MHBench: Demystifying Motion Hallucination in VideoLLMs

Ming Kong[1], Xianzhou Zeng[1], Luyuan Chen[2], Yadong Li[3], Bo Yan[3], Qiang Zhu[1,✉]

[1]Zhejiang University,
[2]Beijing Information Science and Technology University,
[3]Ant Group
{zjukongming, xzzeng, zhuq}@zju.edu.cn; chenly@bistu.edu.cn; {liyadong.lyd, lengyu.yb}@antgroup.com

Abstract

Similar to Language or Image LLMs, VideoLLMs are also plagued by hallucination issues. Hallucinations in videos not only manifest in the spatial dimension regarding the perception of the existence of visual objects (static) but also the temporal dimension influencing the perception of actions and events (dynamic). This paper introduces the concept of Motion Hallucination for the first time, exploring the hallucination phenomena caused by insufficient motion perception capabilities in VideoLMMs, as well as how to detect, evaluate, and mitigate the hallucination. To this end, we propose the first benchmark for assessing motion hallucination **MHBench**, which consists of 1,200 videos of 20 different action categories. By constructing a collection of adversarial triplet types of videos (original/antonym/incomplete), we achieve a comprehensive evaluation of motion hallucination. Furthermore, we present a **M**otion **C**ontrastive **D**ecoding (**MotionCD**) method, which employs bidirectional motion elimination between the original video and its reverse playback to construct an amateur model that removes the influence of motion while preserving visual information, thereby effectively suppressing motion hallucination. Extensive experiments on MHBench reveal that current state-of-the-art VideoLLMs significantly suffer from motion hallucination, while the introduction of MotionCD can effectively mitigate this issue, achieving up to a 15.1% performance improvement. We hope this work will guide future efforts in avoiding and mitigating hallucinations in VideoLLMs.

Benchmark, Code and Appendix —
https://github.com/xzhouzeng/MHBench

Introduction

Language-Vision Large Models (LVLMs) achieve rapid progress by aligning visual information with textual representation space, which has extended to more challenging text-visual tasks, such as video understanding (Yin et al. 2023; Caffagni et al. 2024; Liang et al. 2024).

Hallucination has been recognized as a prominent concerns that impact the reality and applicability of LVLMs across domains (Liu et al. 2024a; Li et al. 2023c). One of the most critical issues is the factuality hallucination regarding the object's appearance, such as identifying *what objects are present in the image?* (Lovenia et al. 2023) However, in video understanding, the factuality hallucination also exists in the temporal dimension, making equally significant of detecting and mitigating hallucinations about the actions/events equally important, such as determining *what actions are performed by the person in the video*.

With the continuous emergence of video large language models, recently proposed video understanding benchmarks have attempted to include action-related evaluation tasks such as recognition, prediction, and sequence (Li et al. 2024;

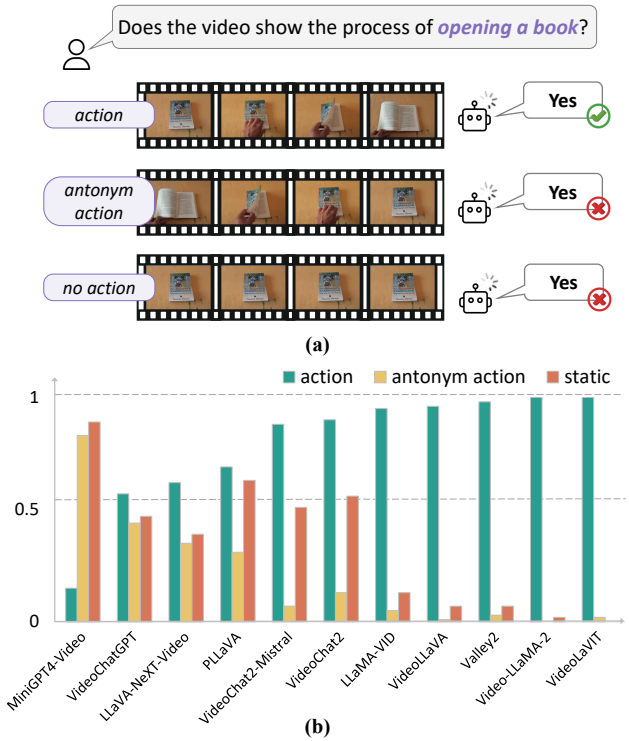

Figure 1: Illustration and preliminary evaluation results for VideoLLMs' motion context understanding ability in adversarial scenarios. a) shows an example of the triple-adversarial video group. b) presents the performance of 11 advanced VideoLLM models on this adversarial group.

Liu et al. 2024b). Despite these efforts, there is an unresolved question that remains: Whether the VideoLLMs recognition capability come from the awareness of static visual elements rather than the continuous motion context?

To validate the existence of the aforementioned issues, we designed a simple preliminary experiment. First, we collected 100 videos containing single action contents from the validation set of something2something-v2 (Goyal et al. 2017), and set the equal-length videos of reversing playback and fixed frames to form adversarial triplets. Figure 1.a demonstrate an example: for a video depicting the action of opening a book, the reversing playback illustrates its antonym action of closing a book, and the fixed frames show a closed book placed without any actions.

Then, we propose the same questions to each video in the adversarial triplets: *"Does the video show the process of* $\langle action \rangle$*?"* where $\langle action \rangle$ refers to the ground truth action from the original video.

Intuitively, if the VideoLLM accurately understands the motion process, it should affirm the action's existence in the original video while denying it in the adversarial videos. However, as shown in Figure 1.b, none of the 11 advanced VideoLLMs gained satisfactory performances. Some of them tend to make random responses with bias (such as MiniGPT4-Video (Ataallah et al. 2024)). Others recognize the actions in the original videos accurately, while falsely affirming the action's existence in the adversarial videos as well. Notably, none of them were able to generate correct answers for any video's adversarial triplet.

