Can Hallucination Correction Improve Video-Language Alignment?

Lingjun Zhao[1]*, **Mingyang Xie**[1],
Paola Cascante-Bonilla[1,2], **Hal Daumé III**[1], **Kwonjoon Lee**[3]
[1]University of Maryland, College Park [2]Stony Brook University [3]Honda Research Institute
lzhao123@umd.edu

Abstract

Large Vision-Language Models often generate hallucinated content that is not grounded in its visual inputs. While prior work focuses on mitigating hallucinations, we instead explore leveraging hallucination correction as a training objective to improve video-language alignment. We introduce HACA, a self-training framework learning to correct hallucinations in descriptions that do not align with the video content. By identifying and correcting inconsistencies, HACA enhances the model's ability to align video and textual representations for spatio-temporal reasoning. Our experimental results show consistent gains in video-caption binding and text-to-video retrieval tasks, demonstrating that hallucination correction-inspired tasks serve as an effective strategy for improving vision and language alignment.

1 Introduction

Aligning representations across modalities involves creating joint embeddings that map visual and linguistic features to a shared space, enabling the model to assess their similarity. This is crucial for tasks including cross-modal retrieval (Xu et al., 2016a), mitigating hallucinations (Jiang et al., 2024), compositional reasoning (Cascante-Bonilla et al., 2024), visual-to-text generation (Li et al., 2022), visual question answering (Shen et al., 2022), and visual-language navigation (Zhao et al., 2021). While progress has been made in aligning image representations with text (Li et al., 2020; Zeng et al., 2022; Wang et al., 2023a; Lin et al., 2025), advancements in video-language alignment remain limited. Videos pose unique challenges due to their rich spatio-temporal information, involving multiple entities and scenes that dynamically interact and change over time. Video-language models (Video-LLMs in §3.1) can compute alignment scores (Lin et al., 2023; Li et al., 2023b) but struggle to distinguish between similar videos and descriptions (Park et al., 2022; Wang et al., 2023b; Saravanan et al., 2024), as illustrated in Figure 1. One promising approach is to fine-tune Video-LLMs on entailment tasks using similar captions (Bansal et al., 2024), where the model is prompted to answer `Yes` or `No` to whether a video is aligned with a given caption (§3.1). However, using a single binary label as a learning signal fails to indicate which parts of the description misalign with the video. Bansal et al. (2024) generates natural language explanations for mismatches but requires costly dataset construction with additional models and annotations.

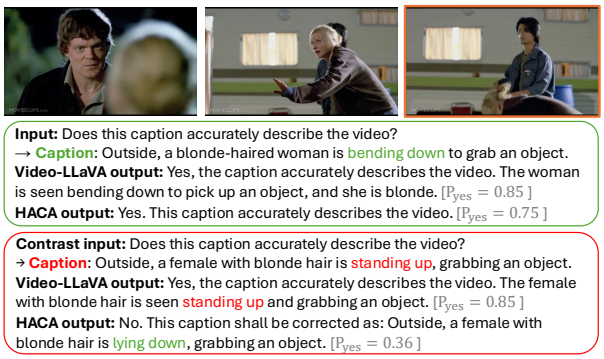


Figure 1: Models tasked with determining whether a given video entails a caption, where the contrast caption closely resembles the correct one. HACA effectively differentiates between the correct caption (top) and the incorrect one (bottom), and corrects hallucination in the latter. In contrast, Video-LLaVA fails to distinguish between those captions or correct the hallucination.

To this end, we introduce **HACA**, a self-training framework grounded in **HA**llucination **C**orrection for video-language **A**lignment (§ 3.2). *Hallucination* (or confabulation) refers to a mismatch between textual descriptions and the corresponding factual content of an image or video (Liu et al., 2024). HACA requires the model to pre-

*This work was partially conducted during an internship at Honda Research Institute.

Findings of the Association for Computational Linguistics: ACL 2025, pages 25636–25646

dict whether a description entails the video content. If the description does not align, the model corrects the hallucinations to better match the video. Instead of relying solely on a binary entailment label, HACA uses hallucination correction as a finer-grained learning signal to enhance the alignment of video and language representations. Given that misalignment between modalities is a key factor in hallucination (Biten et al., 2022; Sun et al., 2024), we hypothesize that introducing a hallucination correction task can improve video-language alignment. HACA also requires no external models or annotations beyond the ground-truth video description. To further enhance HACA, we introduce a masking correction task as data augmentation (§3.3).

We fine-tune two Video-LLMs with HACA, and evaluate these fine-tuned models in a zero-shot manner on two spatio-temporally challenging downstream tasks (§ 4): VELOCITI (Saravanan et al., 2024), a video-caption binding dataset, and SSv2-Temporal (Sevilla-Lara et al., 2021) and SSv2-Events (Bagad et al., 2023), which are text-to-video retrieval datasets emphasizing action recognition. The models fine-tuned with HACA outperform baseline models by up to 17.9% accuracy and 5.7 mAP points, demonstrating that HACA effectively improves video-text alignments, and generalizes beyond in-domain data (§ 5). Our code and data will be available upon acceptance.

