


Figure 7. (Left,Middle) **Relationship between hallucination/omission and model-size.** For most open-source model families (except InternVL2), larger models tend to have lower hallucination (ArgusCost-H) and omission (ArgusCost-O) costs, indicating that scale generally improves performance. (Right) **Relationship between hallucination and frames of the video provided.** Gemini models and SmolVLM2 show a consistent reduction in hallucinations with more frames, while MiniCPM-V and Qwen2.5-VL exhibit fluctuating levels of hallucinations (as measured by ArgusCost-H). Please see Appendix G for results on the effect of frames provided on ArgusCost-O.

observed both of these trend types when we analyzed the relationship between hallucination and omission across model families, as shown in Fig. 1. We find a moderately high Pearson correlation of 0.65, supporting the first hypothesis. Most models lie above the $y = x$ line, indicating they omit more than they hallucinate. Notably, despite a significant improvement in ArgusCost-H for Gemini-2.0-Flash (41%) compared to Gemini-2.0-Pro (48%), its ArgusCost-O remains around 60%. In contrast, GPT-4o and GPT-4o-mini have similar ArgusCost-H (44%), but GPT-4o exhibits a 5% lower ArgusCost-O. LLaVA-Next-Video (DPO), which performed close to proprietary models in hallucination (45%), has a much higher ArgusCost-O (85%), suggesting the model plays it safe and avoids generating uncertain content, often missing important details. In comparison, Qwen 2.5-VL (7B) and MiniCPM-V-2.6 achieve both low ArgusCost-H and ArgusCost-O, making them better opensource choices. In Fig. 1, marker size indicates average number of sentences in captions generated by each of these models; we do not see any trends corresponding to this — *longer captions do not necessarily lead to higher or lower hallucination*. When we perform a correlation analysis, we find no correlation (0.09) between model-generated (average) caption length and ArgusCost-O, while we observe a low positive correlation (0.32) between caption length and ArgusCost-H. We also investigate which other characteristics of the video or the human-generated caption may correlate with hallucination and ArgusCost-O. For instance, we find that both ArgusCost-H and ArgusCost-O have mild positive correlation ($\approx$ 0.25) with the clip duration (in seconds). We present additional results in the Appendix G due to space constraints.

**Does scale help reduce hallucinations and omissions?** In our discussion of Fig. 6, we showed that smaller Video-LLMs do not necessarily hallucinate more and can sometimes even outperform larger models. However, that analysis grouped all models together, meaning factors like architecture, training data, and other differences would have influenced the results—not just scale. To better isolate the effect of the model scale, we analyze ArgusCost-H and ArgusCost-O within 4 model families: SmolVLM2 [1], mPLUG [63], Qwen [62], and InternVL2 [10] with sizes ranging from sub-billion to 8 billion parameters in Fig. 7 (Left, Middle). We find that for most model families, increasing scale improves performance—both ArgusCost-H and ArgusCost-O decrease. For example, in the SmolVLM2 family, scaling from 256 million to 2 billion parameters improves hallucination performance by over 15% and omission performance by 5%. Similarly, Qwen 2.5-L improves by 10% for hallucination and 8% for omission. However, this trend does not hold for the InternVL2 family, where ArgusCost-H and ArgusCost-O initially increase from 1B to 4B parameters before decreasing at 8B, ultimately returning to the level of the 1B-parameter model. It remains an open question what exact components in the InterVL2 training pipeline led to the emergence of such behavior since the smallest InternVL2 (1B) is on par with other smaller 256M and 500M models. We leave this investigation for future work.

