Hallucination Localization in Video Captioning

Anonymous ACL submission

Abstract

We propose a novel task, hallucination localization in video captioning, which aims to identify hallucinations in video captions at the span level (i.e. individual words or phrases). This allows for a more detailed analysis of hallucinations compared to existing sentence-level hallucination detection task. We manually annotate 1,167 hallucination instances from VideoLLM-generated captions to build HLVC-Dataset, a specialized dataset for hallucination localization. We further implement a VideoLLM-based baseline method and conduct quantitative and qualitative evaluations to benchmark current performance on hallucination localization.

1 Introduction

Video platforms, such as Netflix and YouTube, have experienced rapid growth. This has led to unprecedented volumes of video content accompanied by textual data. This expansion has made automatic video understanding an important research area in both computer vision and Natural Language Processing (NLP) (Tang et al., 2025; Madan et al., 2024). Among the various tasks within video understanding, video captioning, which describes video content using natural language, has garnered particular attention (Abdar et al., 2024). Video captioning is highly valuable as it provides summaries for users and facilitates effective video content search and recommendation.

Recently, VideoLLMs have become widely utilized in video captioning tasks (Li et al., 2024, 2023). VideoLLMs are models that integrate a video encoder with a large language model (LLM) (Grattafiori et al., 2024; Achiam et al., 2023; Yang et al., 2025) to perform various natural language tasks, such as answering questions, describing scenes, and summarizing video content. While VideoLLMs generate versatile and fluent captions, they inherit the hallucination problem common in LLMs, producing content that is not supported by or contradicts the input video (Huang et al., 2025; Ma et al., 2024). Such hallucinated captions may mislead users and diminish the system's trustworthiness, particularly when captions serve as official summaries or input for downstream tasks. Therefore, addressing hallucination in video captioning is crucial for deploying VideoLLMs safely and reliably.

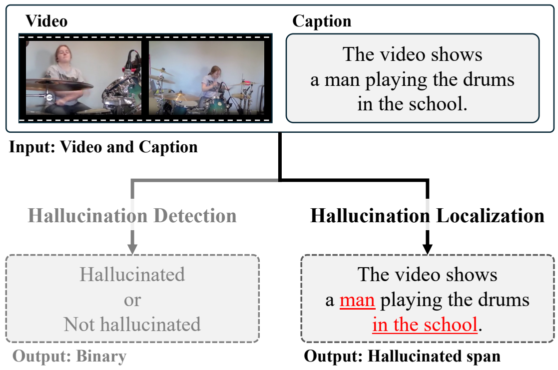


Figure 1: **Comparison between the hallucination detection and hallucination localization.** Given a video and its caption, hallucination detection classifies whether the caption contains hallucinated content (binary classification). In contrast, our proposed hallucination localization identifies the text span responsible for the hallucination.

Researchers have actively explored the issue of hallucinations in video captioning. These efforts include developing dedicated benchmarks (Choong et al., 2024) and designing improved model architectures (Ullah and Mohanta, 2022). Among these research directions, sentence-level hallucination detection, which involves identifying incorrect captions at the sentence level, is particularly important (Shi et al., 2022; Liu and Wan, 2023). Specifically, hallucination detection in video captioning is formulated as a binary classification task, where the model determines whether a caption contains hallucinations based on the video-caption pair. This step is essential for providing feedback about caption errors to users, thereby preserving the overall reliability of video systems.

However, sentence-level hallucination detection suffers from critical limitations due to its coarse granularity. While existing hallucination detection methods operate at the sentence level, hallucinations in video captions typically occur at finer granularity, such as individual words or phrases. For example, Figure 1 illustrates a case where the caption '*The video shows a man playing the drums and singing*' is provided for a video that actually depicts a woman playing drums. Here, hallucinations occur only within specific spans, such as '*man*' and '*and singing*'. Existing sentence-level hallucination detection overlooks these detailed errors, thereby limiting thorough analysis of caption quality. Moreover, providing only sentence-level warnings to users does not specify the exact source of hallucination, making feedback inadequate. Therefore, detailed, fine-grained feedback is critical for precise evaluation and user-oriented services.

To address these issues, we propose a novel task, hallucination localization in video captioning. Hallucination localization aims to precisely identify textual spans (words or phrases) within captions that contradict visual evidence from the corresponding video. By enabling span-level detection, our approach provides accurate feedback to users by marking only erroneous segments, thus preserving correct information. This fine-grained localization not only enhances caption reliability but also provides valuable guidance for model improvement and potentially increases interpretability.

We construct this dataset by generating captions using multiple state-of-the-art VideoLLMs on videos selected from existing datasets such as MSR-VTT (Xu et al., 2016) and FAVD-Bench (Shen et al., 2023), and manually annotating each hallucinated span. The resulting dataset comprises 1,167 video-caption pairs, each containing at least one hallucinated segment. Additionally, we propose a VideoLLM-based baseline model for hallucination localization. This baseline utilizes instruction-tuned VideoLLMs to generate hallucinated spans as output. We conduct extensive experiments using five different VideoLLMs and evaluate their performance quantitatively and qualitatively. In summary, this paper offers three primary contributions:

- We propose hallucination localization in video captioning, enabling identification of hallucinations at word or phrase levels.
- We construct the HLVC-Dataset, enabling researchers to quantitatively evaluate models developed for hallucination localization.
- We develop and evaluate a VideoLLM-based baseline approach for hallucination localization, demonstrating its effectiveness both quantitatively and qualitatively.

