Hallucination Reduction in Video-Language Models via Hierarchical Multimodal Consistency

Jisheng Dang[1,2,3] , **Shengjun Deng**[4] , **Haochen Chang**[2] , **Teng Wang**[5] , **Bimei Wang**[6,3*] , **Shude Wang**[7] , **Nannan Zhu**[2] , **Guo Niu**[4] , **Jingwen Zhao**[2] and **Jizhao Liu**[1]

[1] Lanzhou University, Gansu, China
[2] Sun Yat-sen University, Guangdong, China
[3] National University of Singapore, Singapore
[4] Foshan University, Guangdong, China
[5] The University of Hong Kong, China
[6] Jinan University, Guangdong, China
[7] Lanzhou Institute of Technology, Gansu, China
dangjsh@mail2.sysu.edu.cn, wangbm@stu2021.jnu.edu.cn

Abstract

The rapid advancement of large language models (LLMs) has led to the widespread adoption of video-language models (VLMs) across various domains. However, VLMs are often hindered by their limited semantic discrimination capability, exacerbated by the limited diversity and biased sample distribution of most video-language datasets. This limitation results in a biased understanding of the semantics between visual concepts, leading to hallucinations. To address this challenge, we propose a Multi-level Multimodal Alignment (MMA) framework that leverages a text encoder and semantic discriminative loss to achieve multi-level alignment. This enables the model to capture both low-level and high-level semantic relationships, thereby reducing hallucinations. By incorporating language-level alignment into the training process, our approach ensures stronger semantic consistency between video and textual modalities. Furthermore, we introduce a two-stage progressive training strategy that exploits larger and more diverse datasets to enhance semantic alignment and better capture general semantic relationships between visual and textual modalities. Our comprehensive experiments demonstrate that the proposed MMA method significantly mitigates hallucinations and achieves state-of-the-art performance across multiple video-language tasks, establishing a new benchmark in the field.

1 Introduction

In recent years, the rapid development of large language models (LLMs) has significantly driven the progress of video-language models (VLMs). By integrating visual encoders with LLMs, VLMs can handle various complex multimodal tasks such as video captioning [Li *et al.*, 2023a], visual question answering [Yang *et al.*, 2022], video editing [Dang *et al.*, 2024c; Dang *et al.*, 2023a; Dang *et al.*, 2024a; Dang *et al.*, 2024b], image or video classification [Li *et al.*, 2024; Ma *et al.*, 2024b; Meng *et al.*, 2024; Wang *et al.*, 2024; Meng *et al.*, 2025]. They combine advanced vision and language technologies to provide more accurate and context-aware interpretations of video content. However, video-language models encounter a significant challenge during their development: Hallucination. Hallucination occurs when the content generated by the model does not reflect the actual information in the video, leading to fabricated or incorrect descriptions.

Current research efforts aim to address hallucinations in vision-language models from various perspectives. Some approaches utilize cleaned datasets for instruction tuning to enhance inference performance [Ma *et al.*, 2024a], though this is resource-intensive. Other methods design cross-modal modules to bridge the modality gap, such as learnable interfaces [Liu *et al.*, 2024] and Q-Former [Li *et al.*, 2023a]. Additionally, some models directly correct hallucinations in model outputs through post-processing [Yin *et al.*, 2024] or improved decoding strategies [Leng *et al.*, 2024].

However, many of these methods overlook the critical importance of semantic alignment between modalities. Most visual encoders are trained primarily with language generation loss, producing descriptive visual captions of salient objects, actions, and scenes. This approach limits the discriminative power of visual representations, as they do not learn negative instances of concepts. We identify two core issues of hallucinations: (a) Insufficient semantic discrimination capabilities, which impede accurate comprehension of semantic relationships among visual concepts. (b) Downstream datasets are often small, with limited diversity and uneven sample distribution, leading models to favor predictions of common objects and their co-occurrences.

To address these challenges, we propose a multi-level mul-

[*] Corresponding author.