**Effect of frame sampling rate** Previously, we sampled 16 frames per video for each model since the smallest maximum frame limit among the models in our pool was 16. However, some models can handle up to 64 frames or more, so we now investigate how increasing the total number of frames affects ArgusCost-H and ArgusCost-O. We perform this ablation on a subset of 50 randomly sampled videos from our benchmark. We present the results in Fig. 7. We find that Gemini models and SmolVLM2 consistently improve in both ArgusCost-H and ArgusCost-O as the number of frames increases, suggesting that these models effectively leverage additional frame information to generate more reliable captions—reducing both fabricated and omitted information. For example, the ArgusCost-H for Gemini-2.0-Flash improves by 10.02% when increasing from 2 to 64 frames, while SmolVLM2 improves by 12.38%. However, for some open-source models, such as MiniCPM-V-

| Model | ArgusCost-H | |
|---|---|---|
| | Intra-Prompt | Inter-Prompt |
| SmolVLM2 | $72.7 \pm 2.9$ | $73.0 \pm 3.0$ |
| InternVL2 | $83.4 \pm 1.6$ | $83.4 \pm 1.6$ |
| MiniCPM-V-2.6 | $69.2 \pm 2.4$ | $67.0 \pm 2.4$ |
| Qwen2.5-VL | $62.1 \pm 2.8$ | $62.9 \pm 2.9$ |

Table 2. **Sensitivity to the Prompt.** We sample multiple dense captions by varying prompt parameters and compute ArgusCost-H. The low standard deviation of ArgusCost-H across many Video-LLMs demonstrates its robustness.

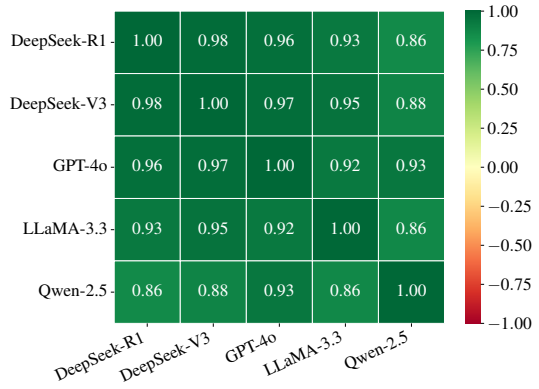


Figure 8. **Sensitivity to different LLMs-as-judge.** High Pearson ranking correlations across different LLMs-as-judge indicate strong agreement in evaluations and robustness in ranking trends.

2.6 and Qwen2.5-VL, the effect on ArgusCost-H is less clear. MiniCPM-V-2.6 initially worsens as the number of frames increases but then drops back to a similar level as with 2 frames at 64 frames. Qwen2.5-VL follows an inverse pattern, where hallucination first improves but then worsens at 64 frames. For omission (Appendix G), all models show consistent improvement as the number of frames increases. This suggests that while using more frames can help models capture more information and reduce omissions, models may also struggle to balance relevant and irrelevant details, leading to greater reliance on learned patterns, abstraction, or speculative inferences rather than staying grounded in the actual video content. Additionally, attention and memory limitations can force models to summarize or interpolate details inaccurately, further increasing the risk of hallucinations.

### 4.3. ARGUS Sensitivity Analysis

Since one of the important steps of our evaluation process involves inherent stochasticity—such as the choice of prompt for generating captions—one concern is whether the observed trends generalize. To address this, we conduct a sensitivity analysis on a subset of 50 randomly sampled clips, evaluating intra-prompt variation, inter-prompt variation, and the choice of LLM-as-judge.

For intra-prompt analyses, we sample three captions for each model using our default prompt ("Describe the video in great detail.") at the default decoding temperature and analyze the variation in ArgusCost-H. For inter-prompt analyses, we run experiments with 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.", and analyze the variation in ArgusCost-H. Tab. 2 shows low standard errors in the range of 1–3 for all model settings. As expected, the error is slightly higher in inter-prompt setting compared to the intra-prompt setting.

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]. Fig. 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. A detailed discussion and additional sensitivity analyses are provided in Appendix D. Additionally, we conducted a human study to assess the reliability of our LLM-based evaluation, where we observed a high average agreement rate of $91.26\%$, with most disagreements arising from fine-grained visual details. Further methodology and analysis are provided in Appendix E.