2 Related Work

2.1 Video Captioning

Video captioning is a task that involves generating descriptive sentences from input videos. Early studies combined CNN-based encoders with LSTM-based decoders (Venugopalan et al., 2015; Yao et al., 2015; Pan et al., 2016). Subsequently, Transformer-based methods such as VideoBERT (Sun et al., 2019), UniVL (Luo et al., 2020), and SwinBert (Lin et al., 2022) were introduced. More recently, LLM-based approaches, such as VideoLLMs, have been applied to video captioning, enabling the generation of more accurate and fluent captions (Li et al., 2023, 2024; Zhang et al., 2023; Cheng et al., 2024; Zhang et al., 2025).

2.2 Hallucination Detection

Hallucination detection is a task that determines whether hallucinations are present within generated text. In NLP, this task has been studied in various domains such as text summarization (Kryscinski et al., 2020), machine translation (Xu et al., 2023), and dialogue systems (Dziri et al., 2021). In computer vision, considerable research has focused on detecting object hallucinations, (i.e., nonexistent or incorrectly identified objects in images) (Sun et al., 2019; Ben-Kish et al., 2024). Additionally, evaluation metrics targeting video content, such as EMScore (Shi et al., 2022) and FactVC (Liu and Wan, 2023), have been proposed and applied specifically to hallucination detection in video captions. To the best of our knowledge, no existing research has localized hallucinations at the span level within video captions.

3 Proposed Task: Hallucination Localization in Video Captioning

In this section, we introduce our proposed task, hallucination localization in video captioning. The goal of this task is to localize spans within captions that contain hallucinated content.

Table 1: **Examples of video–caption pairs from the HLVC-Dataset.** We classified hallucinations into three categories(*Entity*, *Relation*, and *Invented*) and calculated the proportion of each. Original Caption denotes the caption produced by the VideoLLMs, whereas Edited Caption refers to its corrected version.

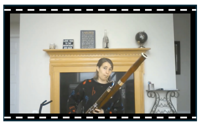
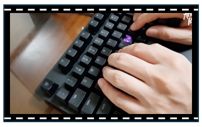
Category	**Example**			**Ratio(%)**
	Video	**Original Caption**	**Edited Caption**	
Entity		a woman playing a flute in a room.	a woman playing a bassoon in a room.	40.5
Relation		a woman in a blue dress standing in front of a camera in a newsroom.	a woman in a blue dress sitting in front of a camera in a newsroom.	23.0
Invented		a person typing on a keyboard and using a mouse.	a person typing on a keyboard.	34.6
Others		a man wearing a shirt with the word "modern".	a man wearing a shirt with the word "pioneering since 1903".	1.97

Figure 1 illustrates the hallucination localization task. The model takes a video and its caption as input and then highlights any spans in the caption that qualify as hallucinations. These spans contain information that either contradicts the video or cannot be verified from it. We can thus perform a more fine-grained analysis of hallucinations in video captioning.

3.1 Task Definition

Let the evaluation set contain M video–caption pairs. For the j-th sample ($1 \leq j \leq M$) we denote the video by $v^{(j)}$ and its caption by $\mathbf{x}^{(j)} = (x_1^{(j)}, \ldots, x_{n_j}^{(j)})$. The objective of hallucination localization is to decide, for every token $x_i^{(j)}$, whether it is grounded in the visual evidence of $v^{(j)}$. We model a hallucination localization system as a function $f(\mathbf{x}, v)$ and write

$$\hat{\mathbf{y}}^{(j)} = f(\mathbf{x}^{(j)}, v^{(j)}) = (\hat{y}_1^{(j)}, \ldots, \hat{y}_{n_j}^{(j)}).$$

Here, each predicted token label is

$$\hat{y}_i^{(j)} = \begin{cases} 1 & \text{if } x_i^{(j)} \text{ is hallucinated,} \\ 0 & \text{otherwise.} \end{cases}$$

Contiguous indices with $\hat{y}_i^{(j)} = 1$ constitute hallucination spans. This task requires fine-grained language and vision alignment along with precise error tagging.

3.2 Evaluation Metrics

For each sample we assume an oracle label sequence $\mathbf{y}^{(j)} = (y_1^{(j)}, \ldots, y_{n_j}^{(j)})$. Following the "exact" and "partial" span-matching criteria popularised in Named Entity Recognition (NER) (Segura-Bedmar et al.), we define three complementary metrics:

Strict Matching Accuracy (SMA). A sample is correct *iff* the predicted and oracle sequences are identical, $\hat{\mathbf{y}}^{(j)} = \mathbf{y}^{(j)}$. The corpus-level score is defined as:

$$\text{SMA} = \frac{1}{M} \sum_{j=1}^{M} \mathbf{1}\left(\hat{\mathbf{y}}^{(j)} = \mathbf{y}^{(j)}\right).$$

Partial Matching Accuracy (PMA). Mirroring the NER "partial match" setting, a sample is counted as correct if the system identifies at least one hallucinated token. Here, n_j denotes the length of the j-th caption:

$$\text{PMA} = \frac{1}{M} \sum_{j=1}^{M} \mathbf{1}\left(\sum_{i=1}^{n_j} \hat{y}_i^{(j)} y_i^{(j)} > 0\right).$$

Edit Distance (ED). To measure overall fidelity we average the Levenshtein distance between predicted and oracle label sequences:

$$\text{ED} = \frac{1}{M} \sum_{j=1}^{M} D_{\text{lev}}\big(\hat{\mathbf{y}}^{(j)}, \mathbf{y}^{(j)}\big).$$

Table 2: **Statistics of HLVC-Dataset.** We report statistics on the presence or absence of hallucinations for each model.