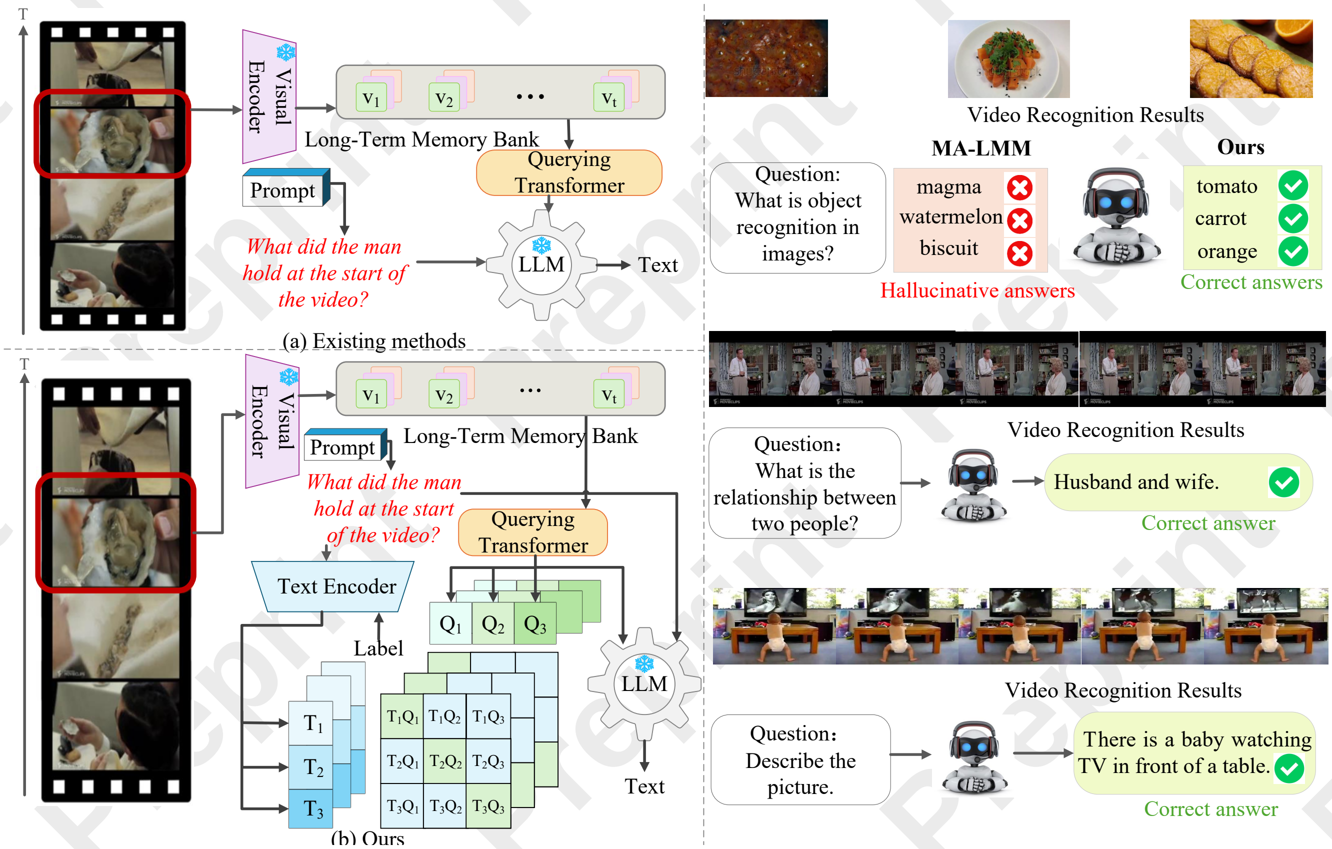
T
Visual Encoder
v_1 v_2 ··· v_t
Long-Term Memory Bank
Prompt
Querying Transformer
What did the man hold at the start of the video?
LLM
Text
(a) Existing methods
T
Visual Encoder
v_1 v_2 ··· v_t
Prompt
Long-Term Memory Bank
What did the man hold at the start of the video?
Querying Transformer
Text Encoder
Label
Q_1 Q_2 Q_3
T_1 T_2 T_3
T_1Q_1 T_1Q_2 T_1Q_3
T_2Q_1 T_2Q_2 T_2Q_3
T_3Q_1 T_3Q_2 T_3Q_3
LLM
Text
(b) Ours
Video Recognition Results
MA-LMM
Ours
Question: What is object recognition in images?
magma
watermelon
biscuit
Hallucinative answers
tomato
carrot
orange
Correct answers
Video Recognition Results
Question: What is the relationship between two people?
Husband and wife.
Correct answer
Video Recognition Results
Question: Describe the picture.
There is a baby watching TV in front of a table.
Correct answer

Figure 1: **(Top-left)** Existing methods utilize visual features that are not semantically aligned for text decoding. **(Bottom-left)** Our MMA employs a multi-level multimodal semantic alignment strategy to mitigate the hallucinations. **(Right)** Our method effectively identifies confusable objects, showcasing its capability to grasp complex semantics. It significantly outperforms MA-LMM in reducing hallucinations and enhancing answer accuracy.

timodal alignment (MMA) strategy to enhance intermediate visual features, combined with final language supervision, guiding the model to generate more accurate and contextually aligned outputs. Specifically, we employ a text encoder to convert text inputs into semantic features, facilitating alignment between visual and textual modalities through a semantic discriminative loss. Our approach goes beyond simple global representations by performing multi-level alignment, aligning semantic features at various levels of the visual and textual modalities. This strategy enables the model to capture both high-level and low-level semantic relationships, reducing hallucinations by establishing precise correspondences between video content and generated language.

To further enhance semantic alignment, we introduce a two-stage progressive training strategy. We leverage larger and more diverse datasets to expand the variety of semantic features and better capture general semantic relationships between visual and textual modalities. By integrating richer semantic information into the model and refining the alignment process, we significantly reduce ambiguity and improve performance across various video-language tasks.

Extensive experimentation demonstrates that our method consistently outperforms existing models in reducing hallucinations, improving multimodal alignment, and achieving superior overall performance across multiple video-language tasks. Our results suggest that the proposed approach effectively mitigates hallucinations in large video-language models, laying a foundation for more reliable and accurate multimodal systems.

We summarize our main contributions as follows:

- We propose a novel multi-level multimodal alignment strategy that incorporates textual semantic supervision during visual encoding. This approach aligns semantic features from both text and vision at multiple levels to address hallucinations in large video-language models.
- We propose a two-stage training strategy that facilitates progressive co-learning from general vision-text semantics to task-specific semantics, utilizing a larger and more diverse dataset.
- We conducted extensive experiments on the LVU and MSVD datasets to compare our methods with other VLMs. Our results show that our approach achieves state-of-the-art performance across various downstream video tasks, significantly improving the quality and reliability of video language models while effectively reducing hallucinations.