2 Related Work

In addition to the video-language alignment approaches discussed in § 1, several methods leverage a contrastive learning objective to learn a shared video-language embedding space (Xue et al., 2023; Rasheed et al., 2023; Girdhar et al., 2023; Zhu et al., 2024), and Bagad et al. (2023) further introduces a contrastive loss in Video-LLMs to enforce time-order consistency. However, most of these models lack robustness to semantically plausible manipulations (Park et al., 2022). Yuksekgonul et al. (2023) also finds that applying a contrastive objective to video-caption datasets does not promote the model's ability to capture fine-grained details. In contrast, our approach encourages Video-LLMs to capture more nuanced semantic mismatches by learning to correct hallucinations, extending beyond sentence-level hallucination detection. More discussion is provided in Appendix B.

3 HACA: Hallucination Correction for Video-language Alignment

To investigate whether hallucination correction can improve video-language alignment, we introduce HACA, a fine-tuning objective for Video-LLM as a sequence-to-sequence generation task.

3.1 Preliminaries: Video-LLMs

Video-LLMs typically consist of three parts: i) a visual encoder to map images and videos to visual representations; ii) an LLM that takes text instructions as inputs to generate text responses; and iii) an adapter between visual and text representations. Our approach finetunes (ii) the text decoder and (iii) the adapter, freezing (i) the visual encoder.

Pre-training. A Video-LLM M_θ parameterized by θ takes a textual question or instruction $\boldsymbol{Q}$, a video $\boldsymbol{V}$ as input, and generates a text response $\boldsymbol{A} = (\boldsymbol{A}_1, \boldsymbol{A}_2, ..., \boldsymbol{A}_T)$ autoregressively using a decoder-based language model (LLM) as output, by estimating a conditional distribution $M(\boldsymbol{A} \mid \boldsymbol{Q}, \boldsymbol{V})$. This is achieved by training the model using the maximum-likelihood estimation (MLE) objective:

$$L(\theta) = \sum_{\mathcal{D}_{\text{train}}} \sum_{t=1}^{T} \log M_\theta(\boldsymbol{A}_t \mid \boldsymbol{A}_{<t}, \boldsymbol{Q}, \boldsymbol{V}) \quad (1)$$

where $\boldsymbol{A}_t$ is t-th word of the text response, and $\boldsymbol{A}_{<t}$ are the first $t-1$ words of the response. The dataset $\mathcal{D}_{\text{train}}$ consists of samples in the form $(\boldsymbol{Q}, \boldsymbol{A}, \boldsymbol{V})$.

Fine-tuning with entailment. Following Bansal et al. (2024), we finetune the Video-LLM using an entailment task, where the text input $\boldsymbol{Q}$ is formatted as an entailment question as $\boldsymbol{Q}(\boldsymbol{W}) =$ `Does this caption accurately describe the video? Caption: {`$\boldsymbol{W}$`}`. In this task, the output of the model $\boldsymbol{A}$ is `Yes` or `No` (Figure 2 (a)). Given a dataset D_{train} consisting of ground-truth answers $\boldsymbol{A}$ for $\boldsymbol{Q}(\boldsymbol{W})$ and $\boldsymbol{V}$, the model is fine-tuned to have a better estimation of $M_\theta(\texttt{Yes} \mid \boldsymbol{Q}(\boldsymbol{W}), \boldsymbol{V})$ and $M_\theta(\texttt{No} \mid \boldsymbol{Q}(\boldsymbol{W}), \boldsymbol{V})$ using the MLE objective:

$$L_{ent}(\theta) = \sum_{\mathcal{D}_{\text{train}}} \log M_\theta(\boldsymbol{A} \mid \boldsymbol{Q}(\boldsymbol{W}), \boldsymbol{V}) \quad (2)$$

3.2 Learning from Hallucination Correction

Building on the work of Bansal et al. (2024), the Video-LLM takes the question $\boldsymbol{Q}$ and the video $\boldsymbol{V}$ as input to determine whether a text description $\boldsymbol{W}$ entails the video (similar to § 3.1). However, in our

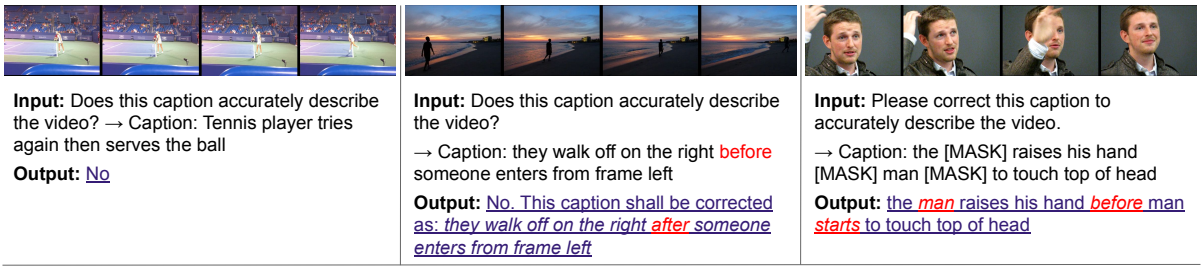


Figure 2: Example of different finetuning objectives. The first column shows an example of the baseline *entailment* task. The second column shows an example of our proposed *HACA* task, where we finetune the model to output hallucination correction to justify the response. The third column shows an example of the *masking correction* task, where we input a masked version of the video description and finetune the model to predict the corrected one.