Model	hallucinated?		Total
	No	Yes	
VideoLLaMA	1410	590	2000
VideoChat	1423	577	2000
Total	2833	1167	4000

While SMA emphasizes strict matching, PMA allows for a more relaxed evaluation. For instance, when multiple hallucinated spans exist within a caption, SMA marks the prediction as incorrect unless all spans are correctly identified, whereas PMA marks it as correct if even one span is identified. Together, these metrics capture both fine-grained and coarse-grained localization performance. PMA, however, does not penalize over-detecting hallucinations. To fill this gap, we also report ED, which offers a balanced metric by penalizing both overpredictions and hallucination misses.

4 Dataset: HLVC-Dataset

In this section, we present HLVC-Dataset, a new benchmark expressly designed for Hallucination Localization in Video Captioning (HLVC). In contrast to existing datasets (Shi et al., 2022) that only indicate hallucinated spans, our dataset also provides corrective annotations explaining how each error should be corrected. Table 1 shows sample entries from the HLVC-Dataset. For each video–caption pair, the dataset also supplies the hallucinated span(s) and the caption after editing. These annotations make the dataset suitable for a broad range of studies.

4.1 Video dataset selection

We collected videos from existing video datasets. We used MSR-VTT (Xu et al., 2016) and FAVD-Bench (Shen et al., 2023) as our sources. MSR-VTT is one of the most widely used corpora for video captioning research and includes diverse, open-domain footage. FAVD-Bench is a video dataset designed for tasks that take audio information into account and offers high audiovisual diversity. We extracted 1,000 clips from each dataset, gathering 2,000 videos in total.

4.2 Video caption generation

Video captions are automatically generated using existing VideoLLMs. We select VideoLLaMA (Zhang et al., 2023) and VideoChat (Li et al., 2023), the most recent models available when our annotation began. By providing each video together with the prompt "Describe this video in one sentence." to the VideoLLMs, we obtain its caption. Applying both VideoLLMs to the 2,000 videos yields 4,000 video–caption pairs in total.

4.3 Annotation Protocol

We annotate hallucinations in video captions through a three-stage workflow.

Caption-level decision. We first determine whether the caption contains *any* hallucination. A binary label is assigned: 1 if at least one hallucination is present, and 0 otherwise. Hallucinations fall into two categories:

- *Invented* — content that cannot be confirmed from the footage (e.g., '*in the school*' when the setting is not discernible);
- *Contradictory* — content that clearly conflicts with the visual evidence (e.g., '*a man*' when the person in the video is a girl).

Span-level marking. For captions labelled 1, each hallucinated span is wrapped in numbered tags `<tagn>...</tagn>`, thereby capturing both the count and precise location of hallucinations. As illustrated in Figure 1, after applying span-level marking, the caption becomes: 'The video shows a <tag1>man</tag1> is playing the drums <tag2>in the school</tag2>.'

Editing. Every tagged span is minimally edited—either by substitution or deletion. Continuing the same example, if the footage simply shows a person drumming in an unspecified room, the corrected caption becomes 'The video shows a <tag1>girl</tag1> is playing the drums <tag2></tag2>.'.

4.4 Annotation Procedure

Because prior studies report low inter-annotator agreement in generic crowdsourcing environments (Shi et al., 2022), we contract a professional annotation firm to perform the labeling. Beyond supplying detailed annotation guidelines, we provide direct instruction to ensure quality control. We first carry out a pilot annotation on a small subset of the data; after verifying satisfactory performance, we scale the process to the full dataset.