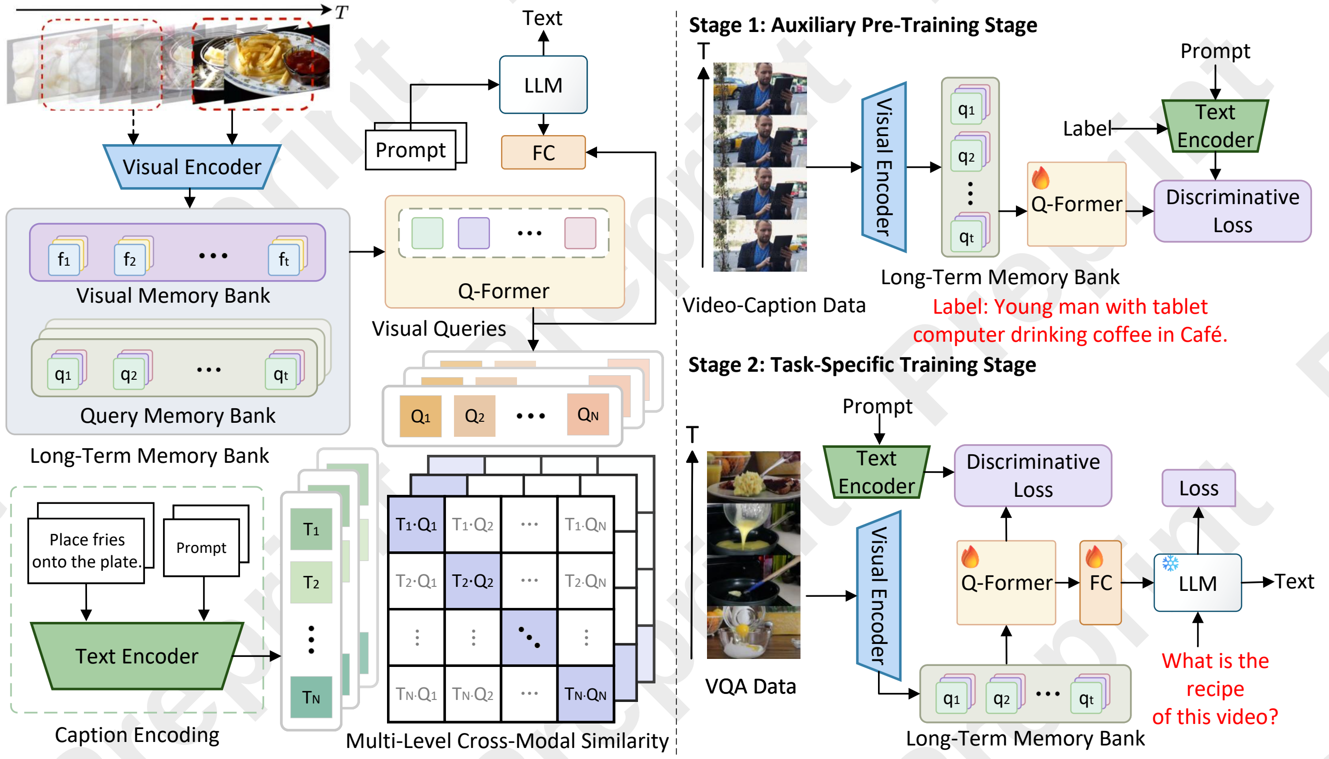

Figure 2: **(Left) Framework overview.** Framework overview. Long-Term Memory Bank and Q-Former are employed to encode visual features. Text encoder extracts joint semantic features from the prompt and text label, achieving multi-level semantic alignment with visual features through contrastive learning. The LLM then generates text outputs for various downstream tasks in video understanding. **(Right) Two-stage training strategy.** In the auxiliary pre-training stage, semantic discriminant loss is utilized on a larger and more diverse video-language dataset. In the task-specific training stage, both semantic discriminant loss and text decoding loss are applied to train on downstream task datasets.

2 Related Work

2.1 Advancements in Long-Term Video Understanding

Recent advancements in long-term video understanding have been driven by multimodal LLMs (MLLMs), memory-augmented architectures, and task-specific methods. Nevertheless, handling long-duration videos remains challenging due to computational inefficiency, temporal dependencies, and redundant information.

The integration of vision and language models has played a key role in this progress. Early models like BLIP-2 [Li *et al.*, 2023a] combined pre-trained vision and language encoders, enabling rich cross-modal reasoning for tasks such as video captioning and visual question answering. Building on this foundation, models like Video-ChatGPT [Maaz *et al.*, 2023] and Video-LLaMA [Zhang *et al.*, 2023] incorporate video transformers to better capture temporal dynamics. However, they still face limitations with extended video sequences due to fixed-size token compression, which leads to loss of critical semantics. TimeChat [Ren *et al.*, 2024] addresses this issue by introducing dynamic token compression, adjusting the compression rate according to video length. This improves temporal event localization and enhances modeling of complex temporal relationships, advancing the capabilities of multimodal video models.

2.2 Memory-Augmented Architectures and Scalability Challenges

Memory-augmented architectures have become a key strategy for long-term video understanding by retaining and referencing past frames to maintain temporal coherence. Models like LongVLM [Weng *et al.*, 2025] balance short- and long-term memory to reduce redundancy in extended videos, though they often struggle to preserve fine-grained visual details under limited computational budgets. Hierarchical models such as MeMViT [Wu *et al.*, 2022] improve attention mechanisms for long-form tasks but often underutilize token-level representation, which remains essential for effective encoding.