setting, if $\boldsymbol{W}$ does not entail $\boldsymbol{V}$, the model generates a *corrected* caption $\hat{\boldsymbol{W}} = (w_1, w_2, ..., w_n)$ to align the description with the video content. During fine-tuning, if $\boldsymbol{W}$ entails $\boldsymbol{V}$, the model is trained to generate the response as $\boldsymbol{A}(\texttt{Yes}) = \texttt{Yes, the caption accurately describes the video}$. If $\boldsymbol{W}$ does not entail $\boldsymbol{V}$, the model is trained to generate a corrected description $\hat{\boldsymbol{W}}$ as its response, formatted as $\boldsymbol{A}(\texttt{No}, \hat{\boldsymbol{W}}) = \texttt{No. This caption shall be corrected as: } \{\hat{\boldsymbol{W}}\}$. We show an example in Figure 2 (b). In contrast to fine-tuning using *entailment* only, our *hallucination correction* objective trains the model to have better estimation of $M_\theta(\boldsymbol{A}(\texttt{Yes}) \mid \boldsymbol{Q}(\boldsymbol{W}), \boldsymbol{V})$ and $M_\theta(\boldsymbol{A}(\texttt{No}, \hat{\boldsymbol{W}}) | \boldsymbol{Q}(\boldsymbol{W}), \boldsymbol{V})$.

Instead of using Eq 2, given a training dataset $\mathcal{D}_{\text{train}}$ that consists of video $\boldsymbol{V}$ and ground-truth text description $\hat{\boldsymbol{W}}$ pairs, we fine-tune the Video-LLM using the MLE objective:

$$L_c(\theta) = \sum_{\mathcal{D}_{\text{train}}} \sum_{t=1}^{T} \log M_\theta(\boldsymbol{A}_t \mid \boldsymbol{A}_{<t}, \boldsymbol{Q}(\boldsymbol{W}), \boldsymbol{V}) \tag{3}$$

where $\boldsymbol{A}_t$ is the t-th word of the text response of $\boldsymbol{A}(\texttt{Yes})$ or $\boldsymbol{A}(\texttt{No}, \hat{\boldsymbol{W}})$, and $\boldsymbol{A}_{<t}$ is the first $t-1$ words of the text response.

3.3 Masking Correction as Augmentation

We also incorporate a masking correction task as data augmentation (Figure 2 (c)), where an instruction $\boldsymbol{Q}$ prompts the Video-LLM to make corrections to a masked caption $\bar{\boldsymbol{W}}$, teaching the model to generate a corrected caption that contains a sequence of words $\hat{\boldsymbol{W}} = (w_1, w_2, ..., w_n)$ as its answer by estimating conditional probability $M(\hat{\boldsymbol{W}} \mid \boldsymbol{Q}(\bar{\boldsymbol{W}}), \boldsymbol{V})$. Specifically, $\boldsymbol{Q}$ is a function that formats the text instruction as $\boldsymbol{Q}(\bar{\boldsymbol{W}}) = \texttt{Please correct this caption to accurately describe the video. Caption: } \bar{\boldsymbol{W}}$, where $\bar{\boldsymbol{W}}$ is masked from $\hat{\boldsymbol{W}}$: $\bar{\boldsymbol{W}} = (w_1, \texttt{[MASK]}, ..., w_n)$, by randomly masking 45% of the content words in the ground truth video description $\hat{\boldsymbol{W}}$.

We finetune the model using two objectives: the MLE objective to estimate the probability for masking correction $M_\theta(\hat{\boldsymbol{W}} | \boldsymbol{Q}(\bar{\boldsymbol{W}}), \boldsymbol{V})$, and the HACA objective (Eq 3). The model is tasked with providing responses corresponding to different instructions.

4 Experimental Setup

Data (detailed in §A.5). We train HACA using videos and their ground-truth and contrastive descriptions from VideoCon (Bansal et al., 2024), generating 115,536 (video, description, correction) triplets for training and 8,312 for validation, which is used for model selection. Synthetic contrast captions are also used to fine-tune the *baseline* entailment task with the same dataset sizes.

We evaluate our trained models on text-to-video retrieval using the temporally-challenging SSv2-Temporal (Sevilla-Lara et al., 2021) and action-intensive SSv2-Events (Bagad et al., 2023) datasets. Additionally, we evaluate our models on compositional ability over time using the VELOCITI benchmark (Saravanan et al., 2024). Each video in the dataset includes a correct caption and an incorrect one.

Baselines. (i) **Pretrained Video-LLMs**: we employ two pre-trained models with different architectures, *Video-LLaVA* (Lin et al., 2023) and *VideoChat2* (Li et al., 2023b). More details in §A.4. (ii) **Entailment**: we fine-tune the pretrained Video-LLMs using the entailment task described in §3.1. More details about the implementation are in §A.1.