## 5. Discussion & Conclusion

In this work, we present ARGUS, a first-of-its-kind evaluation framework for quantifying hallucinations and omissions in dense video captioning, addressing key limitations of previous QA-based approaches. We propose dual metrics, ArgusCost-H and ArgusCost-O, to assess the accuracy and completeness of generated captions. Our experiments show that even top VideoLLMs, such as Gemini-2.0-Flash, produce a significant amount of hallucinated content, highlighting the gap between targeted verification and open-ended generation. Moreover, these trends are consistent across multiple samplings, varied input prompts, and different LLMs-as-judges. These findings emphasize the need for future research to mitigate hallucinations and improve the overall accuracy of dense video captions.

## Acknowledgments

The authors would like to thank Yuxin Wen, Sachin Shah, Sean McLeish, Udit Chugh, Prachi Rawal, and many others who helped us with the human study.

This work was made possible by DARPA TIAMAT and the NSF TRAILS Institute (2229885), NSF IIS 2347592, 2348169, DBI 2405416, CCF 2348306, CNS 2347617. Commercial support was provided by Capital One Bank, the Amazon Research Award program, and Open Philanthropy.

## References

[1] SmolVLM2: Bringing Video Understanding to Every Device — huggingface.co. https://huggingface.co/blog/smolvlm2, 2025. [Accessed 07-03-2025]. 7

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 14

[3] Anonymous. Are language models better at generating answers or validating solutions?, 2025. Under review. 1

[4] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. The snli corpus. 2015. 13

[5] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015. 1

[6] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31, 2018. 13

[7] Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jeng-Neng Hwang, Saining Xie, and Christopher D Manning. Auroracap: Efficient, performant video detailed captioning and a new benchmark. *arXiv preprint arXiv:2410.03051*, 2024. 5, 16, 17

[8] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024. 16

[9] Zeming Chen, Qiyue Gao, and Lawrence S Moss. Neurallog: Natural language inference with joint neural and logical reasoning. *arXiv preprint arXiv:2105.14167*, 2021. 13

[10] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 7

[11] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *CoRR*, abs/2404.16821, 2024. 1

[12] Wey Yeh Choong, Yangyang Guo, and Mohan Kankanhalli. Vidhal: Benchmarking temporal hallucinations in vision llms. *arXiv preprint arXiv:2411.16771*, 2024. 2

[13] Dorottya Demszky, Kelvin Guu, and Percy Liang. Transforming question answering datasets into natural language inference datasets. *arXiv preprint arXiv:1809.02922*, 2018. 13

[14] Chaorui Deng, Shizhe Chen, Da Chen, Yuan He, and Qi Wu. Sketch, ground, and refine: Top-down dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 234–243, 2021. 14

[15] Hugging Face. diffusers/shot-categorizer-v0. https://huggingface.co/diffusers/shot-categorizer-v0, 2025. 20

[16] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19358–19369, 2023. 14

[17] Weixi Feng, Jiachen Li, Michael Saxon, Tsu-jui Fu, Wenhu Chen, and William Yang Wang. Tc-bench: Benchmarking temporal compositionality in text-to-video and image-to-video generation. *arXiv preprint arXiv:2406.08656*, 2024. 5, 16

[18] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 1

[19] Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. Trueteacher: Learning factual consistency evaluation with large language models. *arXiv preprint arXiv:2305.11171*, 2023. 13

[20] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 8, 15

[21] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022. 16

[22] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of*

*the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024. 2

[23] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 3, 8, 15

[24] Hongyu Hu, Jiyuan Zhang, Minyi Zhao, and Zhenbang Sun. Ciem: Contrastive instruction evaluation method for better instruction tuning. *arXiv preprint arXiv:2309.02301*, 2023. 2

[25] Vladimir Iashin and Esa Rahtu. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. *arXiv preprint arXiv:2005.08271*, 2020. 14

[26] Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 958–959, 2020. 14

[27] Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. *arXiv preprint arXiv:2311.01477*, 2023. 2

[28] Prannay Kaul, Zhizhong Li, Hao Yang, Yonatan Dukler, Ashwin Swaminathan, CJ Taylor, and Stefano Soatto. Throne: An object-based hallucination benchmark for the free-form generations of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27228–27238, 2024. 1, 2, 12

[29] Minkuk Kim, Hyeon Bae Kim, Jinyoung Moon, Jinwoo Choi, and Seong Tae Kim. Do you remember? dense video captioning with cross-modal memory retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13894–13904, 2024. 14