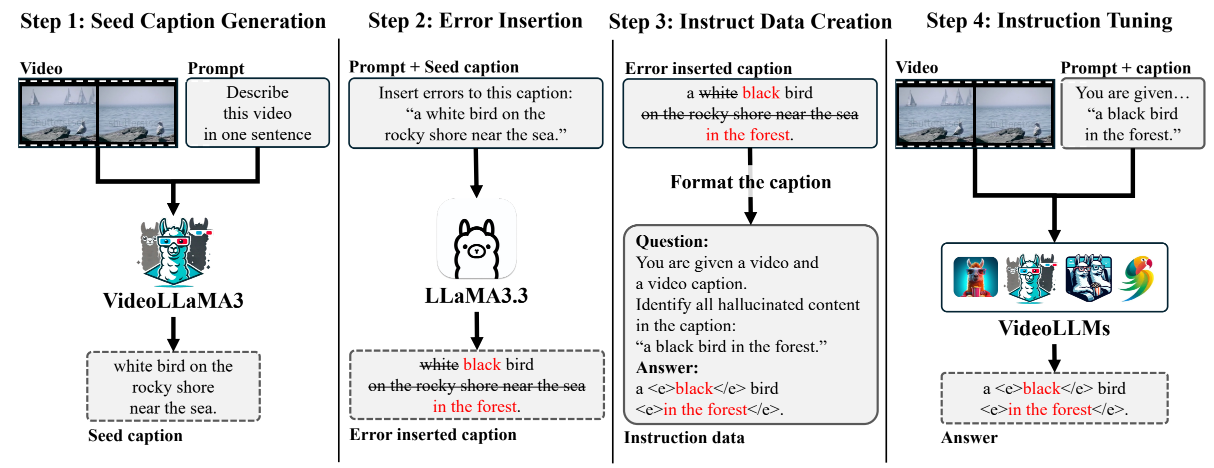


Figure 2: **Overview of our method.** The procedure is as follows: Step 1 generates seed captions using an existing VideoLLM, which are assumed to be free from hallucinations. Step 2 automatically inserts errors into these seed captions using an LLM (LLaMA3.3). Step 3 formats the error-inserted captions as instruction data for VideoLLMs. Step 4 performs instruction tuning on VideoLLMs specifically for hallucination localization, enabling the tuned model to output hallucinated spans in the input video captions.

4.5 Statistics on HLVC-Dataset

Table 2 summarizes the incidence of hallucinations in the 4,000 video–caption pairs that constitute HLVC-Dataset. Overall, 1,167 captions (29.2%) were judged to contain at least one hallucinated span. When comparing the two models, VideoLLaMA produced 590 hallucinated captions out of 2,000 (29.5%), while VideoChat produced 577 out of 2,000 (28.9%). The difference between these two models is modest, at only 0.6 percentage points, suggesting that the VideoLLMs tested here showed a similar level of susceptibility to hallucination in this one-sentence description task.

We grouped each hallucination into one of four mutually exclusive categories: *Entity* (incorrect nouns or noun phrases), *Relation* (incorrect verbs, prepositions, adjectives, or other relation-bearing expressions), *Invented* (information unrelated to the visual content or not verifiably grounded in the video), and *Others* (all remaining cases). If a span exhibited several error types, we assigned it to the single category that most saliently drove the misinterpretation; when no clear primary type could be determined, we defaulted to *Others*. As summarized in Table 1, Entity errors were the most prevalent, accounting for 40.5% of all hallucinations. Invented errors followed at 34.6%, indicating a strong tendency of both VideoLLMs to hallucinate content wholly absent from the video. Relation errors made up 23.0%, reflecting incorrect descriptions of actions, spatial relations, or attributes, while Others constituted only 1.97%.

5 Method

This section presents our baseline for hallucination localization. The approach first prepares an instruction dataset designed to identify hallucinated content and then instruction-tunes VideoLLMs on this data. After tuning, the VideoLLMs can localize hallucinated spans in video captions. Manually authoring such an instruction set is prohibitively expensive. Therefore, following FAVA (Mishra et al., 2024), we prepare instruction data based on an LLM-driven synthetic-error scheme.

5.1 Seed Caption Generation

The objective of this step is to produce a large corpus of video–caption pairs. We sample 500 000 videos from WebVid2M (Bain et al., 2021), a large-scale video dataset. For each video, we prompt VideoLLaMA3 (Zhang et al., 2025) with "Describe this video in one sentence." to generate a corresponding caption. We then compute the video–caption similarity scores with LanguageBind (Zhu et al., 2023), retain the 10 000 highest-scoring pairs, and use them as seed data for instruction tuning.

5.2 Error Insertion

The objective of this step is to generate video captions that deliberately contain hallucinations. Drawing on the FAVA framework, we implement a procedure that uses an LLM to inject synthetic errors. We define three error categories (Entity, Relation, Invented) following the definitions provided in Sec-

Table 3: **Hallucination localization performance on HLVC-Dataset.** The evaluation metrics are Strict Matching Accuracy (SMA), Partial Matching Accuracy (PMA), and Edit Distance (ED). N/A indicates that evaluation was not possible due to the model's output not conforming to the required format.

Model	Vision encoder	LLM	SMA↑	PMA↑	ED↓
Zero-Shot					
VideoChat (Li et al., 2023)	BLIP2	Vicuna-7b	N/A	N/A	N/A
VideoChat2 (Li et al., 2024)	UMT-L	Mistral-7b	1.8	27.2	4.62
VideoLLaMA (Zhang et al., 2023)	BLIP2	LLaMA2-7b	N/A	N/A	N/A
VideoLLaMA2.1 (Cheng et al., 2024)	SigLIP	Qwen2-7b	1.6	3.9	3.25
VideoLLaMA3 (Zhang et al., 2025)	SigLIP	Qwen2.5-7b	2.9	2.6	3.32
Ours (Instruction Tuning)					
VideoChat (Li et al., 2023)	BLIP2	Vicuna-7b	2.6	16.4	3.40
VideoChat2 (Li et al., 2024)	UMT-L	Mistral-7b	10.0	34.3	3.04
VideoLLaMA (Zhang et al., 2023)	BLIP2	LLaMA2-7b	8.5	35.0	3.14
VideoLLaMA2.1 (Cheng et al., 2024)	SigLIP	Qwen2-7b	13.2	42.1	2.92
VideoLLaMA3 (Zhang et al., 2025)	SigLIP	Qwen2.5-7b	**20.5**	**54.9**	**2.64**

tion 4.5. The exact prompts used for each error type are provided in the Appendix A.3. Each error type is inserted into the seed caption with a fixed insertion probability, and we ensure that the different errors do not interfere with one another. In our experiments, we employ LLaMA3.3-70b (Grattafiori et al., 2024) for error injection and set the insertion probability for each error type to 0.5.