Based on the observed phenomena, we propose the concept of **Motion Hallucination** (MH): VideoLLMs failed to adequately learn and perceive dynamic changes, leading to incorrect interpretations of action existence and classification. Analogous to hallucinations in image understanding (Rohrbach et al. 2018) that the responses differ from the spatial existence of objects, motion hallucination can be treated as a discrepancy between the responses and the temporal existence of actions. In particular, if VideoLLMs solely rely on object appearance to infer actions' existence, they will inevitably suffer from motion hallucination.

To systematically evaluate motion hallucination in VideoLLMs, we constructed a benchmark dataset called **MHBench** consists of 1,200 videos and 20 action categories, of which contains a collection of triple-adversarial videos: 1) *original videos* that capture the entire process of the action; 2) *antonym videos* that depict the completion of the antonym action; and 3) *incomplete action videos* that feature the related objects without completing the action. On the basis, we designed two tasks of adversarial binary discrimination and action classification to comprehensively assess the severity and bias of motion hallucination in VideoLLMs.

Recognizing the causes of motion hallucination, we propose a simple yet effective mitigation method named Motion Contrastive Decoding (MotionCD). Specifically, we employ bidirectional motion elimination with the original video and its reversed playback to create an amateur model to remove motion influences while preserving the static appearance. By performing in contrast with the amateur model, VideoLLMs can reduce visual reliance and enhance motion perceptions, thereby mitigating motion hallucination.

We conducted comprehensive experiments to analyze the motion hallucination issue in video understanding. First, we performed extensive evaluations of 11 advanced VideoLLMs on MHBench, confirming they universally suffer from severe motion hallucination issues, which are not adequately reflected in the existing comprehensive VideoLLM benchmarks. We also analyzed the underlying causes of motion hallucination severness from the perspectives of model training and architecture aspects.

Furthermore, we demonstrated that applying MotionCD can effectively mitigate motion hallucination, achieving up to a 15.1% performance improvement. We hope this work can raise awareness of motion hallucination in VideoLLMs and inspire further progress in addressing this challenge.

The main contributions of this paper are summarized as follows:

- We demystify motion hallucination in VideoLLMs for the first time and propose MHBench, the first benchmark dataset for evaluating motion hallucination.
- We propose the MotionCD method as the motion hallucination mitigation baseline, which constructs an amateur model through bidirectional motion elimination using forward-reverse video pairs and performs contrastive decoding to enhance the motion perception of VideoLLMs.
- We conduct an empirical study on advanced VideoLLMs using the MHBench dataset, confirming that existing VideoLLMs significantly suffer from motion hallucination and demonstrating MotionCD can effectively mitigate this issue.

Related Work

Benchmark for VideoLLMs

To evaluate VideoLLMs' perception and understanding of videos, a series of video understanding benchmark datasets have been proposed, such as Video-Bench (Ning et al. 2023), AutoEval-Video (Chen et al. 2023), MMBench-Video (Wang et al. 2023) and VideoMME (Fu et al. 2024). However, most of these benchmarks focus on assessing the appearance of objects or events in videos without measuring the understanding of continuous actions. Some recent efforts have started to incorporate action evaluation to provide a more comprehensive assessment of VideoLLMs' video understanding (Li et al. 2024; Liu et al. 2024b). ActionBench tries to evaluate motion hallucination of videos by setting up *action replacement* and *video reversal* to construct antonym descriptions, and is included by MVBench (Li et al. 2024), currently the most comprehensive benchmark for video understanding. MVBench also considers more action-related aspects, such as sequence, prediction, counting, etc. However, these action evaluations are typically based on single videos, making it difficult to determine whether a model's understanding of actions and events is rooted in the coherent motion process or merely in the typical appearance associated with the actions.

In this paper, we propose an evaluation benchmark MHBench, aiming to make up for the deficiency of motion understanding evaluation of existing VideoLLM benchmarks.

By setting up contrasts with videos of semantic antonym/incomplete actions, our benchmark provides a better assessment of VideoLMM capabilities from the perspective of action understanding.

Hallucination in LMMs

Hallucination has become one of the most critical issues in recent LMM research (Liu et al. 2024a). In image-text modeling, the primary focus is on object hallucination, i.e., the response includes objects that are not present in the image (Rohrbach et al. 2018; Li et al. 2023c). Subsequent research aims to refine its definition and assessment, and extend hallucination evaluation to other visual cues, such as relationships, attributes, counting, and OCR (Jing et al. 2023; Jiang et al. 2024). In the context of VideoLLMs, some text-visual hallucination benchmarks extend image-based hallucination evaluation about static visual attributes to dynamic content, such as actions, events, and narratives, and input videos as sequences of images (Guan et al. 2024; Ravi et al. 2024; Fang et al. 2024). VideoHallucer(Wang et al. 2024) first proposes a comprehensive hallucination detection benchmark for VideoLLMs, using adversarial binary questions to evaluate both factual and non-factual hallucinations related to object relationships, temporal aspects, and semantic details. Despite various hallucination evaluation approaches, these benchmarks primarily focus on hallucinations caused by the appearance of objects and events, neglecting the evaluation of motion-related hallucinations. (Ullah and Mohanta 2022) considers the significance of motion hallucination in video captions and proposes an evaluation metric from the semantic similarity aspect, which is most akin to our concern. However, to date, no work has defined or assessed hallucination in VideoLLMs from the perspective of the correct sequence or completeness of actions or events in videos.

A significant amount of work has been dedicated to LLMs' hallucination mitigation, such as contrastive decoding (Leng et al. 2024; Favero et al. 2024), compare and select (Deng, Chen, and Hooi 2024), or rollback (Huang et al. 2024). Although many text-visual hallucination mitigation strategies can be extended to video, they are not specifically designed to address the motion hallucination issue.