[30] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *CoRR*, abs/2408.03326, 2024. 2

[31] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024. 3

[32] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Jointly localizing and describing events for dense video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7492–7500, 2018. 14

[33] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 2

[34] Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. Revisiting the role of language priors in vision-language models. *arXiv preprint arXiv:2306.01879*, 2023. 2, 12

[35] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 8, 15

[36] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023. 2

[37] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024. 2

[38] Hanmeng Liu, Zhizhang Fu, Mengru Ding, Ruoxi Ning, Chaoli Zhang, Xiaozhang Liu, and Yue Zhang. Logical reasoning in large language models: A survey. *arXiv preprint arXiv:2502.09100*, 2025. 3

[39] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023. 1

[40] Jonghwan Mun, Linjie Yang, Zhou Ren, Ning Xu, and Bohyung Han. Streamlined dense video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6588–6597, 2019. 14

[41] Arsha Nagrani, Mingda Zhang, Ramin Mehran, Rachel Hornung, Nitesh Bharadwaj Gundavarapu, Nilpa Jha, Austin Myers, Xingyi Zhou, Boqing Gong, Cordelia Schmid, et al. Neptune: The long orbit to benchmarking long video understanding. *arXiv preprint arXiv:2412.09582*, 2024. 1

[42] Arsha Nagrani, Sachit Menon, Ahmet Iscen, Shyamal Buch, Ramin Mehran, Nilpa Jha, Anja Hauth, Yukun Zhu, Carl Vondrick, Mikhail Sirotenko, et al. Minerva: Evaluating complex video reasoning. *arXiv preprint arXiv:2505.00681*, 2025. 1

[43] OpenAI. Gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. 1

[44] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[45] Iqra Qasim, Alexander Horsch, and Dilip Prasad. Dense video captioning: A survey of techniques, datasets and evaluation protocols. *ACM Computing Surveys*, 57(6):1–36, 2025. 14

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 14

[47] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 20

[48] Ruchit Rawal, Khalid Saifullah, Miquel Farré, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein.

Cinepile: A long video question answering dataset and benchmark. *arXiv preprint arXiv:2405.08813*, 2024. 1

[49] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018. 2

[50] Alessandro Scirè, Karim Ghonim, and Roberto Navigli. Fenice: Factuality evaluation of summarization based on natural language inference and claim extraction. *arXiv preprint arXiv:2403.02270*, 2024. 13

[51] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023. 2

[52] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1, 14

[53] LAION team. Laion-aesthetics_predictor v1. https://github.com/LAION-AI/aesthetic-predictor, 2022. A linear estimator on top of CLIP to predict the aesthetic quality of images. 20

[54] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 14

[55] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, et al. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023. 2

[56] Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. End-to-end dense video captioning with parallel decoding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6847–6857, 2021. 14

[57] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024. 1

[58] Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv preprint arXiv:2406.16338*, 2024. 2, 12

[59] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017. 13

[60] Adina Williams, Tristan Thrush, and Douwe Kiela. Anlizing the adversarial natural language inference dataset. *arXiv preprint arXiv:2010.12729*, 2020. 13

[61] Huijuan Xu, Boyang Li, Vasili Ramanishka, Leonid Sigal, and Kate Saenko. Joint event detection and description in continuous video streams. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 396–405. IEEE, 2019. 14

[62] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 7, 8, 15

[63] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. In *The Thirteenth International Conference on Learning Representations*, 2024. 7

[64] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 14

[65] Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Jingjing Chen, and Yu-Gang Jiang. Eventhallusion: Diagnosing event hallucinations in video llms. *arXiv preprint arXiv:2409.16597*, 2024. 2

[66] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023. 3

[67] Xingyi Zhou, Anurag Arnab, Shyamal Buch, Shen Yan, Austin Myers, Xuehan Xiong, Arsha Nagrani, and Cordelia Schmid. Streaming dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18243–18252, 2024. 14

[68] Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, et al. Apollo: An exploration of video understanding in large multimodal models. *arXiv preprint arXiv:2412.10360*, 2024. 2, 12