5.3 Instruct Data Creation

We convert captions containing inserted errors into an instruction-based format. Specifically, we create instruction data by concatenating an erroneous caption with the following prompt: "You are given a video and a video caption. Identify all hallucinated content in the caption:". The expected output of this instruction data is the original caption, with inserted errors enclosed within the tags . If a caption contains no errors, the instruction data output remains unchanged from the original caption. An example of instruction data is provided below:

Question
You are given a video and a video caption. Identify all hallucinated content in the caption. Surround each hallucinated word or phrase with <e>...</e>. Video Caption: "This video shows black bird in the forest."
Answer
This video shows <e>black</e> bird <e>in the forest</e>.

5.4 Instruction Tuning

We perform instruction tuning on a pretrained VideoLLMs using the created instruction data. After instruction tuning, the model localizes hallucinated content within video captions by enclosing hallucinated spans with tags. In our experiments, we utilize five VideoLLMs: VideoChat (Li et al., 2023), VideoChat2 (Li et al., 2024), VideoLLaMA (Zhang et al., 2023), VideoLLaMA2 (Cheng et al., 2024), and VideoLLaMA3 (Zhang et al., 2025). The instruction data format and tuning parameters generally follow default settings. Additionally, all models employ a unified LLM architecture of 7 billion parameters.

6 Experiments

6.1 Hallucination Localization

To evaluate the hallucination localization capability of the models, we extract a test set comprising 500 video-caption pairs containing hallucinations from the HLVC-Dataset. We perform evaluations under both zero-shot and instruction-tuned model, employing the same prompt format for consistency across comparisons.

Table 3 summarizes the performance of the models under both evaluation settings. In the zero-shot scenario, models demonstrate significant limitations, frequently failing to output the correct format, and, even when they succeed, they yield relatively low performance. Conversely, our method considerably improves performance, highlighting its

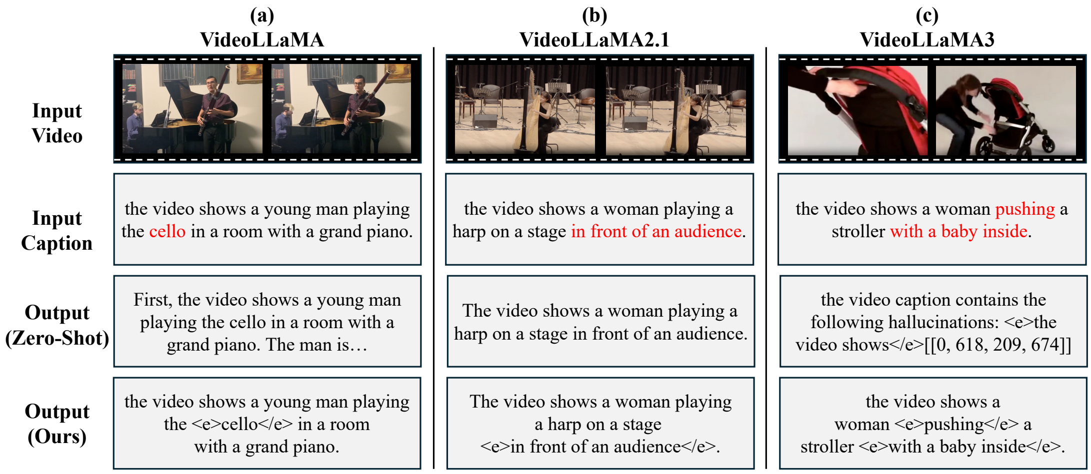


Figure 3: **Qualitative evaluation of hallucination localization.** The first column lists the input video, the second the input caption, the third the model output in the zero-shot setting, and the fourth the model output produced with our instruction-tuned method. The spans highlighted in red within the input caption indicate hallucinated spans.

Table 4: Hallucination detection performance on HLVC-Dataset.

Model	Accuracy	F1 score
Zero-Shot		
VideoChat	N/A	N/A
VideoChat2	49.2	47.2
VideoLLaMA	N/A	N/A
VideoLLaMA2.1	51.3	14.7
VideoLLaMA3	51.0	10.8
Ours (Instruction Tuning)		
VideoChat	50.5	46.2
VideoChat2	60.7	55.5
VideoLLaMA	60.0	60.4
VideoLLaMA2.1	65.1	65.2
VideoLLaMA3	**66.7**	**69.5**

effectiveness in enhancing hallucination localization. We observe a general correlation between the models' overall performance and their performance on hallucination localization. Notably, VideoLLaMA3 successfully localizes approximately 20% of the hallucinated spans. Additionally, there are instances where VideoChat and VideoLLaMA correctly identify their own generated hallucinations, indicating that the instruction tuning process enhances models' introspective capabilities.