Meanwhile, task-specific methods have advanced computational efficiency across various video understanding tasks [Dang *et al.*, 2023b; Dang and Yang, 2022; Dang and Yang, 2021]. Retrieval-augmented generation (RAG) integrates external knowledge with generative models to reduce cost, while approaches like STTS [Bertasius *et al.*, 2021] improve efficiency through early selection and merging of informative tokens. Despite these gains, capturing both local and global temporal dependencies remains a central challenge for comprehensive long-duration video understanding.

2.3 Strategies for Mitigating Hallucinations in Video-Language Models

Hallucination represents a significant challenge in VLMs, significantly limiting their applicability in real-world scenarios. Researchers have proposed a range of correction strategies, broadly categorized into dataset dehallucination, addressing modalities gap, and output correction. However, hallucination mitigation continues to be a significant obstacle, especially in tasks requiring complex multimodal reasoning.

Dataset-level approaches, such as Text Shearing and CIT [Hu *et al.*, 2023], aim to mitigate hallucinations by improving data quality, reducing overconfidence, and disrupting spurious co-occurrences in training data. Modality-level methods, including visual fusion and perceptual reinforcement, enhance cross-modal alignment and reduce semantic misalignment. Output correction strategies, such as post-generation refinement (Woodpecker [Yin *et al.*, 2024]), focus on detecting and correcting hallucinations in model outputs. Despite promising results, these methods often come with high computational costs or rely on extensive curated datasets, limiting scalability. In this work, we propose a novel framework that explicitly integrates language-level alignment and an enhanced training scheme to address hallucinations, achieving stronger semantic consistency and improved robustness across video-language tasks.

3 Method

3.1 Overview

We propose a novel video-language model with multi-level multimodal alignment to mitigate hallucination in long-video understanding tasks. The overall architecture of our method is depicted in Figure 2. During the visual feature extraction stage, the multi-level semantic discriminative loss aligns the video embedding space with the language embedding space through contrastive learning, enabling semantic injection in the visual encoding process. The aligned visual features are then input into the LLM for text decoding. During training, we adopt a two-stage training strategy. In the auxiliary pre-training stage, we use a dataset rich in visual semantics to pre-train the Q-Former, aiming to learn cross-modal general semantics. In the task-specific training stage, with the frozen LLM, we conduct more precise semantic learning on task-specific datasets, thus alleviating the hallucination phenomenon.

3.2 Multi-Level Multimodal Alignment

Visual Feature Extraction. To effectively capture the temporal dynamics of long videos and aggregate the historical information of videos, similar to MA-LMM [He *et al.*, 2024], we obtain video frames sequentially in an autoregressive manner and store the video features in a long-term memory bank. Additionally, we utilize the Querying Transformer [Li *et al.*, 2023a; He *et al.*, 2024] to initially align the visual and text features. Specifically, given a sequence of T video frames, we first input each video frame into a pre-trained visual encoder to extract the corresponding visual features $V = [v_1, v_2, \ldots, v_T], v_t \in \mathbb{R}^{P\times C}$, where P represents the number of patches per frame and C is the channel dimension of the frame features. Subsequently, a position embedding layer is utilized to inject temporal ordering information into the frame-level features, which are then stored in the long-term memory bank to update the visual memory features $F_t = \text{Concat}[f_1, f_2, \ldots, f_t], F_t \in \mathbb{R}^{tP\times C}$ and the query memory features $Z_t = \text{Concat}[z_1, z_2, \ldots, z_t], Z_t \in \mathbb{R}^{tN\times C}$. Here, f_t represents the frame-level feature, t is the current time step, and N is the number of learnable tokens. We employ the Q-Former to interact the video memory features with the learnable Queries, aiming to learn the textual representations of the visual features. In most of the existing video-language models, the alignment of semantics between modalities is merely attempted by relying on the frozen large language model, while the semantic guidance and supervision of natural language during visual encoding are neglected. This results in deviations in the visual encoding process, making it impossible to accurately transform the rich content in the video into feature representations that highly match the textual semantics. When faced with videos containing complex scenes or subtle actions, the model may misinterpret a certain visual concept in the video as another visual concept it has encountered before. Therefore, during subsequent tasks, the model generates inaccurate or irrelevant text descriptions, which leads to the phenomenon of hallucination. We aim to introduce text semantic supervision during the visual encoding process to enhance the semantic discrimination ability of visual encoding.

Inspired by CLIP [Radford *et al.*, 2021], we introduce the bert embedding layer [Devlin, 2018] as the text encoder E_t to extract text semantic features. We concatenate the text prompt and text label and then feed them into the text encoder to obtain the semantic embedding features f_t.

$$f_t = E_t\left([< prompt > + < label >]\right). \tag{1}$$

To achieve the semantic alignment of visual and text modal features, we utilize the semantic discriminative loss to inject text semantics during the training process of the visual Q-Former. Specifically, in a batch of N (video, text) pairs, the positive sample pairs are the N matching video-text pairs, and the negative sample pairs are the $N^2 - N$ non-matching video-text pairs in the batch. The semantic discriminative loss maximizes the cosine similarity of the video and text embeddings of all positive sample pairs in the batch while minimizing the cosine similarity of the embeddings of all negative sample pairs. Essentially, semantic discriminative loss follows the form of InfoNCE loss and learns a common semantic embedding space for the visual and text modalities, which can be expressed as:

$$\mathcal{L}_{semantic} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(f_i^v \cdot f_i^t/\tau\right)}{\sum_{j=1}^{N}\exp\left(f_i^v \cdot f_j^t/\tau\right)}, \tag{2}$$

where f_i^v represents the video semantic features output by Q-Former, and f_i^t represents the text semantic features output by the text encoder. N is the batch size, and τ is the temperature coefficient. During the training process, we update the parameters of the Q-Former and the text encoder, while keeping the parameters of the visual encoder and the LLM frozen.