Evaluation metrics. We report the accuracy on the VELOCITI benchmark as the proportion of examples in which the positive video-caption pair

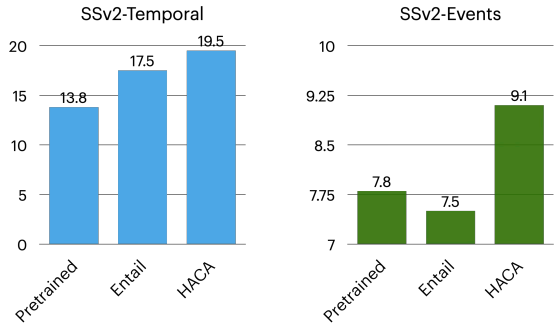


Figure 3: Mean Average Precision (mAP) scores for pre-trained Video-LLaVA and models fine-tuned using various methods on zero-shot text-to-video retrieval tasks.

Model	Agent Tests			Action Tests				Avg
	Iden	Bind	Coref	Adv	Bind	Modif	Chrono	
Random	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
Human	94.7	93.3	96.0	100.0	92.7	91.3	93.3	94.4
CLIP-ViP B/32	75.3	52.4	55.7	70.2	53.5	51.2	48.5	58.1
ViFi-CLIP B/16	82.3	58.7	54.6	63.0	59.3	60.5	49.8	61.2
mPLUG-V	43.0	31.9	51.7	65.0	42.0	49.6	41.3	46.3
PLLaVA	68.6	43.3	60.5	62.4	46.6	56.0	49.6	55.3
VideoCon	67.4	44.6	50.0	73.0	51.1	63.2	45.6	56.4
Video-LLaVA	74.1	50.4	**60.1**	63.6	47.0	47.9	56.0	57.0
+ Entail	73.7	59.7	55.5	68.4	57.3	64.0	57.3	62.3
+ HACA	80.3	**62.6**	57.9	**72.6**	**60.0**	**65.8**	54.5	64.8
+ HACA+Mask	**82.7**	62.1	57.9	71.8	59.0	64.8	**57.9**	**65.2**
VideoChat2	76.8	54.4	53.1	56.0	46.2	59.3	**54.7**	57.2
+ Entail	59.7	56.9	55.5	62.2	53.0	50.1	53.9	55.9
+ HACA	77.2	**60.4**	56.4	65.8	**55.0**	61.7	53.7	61.5
+ HACA+Mask	**79.1**	59.7	**56.9**	**68.2**	54.6	**66.9**	51.1	**62.4**

Table 1: Zero-shot accuracy on VELOCITI for models trained with the baseline entailment task, our proposed HACA objective, and other contrastive (CLIP-ViP (Xue et al., 2023), ViFi-CLIP (Rasheed et al., 2023)) and generative (mPLUG-V (Ye et al., 2023), PLLaVA (Xu et al., 2024)) models.

receives a higher `Yes` entailment probability than the corresponding negative video-caption pair. For SSv2, we compute `Yes` probabilities for each text-video pair, rank their scores, and report mean Average Precision (mAP).

5 Analysis

Performance on text-to-video retrieval. HACA consistently outperforms both the pretrained model and the entailment fine-tuned model, as illustrated in Figure 3. This demonstrates HACA's ability to effectively capture the rich temporal information present in videos. On the SSv2-Events dataset, while the entailment objective yields performance comparable to the pretrained Video-LLaVA, HACA achieves better results on this action-intensive dataset, despite being fine-tuned on the same amount of data. Additional comparisons with other models are provided in §A.2.

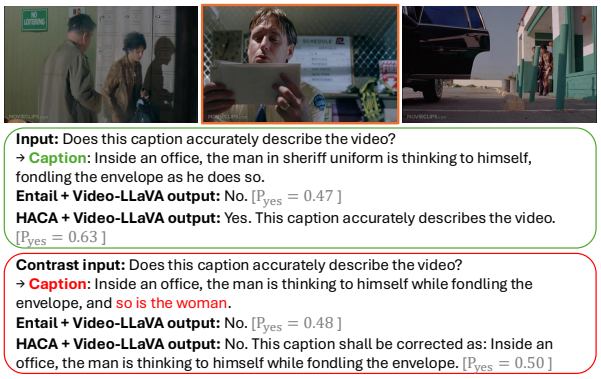


Figure 4: Success on binding and correction: HACA effectively assigns higher entailment probability P_{yes} to the correct caption (top) than the incorrect one (bottom), unlike the entailment-finetuned model. HACA also accurately corrects the incorrect caption in its output.