6.2 Hallucination Detection

Our HLVC-Dataset can also be applied to hallucination detection. Therefore, we conducted additional evaluations specifically targeting hallucination detection performance. To evaluate the hallucination detection capabilities of the models, we construct a comprehensive test set of 1000 video-caption pairs. This dataset comprises 500 pairs containing hallucinations, identical to the set used in the localization task, along with 500 additional pairs without any hallucinations. Predictions are determined based on the presence or absence of error tags in the model outputs. Specifically, if the output contains tags, the model is judged to predict hallucinations; if tags are absent or the format is incorrect, the prediction defaults to no hallucinations. We use Accuracy and F1 Score as metrics.

Table 4 presents the hallucination detection results under zero-shot and instruction tuning conditions. The zero-shot performance is notably limited, approaching random chance levels. Instruction tuning significantly enhances performance, reflecting improved model awareness and detection capability. There is a general correlation between hallucination detection performance and overall model performance as observed in the localization task. However, even the best-performing model, VideoLLaMA3, achieves an accuracy of only 66.7%, highlighting the inherent difficulty of this dataset.

6.3 Qualitative Evaluation

Figure 3 shows qualitative evaluations of hallucination localization results. It compares the zero-shot outputs with those of our instruction-tuned method for VideoLLaMA, VideoLLaMA2, and VideoLLaMA3. (a) shows the results for VideoLLaMA. Although the video depicts a man playing a "bassoon," the caption incorrectly describes him playing a cello. In the zero-shot scenario, the instruc-

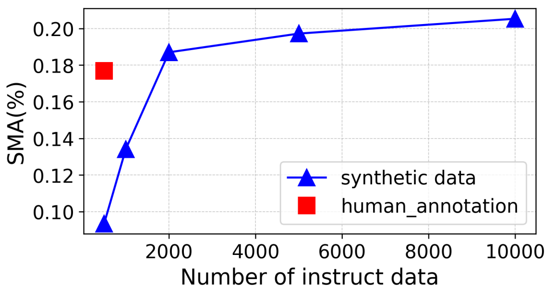


Figure 4: Comparison of human annotation and synthetic data in instruction data.

Table 5: Ablation study on error categories in instruction data.

Error category			Evaluation metrics		
Entity	Relation	Invented	SMA↑	PMA↑	ED↓
✓			13.4	33.3	2.98
	✓		3.1	20.9	3.23
		✓	7.7	17.9	2.92
✓	✓	✓	**13.6**	**41.7**	**2.77**

tion is ignored, and the model merely describes the video content. Conversely, our method accurately localizes the hallucination span. This example demonstrates that VideoLLaMA can identify its own hallucinations, indicating the potential for self-correction in captions. (b) illustrates an example of an "invented" hallucination. The zero-shot approach fails to pinpoint the hallucination accurately, whereas our method successfully localizes it. This demonstrates the model's ability to detect extraneous information generated by VideoLLMs. (c) provides an example of a caption containing multiple hallucinations. In the zero-shot scenario, the output is in an incorrect format, but our method correctly localizes each hallucination span. This indicates that our model can handle not only simple cases such as individual word errors but also more complex hallucinations.

6.4 Ablation Study

Error categories in instruction data. To analyze the impact of different error categories included in instruction data, we perform an ablation study focusing on three specific error categories: Entity, Relation, and Invented. Each setting maintains a dataset size of 5000 samples, systematically varying the presence or absence of these error types. This experiment is conducted using only VideoLLaMA3. Table 5 illustrates the results of this ablation study. The Entity error type alone significantly outperforms Relation and Invented errors individually. The superior performance for Entity likely arises from the prevalence of object-related hallucinations in video datasets and the relative ease of correcting such errors. Conversely, the poorer performance for Relation errors is presumably due to the highly diverse variations within this category, including verbs, prepositions, and adjectives. Combining all error categories yields the highest overall performance, indicating the importance of capturing realistic and diverse error scenarios. This suggests that expanding the diversity of generated errors could enhance model effectiveness.

Human-annotation vs synthetic data. Figure 4 shows how SMA changes when using human-annotated instruction data versus synthetic data on VideoLLaMA3. At 500 samples, human annotation yields 17.68 % SMA, while synthetic data only reaches 9.35 %. As we increase simulated data size, SMA steadily climbs, exceeding 17.68% at around 2,000 samples and reaching 20.53% at 10,000 samples. This simple trend demonstrates that, beyond a modest sample threshold, large-scale simulation can outperform manual annotation in accuracy. Combining a small human-annotated set with additional synthetic examples may therefore offer an efficient path to high-quality instruction data.