Injecting text semantic supervision offers several potential advantages: (1) It improves the learning process of visual encoding in Q-Former, enhancing its semantic discriminability. (2) As a standardized representation, text semantic features exhibit a stronger consistency with visual features, which effectively promotes the interaction of cross-modal information.

Multi-Level Alignment. Semantic alignment at different levels allows for the capture of text and visual information across varying degrees of abstraction. In long-term video understanding tasks, lower levels align basic visual features such as the clothing colors of characters and the geometric shapes of objects with the corresponding color and shape vocabulary in the text. At higher levels, complex visual semantics, like event flow and character relationships, can be matched with more abstract concepts in the text related to story development and character interactions.

To further enhance the semantic discriminability of visual encoding, we extend the semantic discriminative loss to a multi-level framework for improved semantic alignment. The multi-level semantic loss can be expressed as:

$$\begin{aligned}\mathcal{L}_{multi} &= \sum_{l=1}^{L} \mathcal{L}_{semantic}^{l} \\ &= -\frac{1}{N}\sum_{l=1}^{L}\sum_{i=1}^{N} \log \frac{\exp\left(f_i^v \cdot f_i^t/\tau\right)}{\sum_{j=1}^{N}\exp\left(f_i^v \cdot f_j^t/\tau\right)},\end{aligned} \tag{3}$$

where L represents the number of levels. In the experiment, we adopted a two-level alignment scheme of aligning the output features of Q-Former and aligning the input features of the cross-attention mechanism. Finally, we achieve semantic alignment of features across all levels of visual and textual modalities using a multi-level semantic discriminative loss. This enables our model to capture both high-level and low-level semantic relations while effectively reducing hallucinations by establishing precise correspondences between video content and generated language.

3.3 Two-Stage Training

The training dataset is a significant factor contributing to hallucinations in video-language models. On one hand, the lack of diversity in some datasets results in the model having an inadequate understanding of certain visual concepts, complicating the alignment between video and text modalities. On the other hand, the uneven distribution of objects in the training set causes the video-language model to favor predicting common objects or frequently co-occurring combinations of objects. Therefore, we propose a two-stage training scheme that utilizes the extended dataset to improve the optimization of the multi-level semantic discriminative loss. These two stages are the auxiliary pre-training stage and the task-specific training stage respectively.

Auxiliary Pre-Training Stage. In this stage, we utilize a larger amount of data to infuse richer semantics into the training of the video-language model. Specifically, we initially conduct pre-training on the WebVid dataset. Through this process, the model can learn the general semantics between the visual and language modalities. This serves as an auxiliary for the training of the video-language model in specific tasks. The WebVid-5K dataset contains a vast variety of video clips with corresponding textual descriptions. By exposing the model to this extensive and diverse data source, it can capture a wide range of semantic relationships that exist in the real world. This helps the model to generalize better and build a more solid foundation for subsequent task-specific training.

Task-Specific Training Stage. Once the pre-training process of the auxiliary pre-training stage is completed, we proceed with further training on other datasets to achieve semantic alignment for specific task. This two-step approach is designed to leverage the knowledge acquired during the initial pre-training phase and fine-tune the video-language model according to the requirements of the specific task. This progressive learning process from large-scale general semantics to task-specific semantics allows the video-language model to continuously refine its semantic understanding. It gradually narrows down its focus from the broad semantic space learned during pre-training to the specific semantic domain of the target task. Through this iterative process of learning and adaptation, the model can capture more accurate cross-modal semantic relationships, which in turn leads to enhanced performance in generating high-quality outputs for the specific task.

3.4 Training Objectives

We input the output features of the Q-Former, which contains all sequential historical information at the final time step, into the LLM for text decoding. During training, given a labeled dataset consisting of video and text pairs, our model is supervised using the standard cross-entropy loss:

$$\mathcal{L}_{text} = -\frac{1}{S}\sum_{i=1}^{S} \log P(w_i | w_{<i}, V), \tag{4}$$

where V represents the input video and w_i is the i-th ground-truth text token. In the auxiliary pre-training stage, we only use the semantic discriminative loss to train the Q-Former and text encoder, and do not use the LLM for text decoding. However, in the task-specific training stage, we carry out the training using two loss functions simultaneously. The overall loss function can be expressed as

$$\mathcal{L} = \lambda_{multi} \cdot \mathcal{L}_{multi} + \lambda_{text} \cdot \mathcal{L}_{text}, \tag{5}$$

where λ_{multi} and λ_{text} are hyper-parameters to trade off the two parts.