To address the aforementioned gaps, we define motion hallucination and propose MHBench for its evaluation. Besides, we introduce Motion Contrastive Decoding (MotionCD) as the baseline of motion hallucination mitigation.

Motion Hallucination Benchmark

We introduce the Motion Hallucination Benchmark (MHBench), the first benchmark for evaluating motion hallucination of VideoLLMs. In this section, we first introduce the construction process of MHBench and then demonstrate how to define the question-answering tasks for evaluating the severity and bias of motion hallucination in VideoLLMs.

Dataset Construction

The dataset construction process is illustrated in Figure 2. We first identified 20 common actions. To minimize the semantic ambiguity, these actions are clearly defined and not affected by factors like direction or magnitude. Additionally, each action has a clear antonym, such as *lifting* vs. *throwing*, *inserting* vs. *removing*, etc. The detailed definition of action categories can be referred to in the Appendix.

For each action category, we collect three types of videos:

- **Original Videos**: A video of the completed process of executing the specific action, such as rolling a sheet of unfolded paper into a tube.
- **Antonym Videos**: A video of the completed process of executing the antonym action of the original action, such as unrolling a tube of paper.
- **Incomplete Action Videos**: A video of still frames or unfinished action, such as a sheet of unfolded paper placed on the table, or the action of putting a hand on the paper.

Compared to the original action videos, both antonym and incomplete action videos contain similar visual objects but with adversarial motion semantics. Specifically, antonym movies exhibit semantic opposition to the action, reflecting the differences in process and intent; while incomplete action movies show factual opposition to the action, showing the differences in process completeness and final status.

We collect videos through three channels to ensure the variety of content and styles of our dataset: 1) selected from the validation set of something2something-v2 (Goyal et al. 2017), 2) shot by volunteer members, and 3) collected from the internet. All the collected videos are checked and labeled by at least two annotators, those with inappropriate length, insufficient clarity, or ambiguous content are excluded.

Ultimately, MHBench incorporated 1,200 high-quality videos, with each action category including 20 samples per defined action type, ranging in length from 2 to 36 seconds.

Motion Hallucination Evaluation

We set up various question-answering tasks for the motion hallucination evaluation with two kinds of challenges:

Adversarial Binary Discrimination A typical manifestation of motion hallucination is the model's tendency to infer action existence solely based on the static object appearances, regardless of the motion procedure taking place. For example, once a book appears in the video, the probability of responding *"yes"* will raise to both questions of *"Can you see the action of opening the book in the video?"* and *"Can you see the action of closing the book in the video"*.

To address this issue, we designed an adversarial binary discrimination QA pair with affirmative and negative questions. Specifically, for the original and antonymous videos, the question pair is like: *"Does the video show the action of* $\langle action \rangle$/$\langle antonym\ action \rangle$*?"* For the incomplete-action videos, the question pair is like *"Does the video show the action of* $\langle action \rangle$/$\langle antonym\ action \rangle$*?"* versus *"Is the video showing neither* $\langle action \rangle$ *nor* $\langle antonym\ action \rangle$ *in this video?"* We aimed to evaluate the model's ability to recognize both actions' existence and non-exist actions' absence.
Evaluation: Considering the symmetric distribution of the Yes/No answers to the challenge, the average accuracy (AvgAcc) is an adequate metric for evaluating the overall performance. We also introduce the accuracy of the questions with

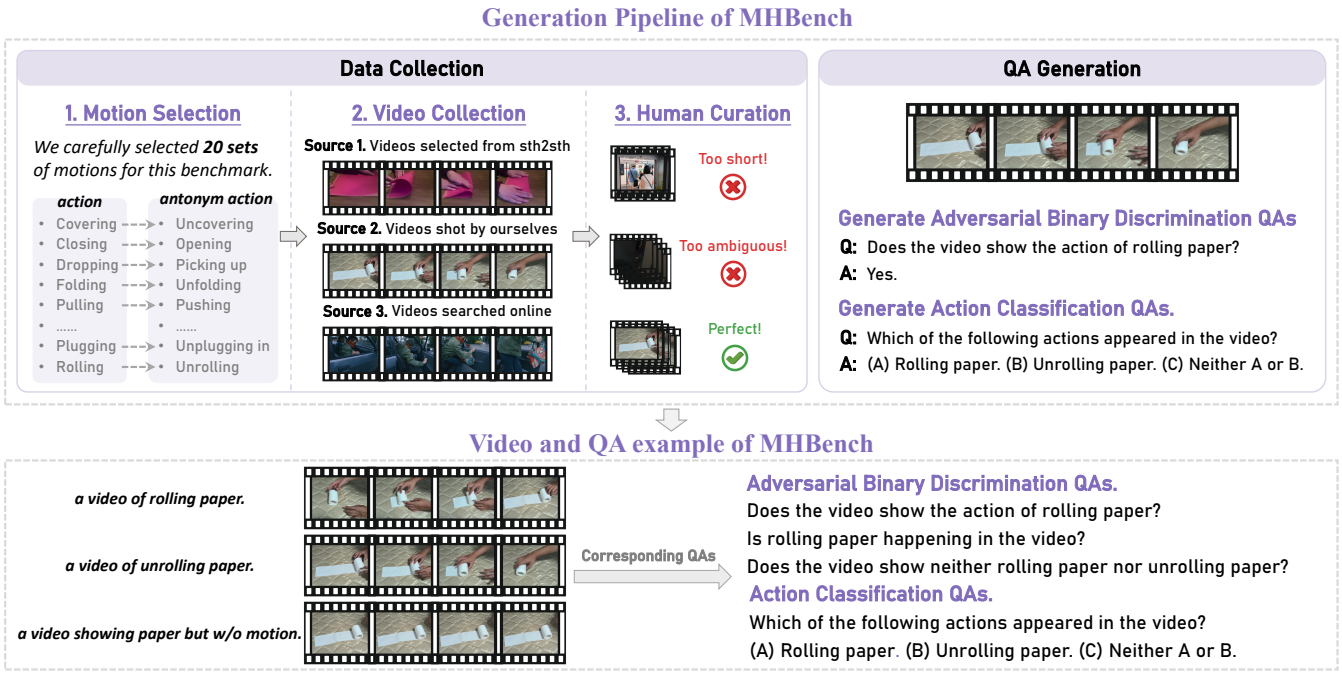

Figure 2: Illustration of MHBench construction process and the definition of the motion hallucination evaluation tasks.