Performance on video-language binding. Table 1 shows that, on average, both Video-LLaVA and VideoChat2 fine-tuned with the HACA objective outperform the pre-trained models and those fine-tuned with the entailment objective. Masking correction further boosts performance through data augmentation. The *Agent Coref* test evaluates a model's ability to link events to specific agents, a misalignment type absent in the VideoCon dataset, where actions are always tied to one agent. Consequently, the pretrained Video-LLaVA outperforms its fine-tuned versions, with HACA marginally exceeding the entailment baseline. The *Chrono* test measures a model's ability to detect reversed event order. While VideoCon includes such data, our results show that models fine-tuned on the entailment objective perform similarly to the pretrained model. Although HACA slightly underperforms the entailment objective, it excels on SSv2-Events, involving multiple events.

HACA consistently outperforms baseline models in all *Action* tests: *Action Adv* (replacing an action with one not in the video), *Action Bind* (replacing an action within the same video), and *Action Modif* (replacing the manner with a plausible modifier). This highlights HACA's robust ability to distinguish actions in videos, requiring understanding of complex spatio-temporal relationships between the video and its description. HACA also excels in *Agent Iden* and *Agent Bind*, showcasing its effectiveness in identifying and binding entities through the right relationship.

Qualitative examples. Figure 4 presents an example where HACA outperforms the entailment baseline on the VELOCITI dataset, and delivers ac-

curate corrections. Additional qualitative examples are provided in §A.3.

HACA does not hinder question answering. To assess whether fine-tuning with the HACA objective affects the multi-task capabilities of Video-LLMs, we conduct a zero-shot evaluation on the MSRVTT-QA dataset (Xu et al., 2016a) using GPT-3.5-turbo. The results, presented in Table 2, are based on a subset of 7,000 samples (10% of the dataset) due to budget constraints. The GPT-evaluated score for the pretrained Video-LLaVA model aligns with previously reported values (Lin et al., 2023). As shown in the table, the HACA-finetuned model performs comparably to the pretrained Video-LLaVA model, despite not involving explicit QA-specific finetuning. In contrast, fine-tuning with the entailment objective alone leads to a significant performance drop on MSRVTT-QA. A possible explanation is that optimizing for a single entailment label may impair the language generation capabilities of video-language models.

Model	GPT Score
Video-LLaVA	3.5
+ Entail	2.8
+ HACA	3.4

Table 2: Zero-shot GPT-assessed score on MSRVTT-QA for the model trained with the baseline entailment task, and our proposed HACA objective. GPT-assessed scores ranges from 0 to 5.

6 Conclusion

Video understanding through language is vital for applications like human-robot interaction and autonomous driving. We propose a novel approach to enhance video-language alignment by connecting it to the hallucination problem in visual-language models, paving the way for future advancements.

Limitations

Our proposed method assumes the availability of ground-truth video caption annotations for fine-tuning using hallucination correction. Additionally, the method assumes a clear separation between the parameters of the video representations and those of the language model, as we freeze the video encoder parameters during fine-tuning to align video-language representations. Another limitation is that our approach has not been evaluated on long videos, due to the limitation of computational resources. We envision future work in this direction.

Acknowledgements

This material is based upon work partially supported by the NSF under Grant No. 2229885 (NSF Institute for Trustworthy AI in Law and Society, TRAILS). We thank Haoqiang Kang for guidance on optimizing the computation speed of fine-tuning video-language models.

References

Piyush Bagad, Makarand Tapaswi, and Cees GM Snoek. 2023. Test of time: Instilling video-language models with a sense of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2503–2516.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1708–1718.

Hritik Bansal, Yonatan Bitton, Idan Szpektor, Kai-Wei Chang, and Aditya Grover. 2024. Videocon: Robust video-language alignment via contrast captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13927–13937.

Ali Furkan Biten, Lluís Gómez, and Dimosthenis Karatzas. 2022. Let there be a clock on the beach: Reducing object hallucination in image captioning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1381–1390.

Paola Cascante-Bonilla, Yu Hou, Yang Trista Cao, Hal Daumé III, and Rachel Rudinger. 2024. Natural language inference improves compositionality in vision-language models. *arXiv preprint arXiv:2410.22315*.

David Dale, Elena Voita, Loïc Barrault, and Marta R Costa-jussà. 2022. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity even better. *arXiv preprint arXiv:2212.08597*.

Pradipto Das, Chenliang Xu, Richard F. Doell, and Jason J. Corso. 2013. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2634–2641.

Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. 2022. Bridging video-text retrieval with multiple choice questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16167–16176.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2018. Localizing moments in video with temporal language. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1380–1390, Brussels, Belgium. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2022. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wen Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023a. Videochat: Chat-centric video understanding. *ArXiv*, abs/2305.06355.

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

Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. 2023c. Lavender: Unifying video-language understanding as masked language modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23119–23129.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2025. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *NeurIPS*.

Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *ArXiv*, abs/2306.05424.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*.

Minesh Mathew, Dimosthenis Karatzas, R. Manmatha, and C. V. Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *WACV*.

Jae Sung Park, Sheng Shen, Ali Farhadi, Trevor Darrell, Yejin Choi, and Anna Rohrbach. 2022. Exposing the limits of video-text models through contrast sets. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3574–3586.

Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Fine-tuned clip models are efficient video

learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6545–6554.