4 Experiments

4.1 Dataset

Experiments are conducted on two widely used long-term video datasets: The **LVU** dataset [Wu and Krahenbuhl, 2021] consists of more than 30,000 video clips, each ranging from 1 to 3 minutes, sourced from approximately 3,000 movies in diverse real-world contexts. The **MSVD** dataset [Chen and

Model	Content			Metadata				Avg
	Relation	Speak	Scene	Director	Genre	Writer	Year	
Obj_T4mer [Wu and Krahenbuhl, 2021]	54.8	33.2	52.9	47.7	52.7	36.3	37.8	45.0
Performer [Choromanski *et al.*, 2020]	50.0	38.8	60.5	58.9	49.5	48.2	41.3	49.6
Orthoformer [Patrick *et al.*, 2021]	50.0	38.3	66.3	55.1	55.8	47.0	43.4	50.8
VideoBERT [Sun *et al.*, 2019]	52.8	37.9	54.9	47.3	51.9	38.5	36.1	45.6
LST [Islam and Bertasius, 2022]	52.5	37.3	62.8	56.1	52.7	42.3	39.2	49.0
VIS4mer [Islam and Bertasius, 2022]	57.1	40.8	67.4	62.6	54.7	48.8	44.8	53.7
MA-LMM [He *et al.*, 2024]	58.2	**44.8**	80.3	74.6	61.0	70.4	51.9	63.0
MMA (Ours)	**62.6**	41.9	**83.0**	**79.9**	**62.1**	**70.6**	**54.9**	**65.0**

Table 1: Comparison with state-of-the-art methods on long-term video understanding task using the LVU dataset.

Dolan, 2011] contains approximately 120K sentences summarizing more than 2,000 video snippets, collected via Mechanical Turk in the summer of 2010.

4.2 Implementation Details

For the visual encoder, we adopt the pre-trained image encoder ViT-G/14 [Alexey, 2020] from EVA-CLIP [Fang *et al.*, 2023], which can also be replaced with other CLIP-based video encoders. We utilize the pre-trained Q-Former weights provided by InstructBLIP [Dai *et al.*, 2023] and use the collected WebVid-5k [Bain *et al.*, 2021] dataset for the first stage of training. Additionally, we employ Vicuna-7B [Chiang *et al.*, 2023] as the large language model (LLM). The hyperparameters of the loss function in our method are set as: $\lambda_{multi} = 0.5$ and $\lambda_{text} = 1$. All experiments are conducted on 4 V100 GPUs.

4.3 Quantitative Results

Long-Term Video Understanding. We compared our method with the state-of-the-art approaches previously reported on the LVU benchmark, and the results are shown in Table 1. It is noteworthy that our method outperforms existing long-term video models (MA-LMM [He *et al.*, 2024], ViS4mer [Islam and Bertasius, 2022], VideoBERT [Sun *et al.*, 2019], and Object Transformer [Wu and Krahenbuhl, 2021]) in both content understanding and metadata prediction tasks. This result in significant improvements in most tasks, with an increase in the average top-1 accuracy by 2.0% compared to the MA-LMM model. The result demonstrates the superior long-term video understanding capability of our method. Unlike previous models, our method performs semantic alignment of video content at multiple levels, enabling the model to achieve a more precise understanding of the questions posed in the video.

Video Question Answering. To compare with existing multimodal video understanding methods, we conducted experiments on the MSVD [Chen and Dolan, 2011] video question answering (VQA) datasets included in Table 2, to validate that using prompts for multi-level alignment enables the model to better understand its task and describe objects in videos.We observed that the introduction of multi-level alignment brings about an enhancement in performance, confirming its role in strengthening the model's ability to align semantics. Conducting two-stage training based on multi-level alignment enables the model to learn more concrete semantic information, reducing model hallucination issues. It is worth noting that our results also surpass the recent MA-LMM on these datasets, highlighting the significant improvement our model provides in video question answering.

Model	MSVD
FrozenBiLM [Yang *et al.*, 2022]	54.8
GiT [Wang *et al.*, 2022]	56.8
mPLUG-2 [Xu *et al.*, 2023]	58.1
UMT-L [Li *et al.*, 2023b]	55.2
VideoCoCa [Yan *et al.*, 2022]	56.9
Video-LLaMA [Zhang *et al.*, 2023]	58.3
MA-LMM [He *et al.*, 2024]	60.6
MMA (Ours)	**60.9**

Table 2: Comparison with state-of-the-art methods on the video question answering task using MSVD dataset.

Ablation Study. To evaluate the effectiveness of the multi-level multimodal alignment strategy and the two-stage training strategy, we conduct comprehensive ablation experiments on the LVU relation dataset. The experimental results are shown in Figure 3.

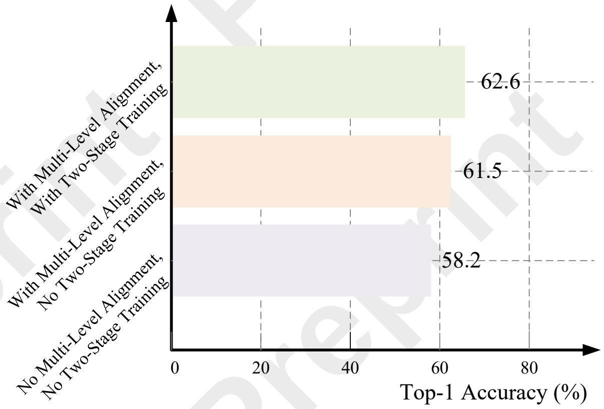

Figure 3: Ablation results of effect of multi-level multimodal alignment and two-stage training on the LVU Relation.

