EchoBat: Echo-Vision Enhancement and Echo-Layered Sampling for Video LLMs Hallucination Mitigation

Shuai Liu[1], Da Chen[1,3], Yiheng Pan[1], Chenwei Tian[1], Qian Li[2], Chenhao Lin[2*]

[1]School of Software Engineering, Xi'an Jiaotong University
[2]School of Cyber Science and Engineering, Xi'an Jiaotong University
[3]ByteDance
sh_liu@mail.xjtu.edu.cn, chenda.0@bytedance.com
{dachen, panyiheng, 2216113609}@stu.xjtu.edu.cn, {linchenhao, qianlix}@xjtu.edu.cn

Abstract

Recent advancements in multimodal large language models (MLLMs) have shown remarkable progress in video understanding. However, video MLLMs (VideoMLLMs) still suffer from hallucinations, generating nonsensical or irrelevant content. This issue partly stems from over-reliance on pre-trained knowledge, sometimes neglecting the rich visual information present in the video. Additionally, many existing methods rely on uniform frame sampling, which can overlook critical visual cues. To address these challenges, we present EchoBat, a novel approach that leverages audio information as well as video temporal and logical consistency to improve preference data construction and keyframe extraction. Our method integrates Direct Preference Optimization (DPO) to mitigate hallucinations by leveraging high-quality, contextually rich preference feedback. Specifically, we use GPT-4o to generate high-quality video descriptions and integrate visually relevant segments from Whisper-derived transcripts to construct preference responses. Correspondingly, we use the reference model itself to describe the reversed video, and use GPT-4o to flashback the text and fill in the hallucination to produce non-preferred responses. This strategy enhances the model's ability to better understand visual content and temporal, logical relationships within videos. Furthermore, we propose an echo-layered sampling strategy for keyframe extraction from videos, which can provide more precise visual supervision compared to uniform sampling. Experimental results on the three latest video hallucination benchmarks demonstrate the effectiveness of our approach.

Introduction

With the development of MLLMs, video understanding has recently gained significant attention (Govindasamy et al. 2025). VideoMLLMs, through techniques such as transfer learning, have evolved from understanding single images to understanding multiple images, and eventually to comprehending videos. By combining visual feature encoders with large language models (LLMs), VideoMLLMs can extract visual features from frame sequences using the visual feature encoder. The model then analyzes them with prompt input into the LLM, ultimately generating responses and demonstrating remarkable video understanding capabilities (Xie et al. 2023). Recently, some works have trained VideoMLLMs on large-scale multimodal instruction datasets, enabling these models to understand videos while also following instructions. This marks an important step towards achieving general artificial intelligence, realizing the true potential of MLLMs, with promising applications in fields such as autonomous driving (Huo et al. 2025; Zhang et al. 2025) and multimodal robotics (Choi, Kim, and Lee 2025; Chirila et al. 2024; Traum et al. 2024).

However, we find that similar to MLLMs, VideoMLLMs are still significantly affected by hallucination issues. To address these issues, numerous approaches have been proposed, including improving the range and quality of training data, developing new alignment strategies between vision and LLM, and optimizing training strategies. Some studies have argued that a significant portion of hallucinations in MLLMs to the models' tendency to rely on pre-existing knowledge while overlooking provided visual information (Ji et al. 2023). Additionally, for VideoMLLMs, understanding videos through uniform frame sampling may cause the omission of crucial frames that are essential for answering the question, thereby introducing hallucinations.

Fortunately, videos inherently contain valuable extra information compared to images, such as the exact timeline and order of scenes. This characteristic makes it easier to design and extract preferred pairs for VideoMLLMs than for image-text MLLMs. Based on the above findings, we construct the preferred pairs consist of high-quality texts with clear timelines, precise scene descriptions and audio-assisted visual localization, while the non-preferred pairs consist of low-quality texts with chaotic timelines, vague scene descriptions and factual errors. In addition, we propose an echo-layered sampling algorithm to optimize the current frame sampling algorithm, which prioritizes key frames with high semantic relevance to the question and large scene feature differences. For the remaining frames, uniform sampling is used to supplement, which retains most of the key visual information.