respective Yes/No answers, i.e., Yes Accuracy (YesAcc) and No Accuracy (NoAcc) to evaluate the model capability of factual confirmation and counterfactual negation. Furthermore, to evaluate the comprehensive understanding ability of the model to the adversarial actions, we introduce a paired accuracy metric (PairAcc), which counts the accuracy of correctly answering both questions in the adversarial pair:

$$PairAcc = \frac{1}{|V|}\sum_{v\in V}\mathbb{I}[\delta(p_{v,1}) \wedge \delta(p_{v,2})] \quad (1)$$

where V is the set of videos, $p_{v,i}$ is the video-question pair of the video v and its i-th question. δ is the discriminant function, which returns 1 when the answer is correct and 0 otherwise. $\mathbb{I}[\cdot]$ is the indicator function.

Considering the semantic prior and the overreliance on static visual perception, we also measure the bias of VideoLLMs toward affirmation and negative responses. Following (Guan et al. 2024), we introduce *Yes Percentage Difference* (PctDiff) and *False Positive Ratio* (FPR) to measure the bias in the model's responses, defined as:

$$PctDiff = \frac{|\{M(p) = \text{yes}\}_{p\in P}| - |GT(p) = \text{yes}\}_{p\in P}|}{|P|}$$
$$FPR = \frac{|\{M(p) = \text{yes}\}_{p\in W}|}{|W|} \quad (2)$$

where P is the set of video-question pairs, $M(p)$ and $GT(p)$ are the prediction and ground-truth answer of the video-question pair p, and W is the set of video-question pairs with incorrect answers. A smaller *PctDff* indicates the model's frequency of responding "yes" is closer to the ground truth, suggesting less discrimination bias. FPR represents the proportion of "yes" responses among incorrect answers, and a value closer to 50% suggests less bias.

Action Classification Due to the semantic biases in VideoLLMs, adversarial binary discrimination cannot evaluate the model's distinguishability in action. To address this limitation, we designed a set of ternary-choice questions for each video: "Which action is included in the following video? A.$\langle action\rangle$, B. $\langle antonym\ action\rangle$ and C. $Neither\ A\ nor\ B$." where the options are randomly shuffled in the actual scenario. The challenge effectively avoids the ambiguity issues in the adversarial binary discrimination and provides a more intuitive assessment of the model's action classification capabilities. Given the similar object's appearance, VideoLLMs must rely on their motion-understanding ability to make decisions.

Evaluation We utilized Macro-Precision, Recall, and F1-scores to evaluate the performance of the action classification task, which comprehensively reflects the model's action recognition ability for the multiple-choice challenge.

Motion Contrastive Decoding

Recognizing the cause of motion hallucination, we propose a simple, training-free hallucination mitigation method called Motion Contrastive Decoding (MotionCD), as the baseline for addressing motion hallucination. Inspired by Visual Contrastive Decoding (VCD) (Leng et al. 2024) for spatial visual hallucination mitigation, MotionCD attempts to counteract the bias and priors of VideoLLMs that overrelying on the static visual perception by contrasting the output probability distributions generated from original and amateur models.

Given a text query x and a visual input v, the contrastive

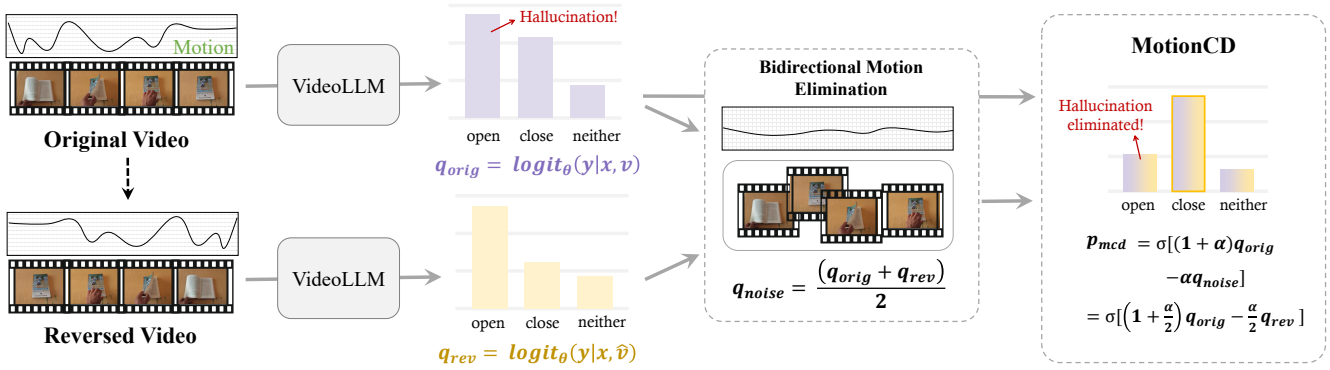

Figure 3: Illustration of the Motion Contrastive Decoding process. By averaging the probability distributions from the original and reversed videos, a distribution is generated that removes motion while preserving visual semantics, and then contrast is applied to suppress motion hallucinations.

decoding allows the model to generate output distributions from both the original input v and the distorted input v'. The contrastive probability distribution reduces hallucination by calculating the difference between the original and hallucination-tend distributions, denoted as:

$$p_{mcd} = softmax[(1+\alpha)q_{orig} - \alpha q_{noise}] \quad (3)$$

where $q_{orig} = f_\theta(y|v,x)$, and $q_{noise} = f_\theta(y|v',x)$, denoting the logit values of LLM prediction of the next token based on various inputs. α represents the intensity of the contrast difference between distributions.