Darshana Saravanan, Darshan Singh, Varun Gupta, Zeeshan Khan, Vineet Gandhi, and Makarand Tapaswi. 2024. Velociti: Can video-language models bind semantic concepts through time? *arXiv preprint arXiv:2406.10889*.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *ECCV*.

Laura Sevilla-Lara, Shengxin Zha, Zhicheng Yan, Vedanuj Goswami, Matt Feiszli, and Lorenzo Torresani. 2021. Only time can tell: Discovering temporal data for temporal modeling. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 535–544.

Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutz. 2022. How much can clip benefit vision-and-language tasks? In *ICLR*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2024. Aligning large multimodal models with factually augmented RLHF. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110, Bangkok, Thailand. Association for Computational Linguistics.

Vicuna Team. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90% chatgpt quality. https://vicuna.lmsys.org/.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning*, pages 23318–23340. PMLR.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2023a. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4581–4591.

Zhenhailong Wang, Ansel Blume, Sha Li, Genglin Liu, Jaemin Cho, Zineng Tang, Mohit Bansal, and Heng Ji. 2023b. Paxion: Patching action knowledge in video-language foundation models. *Advances in Neural Information Processing Systems*, 36:20729–20749.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*.

Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Hao Jiang, Fei Wu, and Linchao Zhu. 2024. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. *arXiv preprint arXiv:2404.14233*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016a. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016b. Msr-vtt: A large video description dataset for bridging video and language. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296.

Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. 2024. Pllava : Parameter-free llava extension from images to videos for video dense captioning. *ArXiv*, abs/2404.16994.

Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. 2023. CLIP-vip: Adapting pre-trained image-text model to video-language alignment. In *The Eleventh International Conference on Learning Representations*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. 2020. Clevrer: Collision events for video representation and reasoning. In *ICLR*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why vision-language models behave like bags-of-words, and what to do about it? In *ICLR*.

Yan Zeng, Xinsong Zhang, and Hang Li. 2022. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *ICML*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *ArXiv*, abs/2306.02858.

Lingjun Zhao, Khanh Xuan Nguyen, and Hal Daumé Iii. 2024. Successfully guiding humans with imperfect instructions by highlighting potential errors and suggesting corrections. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 719–736, Miami, Florida, USA. Association for Computational Linguistics.

Ming Zhao, Peter Anderson, Vihan Jain, Su Wang, Alexander Ku, Jason Baldridge, and Eugene Ie. 2021. On the evaluation of vision-and-language navigation instructions. In *EACL*.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online. Association for Computational Linguistics.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. In *The Twelfth International Conference on Learning Representations*.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Cai Wan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. 2024. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *The Twelfth International Conference on Learning Representations*.

A Appendices

A.1 Implementation Details

Fine-tuning from pretrained models. We use the visual representations for video and language model embeddings pretrained from Video-LLMs to perform instruction fine-tuning using different objectives, including HACA (§ 3.2), entailment (§ 3.1), and masking correction (§ 3.3). During finetuning, visual representations are frozen, and the embeddings from the visual-text adapter layers and LLM are learnable.

Hyperparameters and computation. For Video-LLaVA, we finetune our models for 3 epochs, using a learning rate of $2e^{-4}$ and AdamW optimizer. We also use a LoRA adapter (Hu et al., 2022) of rank 128 and alpha 256. Since we freeze the video encoder, the number of trainable parameters is significantly reduced to 241M for Video-LLaVA. The number of video frames processed per video is 8, with a batch size of 8, using 2 RTXA6000 GPUs, for a total of $\sim$ 72 hours.

For VideoChat2, we finetune our models for 3 epochs, using a learning rate of $2e^{-5}$ and AdamW optimizer. We use a LoRA adapter (Hu et al., 2022) of rank 16 and alpha 32. We also freeze the visual encoders and reduce the number of trainable parameters to 193M. The number of video frames processed per video is 8, with a batch size of 2, using 1 RTXA6000 GPU, for around $\sim$ 72 hours.

Tools. We implement our models with Pytorch 2.0.1, Huggingface Transformers 4.31.0, scikit-learn 1.2.2. We use SciPy 1.6.0 to find content words from ground truth video description by excluding words with part-of-speech tags: AUX, SYM, DET, PUNCT.

A.2 Ablation studies

Performance on text-to-video retrieval. In Table 3, HACA (§ 3.2) demonstrates competitive performance on SSv2 downstream tasks, surpassing the pretrained model by up to 5.7 mAP points and outperforming the model fine-tuned with the entailment task by up to 2.0 mAP points. Masking correction augmentation typically enhances video-language alignment when jointly trained with HACA or the entailment task.

Effect of different mask ratios. Table 4 shows the performance when jointly finetuning Video-LLaVA using HACA and masking correction task (§ 3.3) with different masking ratio. The results indicate that using masking ratio of 45% achieves higher average accuracy.

Comparing HACA and natural language explanations. To assess the effectiveness of HACA as a finetuning task, we compare it against natural language explanations (NLE) generated by external natural language inference models (Bansal et al., 2024), used alongside the entailment task. We fine-tune Video-LLaVA with both the entailment and NLE training objectives and report the results in Table 5. Our findings show that HACA outperforms Video-LLaVA trained with entailment and NLE objectives, even without our proposed masking objective.