7 Conclusion

In this paper, we introduced a novel task, hallucination localization in video captioning, which specifically identifies spans within captions containing content not grounded in the corresponding visual evidence. To facilitate research on this task, we created the HLVC-Dataset, a carefully annotated dataset consisting of 1,167 video-caption pairs with marked hallucination spans. Additionally, we proposed an instruction-tuned VideoLLM baseline designed to accurately predict hallucinated spans. Our experiments demonstrated that instruction tuning significantly enhances the ability of VideoLLMs to localize hallucinations, achieving substantial improvements over zero-shot approaches. We further provided extensive qualitative and quantitative evaluations, illustrating our method's effectiveness and highlighting areas for potential improvement. Future work should incorporate a broader range of error scenarios and enhance overall model reliability in video captioning applications.

8 Limitation

Limitations of VideoLLM-Based Hallucination Localization In our proposed method, we instruction-tune VideoLLMs to perform hallucination localization. This allows the model to identify hallucinated spans in the video caption. However, one of the main limitations of this approach lies in the assumption that the model's output format will always match the expected structure. In practice, VideoLLMs may generate responses that deviate from the intended format, which can hinder the accurate detection of hallucinations. To address this issue, future work may consider incorporating mechanisms such as format-aware loss functions or additional constraints to guide the model toward producing consistently structured outputs.

Scope Limitation: Excluding Hallucination Editing The HLVC-Dataset could also serve as a resource for studying hallucination editing, where hallucinated content is replaced with accurate, video-grounded information. However, we have intentionally left that task outside the scope of this paper. Editing is far harder than localization because it requires first detecting the errors and then producing authoritative corrections tied to the video. Evaluation is also difficult, as multiple reasonable fixes may exist and there is often no single "right" answer. For these reasons, we limit our work to hallucination localization and present it as a practical first step toward improving the reliability of VideoLLMs.

References

Moloud Abdar, Meenakshi Kollati, Swaraja Kuraparthi, Farhad Pourpanah, Daniel McDuff, Mohammad Ghavamzadeh, Shuicheng Yan, Abduallah Mohamed, Abbas Khosravi, Erik Cambria, and 1 others. 2024. A review of deep learning for video captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*.

Assaf Ben-Kish, Moran Yanuka, Morris Alper, Raja Giryes, and Hadar Averbuch-Elor. 2024. Mitigating open-vocabulary caption hallucinations. In *Proceedings of the Empirical Methods in Natural Language Processing*.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and 1 others. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

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

Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proceedings of the Empirical Methods in Natural Language Processing*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the Empirical Methods in Natural Language Processing*.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. Swinbert: End-to-end transformers with sparse attention for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Hui Liu and Xiaojun Wan. 2023. Models see hallucinations: Evaluating the factuality in video captioning. In *Proceedings of the Empirical Methods in Natural Language Processing*.

Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. 2020. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*.

Fan Ma, Xiaojie Jin, Heng Wang, Yuchen Xian, Jiashi Feng, and Yi Yang. 2024. Vista-llama: Reducing hallucination in video language models via equal distance to visual tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Neelu Madan, Andreas Møgelmose, Rajat Modi, Yogesh S Rawat, and Thomas B Moeslund. 2024. Foundation models for video understanding: A survey. *Authorea Preprints*.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. *arXiv preprint arXiv:2401.06855*.

Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang. 2016. Hierarchical recurrent neural encoder for video representation with application to captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*.

Xuyang Shen, Dong Li, Jinxing Zhou, Zhen Qin, Bowen He, Xiaodong Han, Aixuan Li, Yuchao Dai, Lingpeng Kong, Meng Wang, and 1 others. 2023. Fine-grained audible video description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yaya Shi, Xu Yang, Haiyang Xu, Chunfeng Yuan, Bing Li, Weiming Hu, and Zheng-Jun Zha. 2022. Emscore: Evaluating video captioning via coarse-grained and fine-grained embedding matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, and 1 others. 2025. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*.

Nasib Ullah and Partha Pratim Mohanta. 2022. Thinking hallucination for video captioning. In *Proceedings of the Asian Conference on Computer Vision*.

Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. 2015. Translating videos to natural language using deep recurrent neural networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Weijia Xu, Sweta Agrawal, Eleftheria Briakou, Marianna J Martindale, and Marine Carpuat. 2023. Understanding and detecting hallucinations in neural machine translation via model introspection. *Transactions of the Association for Computational Linguistics*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE International Conference on Computer Vision*.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, and 1 others. 2025. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the Empirical Methods in Natural Language Processing: System Demonstrations*.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, and 1 others. 2023. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *International Conference on Learning Representations*.

Figure 5: Part of the actual explanatory materials that were submitted to the annotation company.

A Appendix

A.1 Details of Instruct Tuning

We prepared 500 validation samples separately from the test data and used them to select the model weights for evaluation. During training, we adopted the model weights from the iteration that achieved the highest averaged score of SMA and PMA on the validation set as the final weights. We did not perform any hyperparameter tuning in this experiment. We used 8 A100 80GB GPUs for all instruct tuning of VideoLLMs.

A.2 Details of Annotation Procedure

We outsourced the annotation work required to construct the HLVC-Dataset to a professional annotation company in Japan. As illustrated in Figure 5, we prepared detailed guideline materials for the annotators and supplied these to the domestic vendor; additionally, we scheduled an online session to explain the task in real time. The remuneration agreed upon prior to the annotation phase was paid in full. The agreement on data handling and the ethics review were completed at the contract time. Our annotation procedure did not entail any of the ethical concerns often raised in crowdsourced settings.