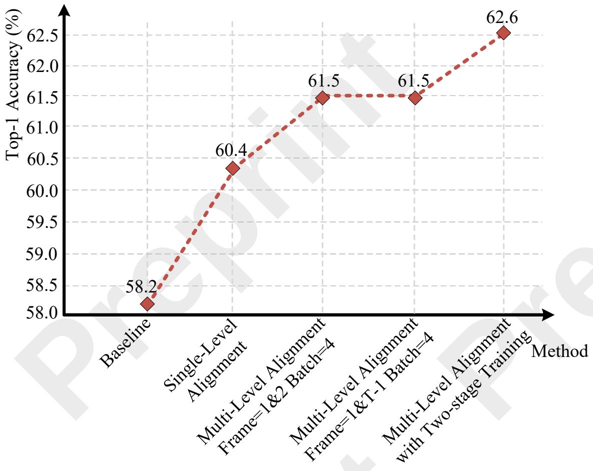

Figure 4: Ablation experiment of multi-level alignment module.

The results show that employing a multi-level alignment strategy based on the baseline yields performance improvements of 3.3%. This indicates that multi-level multimodal alignment effectively integrates textual semantics into the training of the video language model, enhancing the training process of Q-Former and significantly improving recognition accuracy in long video understanding tasks.

The results in Figure 3 also show that applying the combination of the multi-level multimodal alignment strategy and the two-stage training approach, applied on top of the baseline, results in performance improvements of 4.4% on the LVU relation dataset. Furthermore, the experimental accuracy achieved with this combined approach surpasses that of using only the multi-level multimodal alignment strategy. This demonstrates the effectiveness of the two-stage training approach, which enables VLMs to learn richer semantics.

Figure 4 shows the results regarding the effects of various alignment strategies and sampling frames. When performing single-level alignment on only using the first three frames, the performance improved compared to the baseline. We use multi-level alignment, achieving a score of 61.5%, which further validates the effectiveness of the alignment strategy. We attempted to modify the frame sampling method by testing on the first two frames and the first and last frames, respectively, and the results showed no change in top-1 performance. Additionally, using more frames in multi-level alignment does not bring any performance gains. Multi-level alignment improved model performance compared to single-level alignment, demonstrating the superiority of the multi-level alignment method. Finally, with the addition of two-stage training, the performance improved again, indicating that the alignment of additional data enables the model to better mitigate hallucination phenomena.

4.4 Visualization

Figure 1 (right) shows the comparison results between our method and the baseline method MA-LMM on the long-term video understanding task.

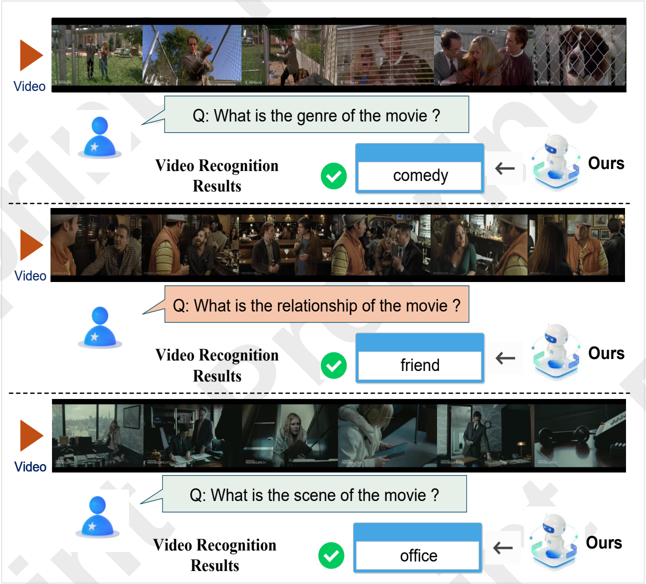
Video
Q: What is the genre of the movie ?
Video Recognition Results
comedy
Ours
Video
Q: What is the relationship of the movie ?
Video Recognition Results
friend
Ours
Video
Q: What is the scene of the movie ?
Video Recognition Results
office
Ours

Figure 5: Visualization results of our method on long-term video recognition task on LVU dataset.

Figure 5 shows the qualitative results of our method on the long-term video understanding task of the LVU dataset. In the "comedy" genre judgment (top), the model accurately determined the movie genre based on the video content, demonstrating its understanding of visual elements such as the plot and its capacity to align with semantic concepts. In the "friend" relationship recognition (middle), the model successfully inferred the relationship between characters, showcasing its ability to effectively capture and analyze visual information, including character interactions. In the "office" scene recognition (bottom), the model correctly identified the scene, illustrating its proficiency in analyzing and classifying visual elements like the video background and accurately outputting semantic information. These three examples collectively demonstrate that our model effectively captures complex semantic information while minimizing the occurrence of hallucinations.

5 Conclusions

In this work, we present a novel framework that directly addresses the challenge of mitigating hallucinations in large video-language models. By incorporating language-level supervision and alignment during training, our approach enhances semantic consistency between video and text modalities, effectively reducing the impact of noisy or misaligned data. The use of an expanded dataset and improved semantic discrimination loss further strengthens cross-modal alignment by introducing more diverse and semantically rich representations. Experimental results across various video-language tasks show that our method not only significantly reduces hallucinations but also achieves state-of-the-art performance, setting a new benchmark for future research. This work paves the way for more accurate and robust video-language understanding, with broad applications in video analysis, multimodal learning, and beyond.

References

[Alexey, 2020] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020.

[Bain *et al.*, 2021] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1728–1738, 2021.

[Bertasius *et al.*, 2021] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

[Chen and Dolan, 2011] David L. Chen and William B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, page 190–200, 2011.