Model	SSv2-Temporal	SSv2-Events
Random	7.3	3.3
ImageBind (Girdhar et al., 2023)	10.5	5.5
TACT (Bagad et al., 2023)	-	7.8
mPLUG-V (Ye et al., 2023)	10.9	6.8
VideoCon (Bansal et al., 2024)	15.2	11.4
Video-LLaVA	13.8	7.8
+ Entail	17.5	7.5
+ Entail+Mask	18.0	9.9
+ HACA	**19.5**	9.1
+ HACA+Mask	15.3	**10.3**

Table 3: Mean Average Precision (mAP) scores for the tested models in the zero-shot text-to-video retrieval tasks.

Model	Agent Tests			Action Tests			Chrono	Avg
	Iden	Bind	Coref	Adv	Bind	Modif		
Video-LLaVA	74.1	50.4	**60.1**	63.6	47.0	47.9	56.0	57.0
+ HACA	80.3	**62.6**	57.9	**72.6**	**60.0**	**65.8**	54.5	64.8
+ HACA+Mask 15%	77.9	58.6	58.4	71.4	57.7	61.7	57.4	63.3
+ HACA+Mask 30%	81.4	60.3	54.3	70.0	58.0	65.4	**59.5**	64.1
+ HACA+Mask 45%	**82.7**	62.1	57.9	71.8	59.0	64.8	57.9	**65.2**
+ HACA+Mask 60%	**82.7**	62.1	56.5	70.6	57.8	62.3	55.7	64.0

Table 4: Accuracy of Video-LLaVA jointly finetuned with HACA and masking correction task using different masking ratio on VELOCITI (zero-shot).

Model	Agent Tests			Action Tests			Chrono	Avg
	Iden	Bind	Coref	Adv	Bind	Modif		
Video-LLaVA	74.1	50.4	**60.1**	63.6	47.0	47.9	56.0	57.0
+ Entail + NLE	77.1	60.0	58.6	66.8	55.9	**66.3**	**57.9**	63.2
+ HACA	80.3	**62.6**	57.9	**72.6**	**60.0**	65.8	54.5	64.8
+ HACA+Mask	**82.7**	62.1	57.9	71.8	59.0	64.8	**57.9**	**65.2**

Table 5: Zero-shot accuracy on VELOCITI for models trained with the baseline entailment task, mixture of entailment and natural language explanation tasks, and our proposed HACA objective.

A.3 Additional Qualitative Analysis

Figure 5 shows additional success and failure cases of HACA and the other models we tested.

A.4 Pretrained Video-LLMs.

We use two pre-trained Video-LLMs with different model architectures.

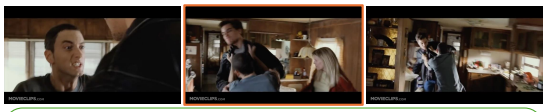
Video-LLaVA. Video-LLaVA (Lin et al., 2023) consists of LanguageBind (Zhu et al., 2024) encoders for the visual inputs, a large language model (Team, 2023), visual projection layers and a word embedding layer. It is finetuned via visual instruction tuning with 665k image-text pairs from LLaVA 1.5 (Liu et al., 2023b) and a 100k video-text instruction set from Video-ChatGPT (Maaz et al., 2023). We use this model under their Apache License 2.0.

VideoChat2. VideoChat2 (Li et al., 2023b) performs a progressive multi-modal training for three stages. In the first stage, it is trained to aling the visual encoder with a Querying Transformer (Q-Former) (Li et al., 2022) which acts as an information bottleneck between the image and textual encoders and distill relevant information to the textual context. The second stage connects the visual encoder with a pretrained LLM. In the third stage, finetunes the model via instruction tuning, using 5 different tasks including: captioning, conversations, visual question answering, reasoning and classification, with data coming from LLaVA (Liu et al., 2023b), VideoChat (Li et al., 2023a), VideoChatGPT (Maaz et al., 2023), COCO Captions (Lin et al., 2014), WebVid (Bain et al., 2021), YouCook (Das et al., 2013), OK-VQA (Marino et al., 2019), AOK-VQA (Schwenk et al., 2022), DocVQA (Mathew et al., 2021), CLEVR (Johnson et al., 2017), CLEVRER (Yi et al., 2020) and NExT-QA (Xiao et al., 2021) among others. We use this model under their MIT License.

A.5 Datasets

VideoCon. VideoCon is constructed by generating contrastive video captions and explanations for different subset of videos (Xu et al., 2016b; Wang et al., 2019; Hendricks et al., 2018). This dataset contains seven misaligned types that include replacement of objects, actions, attributes, counts and relations, and adds hallucinations (i.e. unrelated but plausible information). We use this dataset under their MIT License.