Visual contrastive decoding constructs an amateur model by adding Gaussian noise or Diffusion noise to the original visual input, increasing the uncertainty of visual representations to amplify hallucinations. For motion contrastive decoding, the main challenge is generating a distorted probability distribution that disrupts the motion context of the original video while preserving the static visual semantics. The most intuitive approach is to randomly shuffle the video frames; however, this method makes it difficult to precisely control the noise intensity. Specifically, random shuffling indiscriminately weakens the logits associated with non-action-related vocabulary. Another approach is to reduce the video sampling rate or omit the video entirely, but these methods risk losing the appearance of static objects.

As illustrated in Figure 3, we propose a Bidirectional Motion Elimination (BME) strategy to generate the distorted distribution. Specifically, we treat the distorted distribution as the mean of the output distributions from the original video and the reversed video, expressed as:

$$q_{noise} = \frac{f_\theta(y|v,x) + f_\theta(y|\hat{v},x)}{2} \quad (4)$$

where $\hat{v}$ represents the reversed playback of the video.

Compared to the original video, the logit distribution in the reversed video shows similar static awareness and mutual exclusive motion perceptions. Therefore, Bidirectional Motion Elimination can offset the influence of motion context awareness, thereby amplifying the motion hallucinations that VideoLLMs infer based on object awareness. The contrastive probability distribution can then be expressed as:

$$p_{mcd} = softmax[(1+\frac{\alpha}{2})f_\theta(y|v,x) - \frac{\alpha}{2}f_\theta(y|\hat{v},x)] \quad (5)$$

Notably, by simplifying, the contrastive probability distribution depends only on the logits of the original and reversed playback inputs, making it easy to obtain.

Following VCD, MotionCD further introduces an adaptive plausibility constraint to mitigate indiscriminate language biases and common-sense reasoning penalties. This ensures that the candidate pool is simplified when the output corresponding to the original input has high confidence.

$$\begin{aligned} \mathcal{V}_{\text{head}}(y_{<t}) = \{y_t \in \mathcal{V} : \\ f_\theta(y_t \mid v,x,y_{<t}) \geq \beta \max_w f_\theta(w \mid v,x,y_{<t})\}, \\ p_{mcd}(y_t|v,\hat{v},x) = 0, \text{ if } y_t \notin \mathcal{V}_{\text{head}}(y_{<t}), \end{aligned} \quad (6)$$

where $\mathcal{V}$ is the output vocabulary of the VideoLLM, and $\beta \in [0,1]$ is a hyperparameter for controlling the truncation of the next token distribution. A larger β indicates more aggressive truncation to keep only high-probability tokens.

In summary, the output prediction probability of MotionCD can be expressed as:

$$\begin{aligned} y_t \sim & softmax[(1+\frac{\alpha}{2})f_\theta(y|v,x,y_{<t}) - \frac{\alpha}{2}f_\theta(y|\hat{v},x,y_{<t})] \\ & \text{subject to } y_t \in \mathcal{V}_{\text{head}}(y_{<t}) \end{aligned} \quad (7)$$

where y_t represents the output of the t-th token sampled from the logit distribution.

Experiments

In this chapter, we first evaluate the latest advanced VideoLLMs on MHBench to reveal their existing motion hallucination issues, and then we demonstrate the effectiveness of the proposed MotionCD method in mitigating these motion hallucinations.

Results on Main Baselines

As shown in Table 1, we conduct a motion hallucination evaluation on MHBench for 11 advanced open-source VideoLLMs (all standardized to the 7B model size, retaining the

Models	LLM	MVBench	MHBench Adversarial Binary Discrimination						MHBench Action Classification		
		Acc	PctDiff	FPR	YesAcc	NoAcc	PairAcc	AvgAcc	Macro Precision	Macro Recall	Macro F1
Random	-	27.3	0.00	0.50	50.0	50.0	25.0	50.0	33.3	33.3	33.3
VideoChatGPT(Maaz et al. 2023)	LLaMA-7B	33.3	-0.05	0.36	53.3	63.9	29.6	58.6	11.1	33.3	16.6
Valley2(Luo et al. 2023)	LLaMA2-7B	30.1	0.16	1.0	97.2	0.1	0.0	42.6	26.6	31.7	20.4
VideoChat2(Li et al. 2023a)	Vicuna-7B-v0	50.0	0.10	0.56	**73.9**	**43.9**	31.5	**59.3**	44.9	39.0	34.9
Video-LLaVA(Lin et al. 2023)	Vicuna-7B-v1.5	43.1	0.17	0.80	53.8	20.4	7.5	31.7	32.5	36.5	30.8
LLaMA-VID(Li, Wang, and Jia 2023)	Vicuna-7B-v1.5	40.6	0.27	0.91	62.4	8.7	3.9	35.5	22.5	33.8	27.0
VideoLaVIT(Jin et al. 2024)	LLaMA2-7B	43.2	0.46	1.0	99.6	0.0	0.0	47.8	34.2	33.7	29.8
Video-LLaMA2(Cheng et al. 2024)	LLaMA2-7B	36.7	0.06	0.97	93.5	3.0	1.4	48.1	19.5	33.3	17.0
MiniGPT4-Video(Ataallah et al. 2024)	Mistral-7B	36.5	-0.28	0.09	38.3	91.2	**33.5**	48.1	38.0	38.0	37.4
VideoChat2-Mistral(Li et al. 2024)	Mistral-7B	**60.6**	-0.38	0.80	96.8	19.9	19.4	58.3	**55.4**	**51.2**	**50.1**
PLLaVA(Xu et al. 2024)	Vicuna-7B-v1.5	45.2	-0.33	0.20	14.2	80.5	7.4	47.3	40.3	35.4	28.6
LLaVA-NeXT-Video-DPO(Zhang et al. 2024)	Vicuna-7B-v1.5	40.0	-0.34	0.19	13.2	81.5	5.8	47.3	37.8	31.2	26.3