[Chiang *et al.*, 2023] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

[Choromanski *et al.*, 2020] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

[Dai *et al.*, 2023] Wenliang Dai, Junnan Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. *arXiv preprint arXiv:2305.06500*, 2, 2023.

[Dang and Yang, 2021] Jisheng Dang and Jun Yang. Higcnn: Hierarchical interleaved group convolutional neural networks for point clouds analysis. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2825–2829. IEEE, 2021.

[Dang and Yang, 2022] Jisheng Dang and Jun Yang. Lhphgcnn: Lightweight hierarchical parallel heterogeneous group convolutional neural networks for point cloud scene prediction. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):18903–18915, 2022.

[Dang *et al.*, 2023a] Jisheng Dang, Huicheng Zheng, Jinming Lai, Xu Yan, and Yulan Guo. Efficient and robust video object segmentation through isogenous memory sampling and frame relation mining. *IEEE Transactions on Image Processing*, 32:3924–3938, 2023.

[Dang *et al.*, 2023b] Jisheng Dang, Huicheng Zheng, Xiaohao Xu, and Yulan Guo. Unified spatio-temporal dynamic routing for efficient video object segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 25(5):4512–4526, 2023.

[Dang *et al.*, 2024a] Jisheng Dang, Huicheng Zheng, Bimei Wang, Longguang Wang, and Yulan Guo. Temporo-spatial parallel sparse memory networks for efficient video object segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[Dang *et al.*, 2024b] Jisheng Dang, Huicheng Zheng, Xiaohao Xu, Longguang Wang, and Yulan Guo. Beyond appearance: Multi-frame spatio-temporal context memory networks for efficient and robust video object segmentation. *IEEE Transactions on Image Processing*, 2024.

[Dang *et al.*, 2024c] Jisheng Dang, Huicheng Zheng, Xiaohao Xu, Longguang Wang, Qingyong Hu, and Yulan Guo. Adaptive sparse memory networks for efficient and robust video object segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[Devlin, 2018] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Fang *et al.*, 2023] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19358–19369, 2023.

[He *et al.*, 2024] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13504–13514, 2024.

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

[Islam and Bertasius, 2022] Md Mohaiminul Islam and Gedas Bertasius. Long movie clip classification with state-space video models. In *European Conference on Computer Vision*, pages 87–104. Springer, 2022.

[Leng *et al.*, 2024] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13872–13882, 2024.

[Li *et al.*, 2023a] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742. PMLR, 2023.

[Li *et al.*, 2023b] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19948–19960, 2023.

[Li *et al.*, 2024] Xiangxian Li, Yuze Zheng, Haokai Ma, Zhuang Qi, Xiangxu Meng, and Lei Meng. Cross-modal learning using privileged information for long-tailed image classification. *Computational Visual Media*, 10(5):981–992, 2024.

[Liu *et al.*, 2024] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306, 2024.

[Ma *et al.*, 2024a] Fan Ma, Xiaojie Jin, Heng Wang, Yuchen Xian, Jiashi Feng, and Yi Yang. Vista-llama: Reducing hallucination in video language models via equal distance to visual tokens. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13151–13160, 2024.

[Ma *et al.*, 2024b] Haokai Ma, Ruobing Xie, Lei Meng, Xin Chen, Xu Zhang, Leyu Lin, and Jie Zhou. Triple sequence learning for cross-domain recommendation. *ACM Transactions on Information Systems*, 42(4):1–29, 2024.

[Maaz *et al.*, 2023] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

[Meng *et al.*, 2024] Lei Meng, Zhuang Qi, Lei Wu, Xiaoyu Du, Zhaochuan Li, Lizhen Cui, and Xiangxu Meng. Improving global generalization and local personalization for federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[Meng *et al.*, 2025] Lei Meng, Xiangxian Li, Xiaoshuo Yan, Haokai Ma, Zhuang Qi, Wei Wu, and Xiangxu Meng. Causal inference over visual-semantic-aligned graph for image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19449–19457, 2025.

[Patrick *et al.*, 2021] Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and Joao F Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. *Advances in Neural Information Processing Systems*, 34:12493–12506, 2021.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[Ren *et al.*, 2024] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14313–14323, 2024.

[Sun *et al.*, 2019] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7464–7473, 2019.

[Wang *et al.*, 2022] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022.

[Wang *et al.*, 2024] Yuqing Wang, Lei Meng, Haokai Ma, Yuqing Wang, Haibei Huang, and Xiangxu Meng. Modeling event-level causal representation for video classification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3936–3944, 2024.

[Weng *et al.*, 2025] Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. In *European Conference on Computer Vision*, pages 453–470. Springer, 2025.

[Wu and Krahenbuhl, 2021] Chao-Yuan Wu and Philipp Krahenbuhl. Towards long-form video understanding. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1884–1894, 2021.

[Wu *et al.*, 2022] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13587–13597, 2022.

[Xu *et al.*, 2023] Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, et al. mplug-2: a modularized multimodal foundation model across text, image and video. In *Proceedings of the 40th International Conference on Machine Learning*, pages 38728–38748. PMLR, 2023.

[Yan *et al.*, 2022] Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. Videococa: Video-text modeling with zero-shot transfer from contrastive captioners. *arXiv preprint arXiv:2212.04979*, 2022.

[Yang *et al.*, 2022] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *Advances in Neural Information Processing Systems*, 35:124–141, 2022.

[Yin *et al.*, 2024] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024.

[Zhang *et al.*, 2023] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.