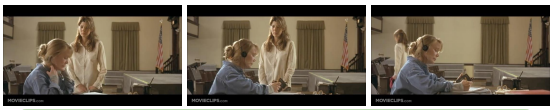
VELOCITI. The duration of the video clips in the dataset is 10 seconds, and has dense text annotations on action and role descriptions. The perception-based tests require discriminating video-caption pairs



Input: Does this caption accurately describe the video?
→ **Caption:** In the kitchen, a man in a blue shirt is pushing a man in a black jacket back angrily.
Entail + Video-LLaVA output: No. [$P_{yes} = 0.19$]
HACA + Video-LLaVA output: Yes. This caption accurately describes the video. [$P_{yes} = 0.71$]

Contrast input: Does this caption accurately describe the video?
→ **Caption:** In the kitchen, the man in a blue shirt gently pushes the man in a black jacket back.
Entail + Video-LLaVA output: No. [$P_{yes} = 0.21$]
HACA + Video-LLaVA output: Yes. This caption accurately describes the video. [$P_{yes} = 0.64$]

(a) Success on binding, failure on correction: HACA successfully assigns a higher entailment probability (P_{yes}) to the correct caption (top) compared to the incorrect one (bottom), outperforming the entailment-finetuned model in this regard. However, HACA fails to produce a correction, as it erroneously indicates that the incorrect caption accurately describes the video.



Input: Does this caption accurately describe the video?
→ **Caption:** First, At a table, the lady in blue is slowly covering her ears. Then, By a table, the person in white is slowly walking away.
Video-LLaVA output: Yes, the caption accurately describes the video. The lady in blue is covering her ears, and the person in white is walking away slowly. [$P_{yes} = 0.75$]
HACA + Video-LLaVA output: Yes. This caption accurately describes the video. [$P_{yes} = 0.67$]

Contrast input: Does this caption accurately describe the video?
→ **Caption:** First, By a table, the person in white is slowly walking away. Then, At a table, the lady in blue is slowly covering her ears.
Video-LLaVA output: Yes, the caption accurately describes the video. [$P_{yes} = 0.81$]
HACA + Video-LLaVA output: Yes. This caption accurately describes the video. [$P_{yes} = 0.76$]

(b) Failure on binding and correction: both HACA and the pre-trained Video-LLaVA model incorrectly assign a higher entailment probability (P_{yes}) to the incorrect caption (bottom) than to the correct caption (top). Additionally, HACA fails to provide a correction, mistakenly asserting that the incorrect caption accurately describes the video.

Figure 5: Some successful and failure cases of HACA and the other models on the VELOCITI dataset. The red color in text indicates the incorrect text description.

that share similar entities, and the binding tests require models to associate the correct entity to a given situation while ignoring the different yet plausible entities that also appear in the same video. There are 1000 tests using 643 videos for Agent Iden, 1676 tests using 707 videos for Agent Bind, 418 tests using 270 videos for Agent Coref, 500 tests using 400 videos for Action Adv, 1625 tests using 590 videos for Action Bind, 500 tests using 411 videos for Action Mod, and 1908 tests using 669 videos for Chrono. We use this dataset under their Creative Commons Public Licenses.

SSv2-Temporal and SSv2-Events. SSv2-Temporal contains a list of 18 actions that require models to capture rich temporal information in the video, consisting of 216 (18 ×12) candidate videos for every text action query. SSv2-Events has 49 actions that consist two verbs in the action templates that are indicative of multiple events in the video, consisting of 2888 (49×12) candidate videos for every text action query.

B Related Work

Alignment in Video-Language Models is fundamental for the logical integration of video and textual information. To align both modalities, prior work has focused on pre-training models with different objectives to capture the temporal dynamics in video. While these self-supervised correction objectives are highly effective during pre-training (Li et al., 2023c; Wang et al., 2022; Zhu et al., 2024; Ge et al., 2022), fine-tuning is typically required to adapt Video-LLMs to specific downstream tasks (Li et al., 2023a; Zhang et al., 2023; Bansal et al., 2024) (e.g., classification, retrieval, or question answering). However, these objectives rely on coarse-grained alignment labels and do not provide detailed feedback for resolving inconsistencies between video and language.

Hallucination Correction methods aim to mitigate the generation of content that does not align with the data a model was trained on, or the model describes content that does not exist in the provided input (Huang et al., 2024). Orthogonal to our proposed method, LURE (Zhou et al., 2024) uses statistical analysis to identify and rectify errors in generated descriptions, addressing co-occurrence, uncertainty, and positional factors via masking. In our work, we randomly mask the video description so that the model is required to output the corrected sentence, which is also conditioned in the input video via visual entailment. Yin et al. (2023); Wang et al. (2023b) uses external models and measures to correct hallucinations to be consistent with images or videos. Zhou et al. (2021); Liu et al. (2023a); Xiao et al. (2024); Zhao et al. (2024) create a synthetic dataset to train a specialized model to detect and correct hallucinations. Dale et al. (2022); Huang et al. (2024) shows promising results on correcting hallucinations without an external model for machine translation and image captioning. In our work, we investigate leveraging hallucination as a training objective to improve video-language alignment, by exploring the potential of using a video-LLM model itself to correct hallucinations through fine-tuning on a synthetic dataset.