A.3 Prompts for Error Insertion

Table 6 shows the prompts used to generate synthetic errors in our baseline method. The errors are categorized into three types: Entity, Relation, and Invented. Each prompt includes instructions, definitions of the error types, the output format, and illustrative examples.

A.4 Licence

The licenses for the datasets and models used in this study are summarized below. All resources were employed strictly for research purposes only.

- MSR_VTT: Unknown (research-only use)
- FAVD_Bench: Apache 2.0
- VideoChat: MIT / base model under research-only license
- VideoChat2: MIT / base model under research-only license
- VideoLLaMA: BSD 3-Clause / base model under research-only license
- VideoLLaMA2: Apache 2.0
- VideoLLaMA3: Apache 2.0 / model weights for research-only use (non-commercial)

Category	Prompt
Entity	You are provided with a video caption extracted from a video-text dataset. Your task is to simulate a pseudo-hallucination by deliberately introducing an entity error into the caption. This means that a small portion of the caption—typically a noun or noun phrase (such as an object, person, or location)—should be replaced with an incorrect (hallucinated) entity, mimicking a realistic mistake. Follow these guidelines: - **Modify only one entity per caption.** - The alteration should target a short segment (usually 1-3 words) that forms a noun or noun phrase. - Use the following markup to annotate your change: - ‘<delete>...</delete>‘: Enclose the original text that is being removed or replaced. - ‘<mark>...</mark>‘: Enclose the new, hallucinated text that replaces the original. - ‘<entity>...</entity>‘: Wrap both ‘<delete>‘ and ‘<mark>‘ tags to indicate that the change is an entity error. - **Output only the final edited caption enclosed within ‘<s>‘ and ‘</s>‘ tags.** Do not include any additional commentary or explanation. Good Example 1: - Before: ‘<s>woman narrating a description of sauteed shrimp and penne pasta with cooking instructions.</s>‘ - After: ‘<s>woman narrating a description of <entity><delete>sauteed shrimp</delete><mark>glazed strawberries</mark></entity> and penne pasta with cooking instructions.</s>‘ Good Example 2: - Before: ‘<s>bunch of volleyball players in black and white jerseys playing against each other in a match.</s>‘ - After: ‘<s>bunch of <entity><delete>volleyball</delete><mark>basketball</mark></entity> players in black and white jerseys playing against each other in a match.</s>‘
Relation	You are provided with a video caption extracted from a video-text dataset that may already have error tokens inserted. Your task is to simulate a pseudo-hallucination by deliberately introducing a relation error into the caption. This means that a small portion of the caption—typically a verb or relational phrase—should be replaced with an incorrect (hallucinated) relational token, mimicking a realistic mistake. Follow these guidelines: - **Modify only one relation per caption.** - The alteration should target a short segment (usually 1-3 words) that forms a relational phrase. - Use the following markup to annotate your change: - ‘<delete>...</delete>‘: Enclose the original text that is being removed or replaced. - ‘<mark>...</mark>‘: Enclose the new, hallucinated text that replaces the original. - ‘<relation>...</relation>‘: Wrap both ‘<delete>‘ and ‘<mark>‘ tags to indicate that the change is a relation error. - **Output only the final edited caption enclosed within ‘<s>‘ and ‘</s>‘ tags.** Do not include any additional commentary or explanation. Good Example 1: - Before: ‘<s>woman narrating a description of sauteed shrimp and penne pasta with cooking instructions.</s>‘ - After: ‘<s>woman <relation><delete>narrating</delete><mark>describing</mark></relation> a description of sauteed shrimp and penne pasta with cooking instructions.</s>‘ Good Example 2: - Before: ‘<s>bunch of volleyball players in black and white jerseys playing against each other in a match.</s>‘ - After: ‘<s>bunch of volleyball players in black and white jerseys <relation><delete>playing</delete><mark>competing</mark></relation> against each other in a match.</s>‘
Invented	You are provided with a video caption extracted from a video-text dataset that may already have error tokens inserted. Your task is to simulate a pseudo-hallucination by deliberately introducing an invented information error into the caption. This means that an additional sentence or phrase containing fabricated information should be inserted into the caption, mimicking a realistic mistake. Follow these guidelines: - **Insert the invented information error outside any already existing error tokens.** - The fabricated information should be a short phrase that introduces an incorrect fact. - Use the following markup to annotate your change: - ‘<invented>...</invented>‘: Wrap the fabricated information error. - **Output only the final edited caption enclosed within ‘<s>‘ and ‘</s>‘ tags.** Do not include any additional commentary or explanation. Good Example 1: - Before: ‘<s>woman narrating a description of sauteed shrimp and penne pasta with cooking instructions.</s>‘ - After: ‘<s>woman <invented>and man</invented> narrating a description of sauteed shrimp and penne pasta with cooking instructions.</s>‘ Good Example 2: - Before: ‘<s>bunch of volleyball players in black and white jerseys playing against each other in a match.</s>‘ - After: ‘<s>bunch of volleyball players in black and white jerseys playing against each other in a match <invented>in a park</invented>.</s>‘

Table 6: Prompts for error insertion.