Table 1: Motion hallucination evaluation on MHBench for advanced VideoLLMs. MVBench results are our reproduced outcomes. The bolded result represents the best outcome, with accuracy presented as a percentage.

original settings of all baselines for fairness). Additionally, we include the results on the general video understanding benchmark MVBench as a reference.

Although these models all exceeded random levels on MVBench, most of them gain unsatisfactory performance on both adversarial binary discrimination and action classification tasks on MHBench, with some even failing to surpass random guessing. Most methods show severe bias in adversarial binary discrimination, i.e., they all tend to make the same responses to $\langle action \rangle$, $\langle antonym\ action \rangle$ and $\langle incompleteaction \rangle$ videos, resulting in extremely low performance on PairAcc. The results reveal their high dependence on object appearance and semantic priors, failing to take the impacts of motion perception into account.

Among the models, VideoChat2-Mistral and MiniGPT4-Video achieve better performance on the action classification task, showing a stronger capability of action discrimination. However, they still suffer from severe bias in adversarial binary discrimination, indicating their action presence discriminability is affected by semantic priors. We expect hallucination mitigation methods can show significant improvements in these models. For more detailed experiments and analyses, please refer to the Appendix.

Discussion Based on the results, we explored potential strategies for constructing VideoLLMs with reduced motion hallucinations. Firstly, the top-performing models leverage advanced visual encoders that are inherently well-suited for video encoding, such as UMT-L (Li et al. 2023b) for VideoChat2 and EVA-CLIP (Sun et al. 2023) for MiniGPT4-Video. Although some approaches, such as VideoChatGPT and VideoLaVIT, incorporate mechanisms to align video context and motion vectors, they are not observed highly effective in addressing motion hallucinations. Additionally, incorporating a sufficient quantity and diversity of video data during instruction fine-tuning is essential. For instance, the VideoChat2 series used 1.9 million samples generated by ChatGPT, drawn from 34 datasets across 6 categories, covering various temporal tasks, which contributed to its superior performance. Lastly, the choice of the large language model backbone significantly impacts video hallucinations. Note that models using Mistral-7B (Jiang et al. 2023) demonstrate a clear performance edge over those using Vicuna-7B v0 (Chiang et al. 2023) (e.g., VideoChat2-Mistral and MiniGPT4-Video vs VideoChat2).

Models	Dis Ques		MC Ques
	PairAcc	AvgAcc	Macro F1
Video-LLaVA	7.5	37.1	30.8
Video-LLaVA + **MotionCD**	**10.4**	**45.2**	**33.9**
VideoChat2	**31.5**	59.3	34.9
VideoChat2 + **MotionCD**	29.2	**59.7**	**40.9**
VideoChat2-Mistral	19.4	58.3	50.1
VideoChat2-Mistral + **MotionCD**	**36.3**	**65.2**	**65.2**

Table 2: Experimental results of MotionCD on representative VideoLLM baselines on MHBench. MC stands for multiple-choice. For more results, refer to the appendix.

Results with MotionCD

Effectiveness We verify the motion hallucination mitigation ability of MotionCD to several representative VideoLLMs. We select the optimal hyperparameter settings through a grid search on the VideoChat2-Mistral model and extend it to other models. Specifically, the noise intensity α=20 and the adaptive rationality constraint β=0.1.

As shown in Table 2, applying MotionCD can significantly improve performance on various models. In particular, for VideoChat2-Mistral, the best-performing model, MotionCD can increase the average accuracy against binary action discrimination by 6.9%, pair accuracy by 16.9%, and macro f1-score for action classification by 15.1%. Besides, MotionCD can effectively suppress the tendency problem in binary action discrimination, which fully demonstrates its motion perception enhanced ability.

We conduct further experiments for the effectiveness of the Bidirectional Motion Elimination strategy by comparing the token generation from the motion-eliminated and original inputs generated probability distributions and analyze the proportion of different types of errors on the MHBench multiple-choice questions.

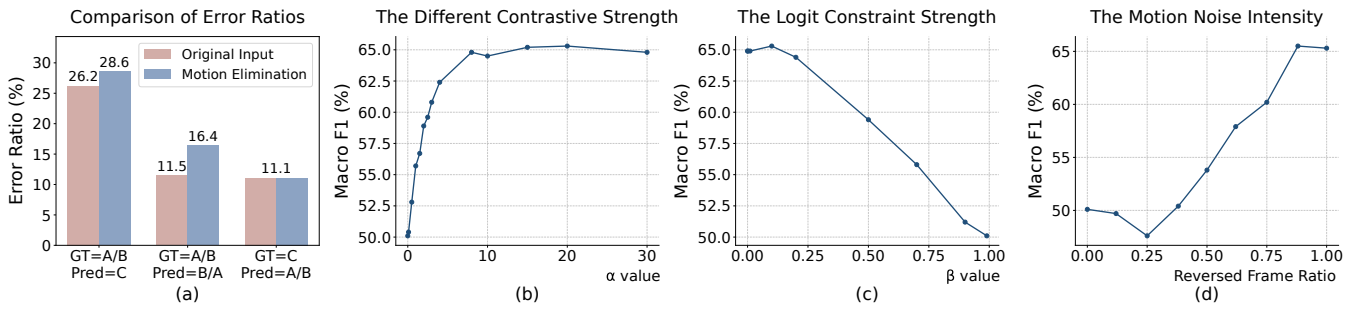

Figure 4: Evaluation of Bidirectional Motion Elimination and the parameter sensitivity on MHBench multiple-choice questions using the VideoChat2-Mistral model. (a) Error distribution comparison between motion elimination and original inputs. A/B options indicate actions, while C represents incomplete or no action. (b-d) Impact of noise intensity α, adaptive rationality constraint β, and reversed frame ratio on model performance.

As shown in Figure 4.a, our strategy 1) tends to predict no-action responses, meaning the motions are effectively eliminated; 2) increases the error rate in distinguishing between actions and antonym actions, indicating more pronounced hallucinations; and 3) maintains the error rate for predicting no action videos as having actions, which suggests that the strategy effectively preserves visual information without introducing additional uncertain motion noise.

Ablation Study To demonstrate the effectiveness of our proposed amateur model construction strategy of Bidirectional Motion Elimination, we compared different strategies for adding motion noise to input videos (amateur model construction methods) on VideoChat2-Mistral. As shown in Table 3, the amateur model construction method based on BME achieved the best results. We believe that adding noise to video frame images, static video or no video will affect the static visual information of the video, and shuffling will indiscriminately weaken the vocabulary logits that do not contain action-related information, thereby affecting the discrimination of videos that do not contain action.

We also experimented with different model parameter values, as shown in Figure 4.b-d. b) When the noise intensity α = 20, the model performance reaches its peak, indicating that the model has some motion perception ability, but it is relatively weak and requires significant amplification to perform optimally. c) The optimal value for the adaptive rationality truncation hyperparameter β was 0.1, suggesting that appropriate constraints can mitigate errors caused by low-confidence predictions. d) Additionally, we performed an ablation study on the reversed frames ratio, which corresponds to the degree of motion interference. The overall upward trend indicates that the design of amateur models with high-intensity motion interference/elimination is crucial for achieving strong comparative results.

Amateur Models	Dis Ques		MC Ques
	PairAcc	AvgAcc	Macro F1
w/o CD	19.4	58.3	50.1
shuffle	24.3	58.2	58.1
w/o video	16.7	52.2	48.2
add visual noise	30.0	60.2	51.9
static video	9.7	46.6	54.4
BME (ours)	**36.3**	**65.2**	**65.2**

Table 3: Comparison of different motion noise strategies for amateur model construction on VideoChat2-Mistral. *w/o CD*: without contrastive decoding; *shuffle*: apply a random frame order; *w/o video*: remove the video input; *add visual noise*: simply follow VCD (Leng et al. 2024); and *static video*: repeat a single frame.

Conclusion

For VideoLLMs, hallucinations extend beyond the static appearance on spatial dimension to encompass more significant challenges in the temporal dimension, known as Motion Hallucinations. This paper establishes MHBench, a benchmark designed to evaluate motion hallucinations in VideoLLMs through various tasks that objectively assess the models' motion perceptual capabilities. On this basis, we propose a simple yet effective motion contrastive decoding method MotionCD to alleviate the motion hallucinations by constructing an amateur model using bidirectional motion elimination to mitigate statistical biases from static appearances. Evaluation of 11 advanced VideoLLMs reveals the widespread prevalence of motion hallucinations and underscores the limitations of current video understanding benchmarks. Furthermore, we demonstrate that MotionCD significantly enhances the models' ability to understand actions in motion contexts, effectively reducing motion hallucinations.

It is worth noting that we recognize the scale and evaluation granularity of MHBench are still limited. We will keep working on it to improve its action category, video amount, and challenge diversity. We also don't count on MotionCD to thoroughly solve the motion hallucination problem of the videoLLMs. Instead, we hope this work can raise awareness of motion hallucinations among researchers and provide a solid foundation for future studies in this area.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 42394060 and 42394064, Ant Group, and the Earth System Big Data Platform of the School of Earth Sciences, Zhejiang University.

References

Ataallah, K.; Shen, X.; Abdelrahman, E.; Sleiman, E.; Zhu, D.; Ding, J.; and Elhoseiny, M. 2024. MiniGPT4-Video: Advancing Multimodal LLMs for Video Understanding with Interleaved Visual-Textual Tokens. *arXiv preprint arXiv:2404.03413*.

Caffagni, D.; Cocchi, F.; Barsellotti, L.; Moratelli, N.; Sarto, S.; Baraldi, L.; Cornia, M.; and Cucchiara, R. 2024. The (r) evolution of multimodal large language models: A survey. *arXiv preprint arXiv:2402.12451*.

Chen, X.; Lin, Y.; Zhang, Y.; and Huang, W. 2023. Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering. *arXiv preprint arXiv:2311.14906*.

Cheng, Z.; Leng, S.; Zhang, H.; Xin, Y.; Li, X.; Chen, G.; Zhu, Y.; Zhang, W.; Luo, Z.; Zhao, D.; and Bing, L. 2024. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476*.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Deng, A.; Chen, Z.; and Hooi, B. 2024. Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding. *arXiv preprint arXiv:2402.15300*.

Fang, X.; Mao, K.; Duan, H.; Zhao, X.; Li, Y.; Lin, D.; and Chen, K. 2024. MMBench-Video: A Long-Form Multi-Shot Benchmark for Holistic Video Understanding. *CoRR*, abs/2406.14515.

Favero, A.; Zancato, L.; Trager, M.; Choudhary, S.; Perera, P.; Achille, A.; Swaminathan, A.; and Soatto, S. 2024. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14303–14312.

Fu, C.; Dai, Y.; Luo, Y.; Li, L.; Ren, S.; Zhang, R.; Wang, Z.; Zhou, C.; Shen, Y.; Zhang, M.; et al. 2024. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. *arXiv preprint arXiv:2405.21075*.

Goyal, R.; Ebrahimi Kahou, S.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fruend, I.; Yianilos, P.; Mueller-Freitag, M.; et al. 2017. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, 5842–5850.

Guan, T.; Liu, F.; Wu, X.; Xian, R.; Li, Z.; Liu, X.; Wang, X.; Chen, L.; Huang, F.; Yacoob, Y.; Manocha, D.; and Zhou, T. 2024. HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models. arXiv:2310.14566.

Huang, Q.; Dong, X.; Zhang, P.; Wang, B.; He, C.; Wang, J.; Lin, D.; Zhang, W.; and Yu, N. 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*, 13418–13427.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Jiang, C.; Ye, W.; Dong, M.; Jia, H.; Xu, H.; Yan, M.; Zhang, J.; and Zhang, S. 2024. Hal-eval: A universal and fine-grained hallucination evaluation framework for large vision language models. *arXiv preprint arXiv:2402.15721*.

Jin, Y.; Sun, Z.; Xu, K.; Chen, L.; Jiang, H.; Huang, Q.; Song, C.; Liu, Y.; Zhang, D.; Song, Y.; et al. 2024. Video-lavit: Unified video-language pre-training with decoupled visual-motional tokenization. *arXiv preprint arXiv:2402.03161*.

Jing, L.; Li, R.; Chen, Y.; Jia, M.; and Du, X. 2023. Faithscore: Evaluating hallucinations in large vision-language models. *arXiv preprint arXiv:2311.01477*.

Leng, S.; Zhang, H.; Chen, G.; Li, X.; Lu, S.; Miao, C.; and Bing, L. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13872–13882.

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; et al. 2023a. MVBench: A Comprehensive Multi-modal Video Understanding Benchmark. *CoRR*.

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; et al. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22195–22206.

Li, K.; Wang, Y.; Li, Y.; Wang, Y.; He, Y.; Wang, L.; and Qiao, Y. 2023b. Unmasked Teacher: Towards Training-Efficient Video Foundation Models. arXiv:2303.16058.

Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J.-R. 2023c. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Li, Y.; Wang, C.; and Jia, J. 2023. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*.

Liang, T.; Huang, J.; Kong, M.; Chen, L.; and Zhu, Q. 2024. Querying as Prompt: Parameter-Efficient Learning for Multimodal Language Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26855–26865.

Lin, B.; Zhu, B.; Ye, Y.; Ning, M.; Jin, P.; and Yuan, L. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Liu, H.; Xue, W.; Chen, Y.; Chen, D.; Zhao, X.; Wang, K.; Hou, L.; Li, R.; and Peng, W. 2024a. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Liu, Y.; Li, S.; Liu, Y.; Wang, Y.; Ren, S.; Li, L.; Chen, S.; Sun, X.; and Hou, L. 2024b. TempCompass: Do Video LLMs Really Understand Videos? *arXiv preprint arXiv: 2403.00476*.

Lovenia, H.; Dai, W.; Cahyawijaya, S.; Ji, Z.; and Fung, P. 2023. Negative object presence evaluation (nope) to measure object hallucination in vision-language models. *arXiv preprint arXiv:2310.05338*.

Luo, R.; Zhao, Z.; Yang, M.; Dong, J.; Qiu, M.; Lu, P.; Wang, T.; and Wei, Z. 2023. Valley: Video Assistant with Large Language model Enhanced abilitY. arXiv:2306.07207.

Maaz, M.; Rasheed, H.; Khan, S.; and Khan, F. S. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

Ning, M.; Zhu, B.; Xie, Y.; Lin, B.; Cui, J.; Yuan, L.; Chen, D.; and Yuan, L. 2023. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv preprint arXiv:2311.16103*.

Ravi, S. S.; Mielczarek, B.; Kannappan, A.; Kiela, D.; and Qian, R. 2024. Lynx: An Open Source Hallucination Evaluation Model. *arXiv preprint arXiv:2407.08488*.

Rohrbach, A.; Hendricks, L. A.; Burns, K.; Darrell, T.; and Saenko, K. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Sun, Q.; Fang, Y.; Wu, L.; Wang, X.; and Cao, Y. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Ullah, N.; and Mohanta, P. P. 2022. Thinking Hallucination for Video Captioning. In Wang, L.; Gall, J.; Chin, T.; Sato, I.; and Chellappa, R., eds., *Computer Vision - ACCV 2022 - 16th Asian Conference on Computer Vision, Macao, China, December 4-8, 2022, Proceedings, Part IV*, volume 13844 of *Lecture Notes in Computer Science*, 623–640. Springer.

Wang, Y.; Wang, Y.; Zhao, D.; Xie, C.; and Zheng, Z. 2024. VideoHallucer: Evaluating Intrinsic and Extrinsic Hallucinations in Large Video-Language Models. *arXiv preprint arXiv:2406.16338*.

Wang, Z.; Blume, A.; Li, S.; Liu, G.; Cho, J.; Tang, Z.; Bansal, M.; and Ji, H. 2023. Paxion: Patching Action Knowledge in Video-Language Foundation Models. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Xu, L.; Zhao, Y.; Zhou, D.; Lin, Z.; Ng, S. K.; and Feng, J. 2024. PLLaVA : Parameter-free LLaVA Extension from Images to Videos for Video Dense Captioning. arXiv:2404.16994.

Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; and Chen, E. 2023. A Survey on Multimodal Large Language Models. arXiv:2306.13549.

Zhang, Y.; Li, B.; Liu, h.; Lee, Y. j.; Gui, L.; Fu, D.; Feng, J.; Liu, Z.; and Li, C. 2024. LLaVA-NeXT: A Strong Zero-shot Video Understanding Model.