Taking LLaVA-video-7B as an example, the use of our proposed EchoBat algorithm greatly reduces its hallucinations, elevating LLaVA-video-7B from the worst perfor-

*Corresponding Author

mance among 7B models to the best within models of the same parameter size. Experimental results on three benchmarks demonstrate that EchoBat effectively improves the accuracy and hallucination resistance of the base model. By using EchoBat to construct preference pairs of data and training with DPO, the model achieves the best or second best results on multiple datasets.

The contributions can be summarized as follows:

• We propose EchoBat, a new strategy for constructing high-quality preference pairs data and extracting keyframes from videos. With these pairs for preference alignment, hallucinations in VideoMLLMs can be significantly reduced.

• EchoBat demonstrates excellent generalizability and can be seamlessly integrated into existing VideoMLLMs to mitigate hallucinations.

• Extensive experimental results validate the practical effectiveness of EchoBat across multiple hallucination benchmarks. For instance, when applied to LLaVA-Video-7B, EchoBat reduces the hallucinations by 7.9% in VideoHalluce. On EventHallusion, it boosts the Binary and Desc scores by 11.3% and 17.75%, respectively, over the base model. Notably, on the LangDominance metric in CMM (video), EchoBat achieves performance comparable to the closed-source GPT-4o, highlighting its ability to resist Hallucinations.

Related Works

Causes of hallucinations in VideoMLLMs

Hallucinations in VideoMLLMs mainly stem from three aspects: hallucinations from data, from the model, and from modality alignment.

• **Hallucinations from Data.** The quality and distribution of data are primary factors leading to hallucinations. For instance, in factual question answering, an imbalanced data distribution often results in erroneous answers (Liu et al. 2024). Moreover, data homogeneity the model's ability to adapt to different environments (You et al. 2024). The performance of models like MiniGPT-4 and LLaVA highlights that when training lacks diverse learning instructions and visual information, insufficient data can make it difficult for models to describe local visual relationships, thereby resulting in hallucinations.

• **Hallucinations from model.** VideoMLLMs often use CLIP as the visual encoder, which uses contrastive learning to map visual and textual features to a shared space. Despite strong performance in many tasks, CLIP struggles with fine-grained visual details—e.g., limited image resolution hinders detail extraction, though higher resolutions help (Jain, Yang, and Shi 2024; Cho et al. 2022). As another component of VideoMLLMs, LLMs significantly enhance the capability of MLLMs in handling complex tasks but also introduce their inherent hallucinations into vision-language models. Factors such as biases in LLM training data and over-reliance on prior knowledge can lead to hallucinations. Insufficient attention mechanisms (Lee et al. 2024) and errors caused by random sampling during decoding (Chuang et al. 2024) can also cause hallucinations.

• **Hallucinations from modality aligning.** Hallucinations from modality alignment: Multimodal tasks require aligning visual features with the LLM's embedding space, typically via simple alignment modules like linear layers. Their simplicity limits comprehensive multimodal integration, increasing hallucination risks (Sun et al. 2024b; Zhao et al. 2023). For example, Q-Former—a widely used module in models like InstructBLIP and MiniGPT-4—encodes a fixed number of randomly initialized tokens into text-aligned visual features. Yet limited tokens fail to fully capture information, causing loss and higher hallucination risks (Yin et al. 2023; Chen et al. 2023).

While VideoMLLMs have demonstrated strong video understanding capabilities, research on their hallucinations remains nascent. Zhang et al. (Zhang et al. 2024a) proposed the EventHallusion benchmark and Temporal Contrastive Decoding (TCD). Sun et al. (Sun et al. 2024a) targeted hallucinations in long video processing to boost long-term video understanding. We argue that resolving hallucinations is critical for building practical VideoMLLMs. Thus, we propose EchoBat: a method leveraging video data to construct high-quality preference pairs and extract key frames for enhanced visual supervision, aiming to alleviate hallucinations.

Methodology

The overview framework of EchoBat is illustrated in Figure 1, it mainly consists of three key components: Echo-Vision Enhancement, Video reverse non-preference, and Echo-Layered Sampling. The framework can be used to construct high-quality (chosen) and low-quality (rejected) video responses to form preference pairs. In addition, the preference pairs generated by EchoBat can be used to enhance model performance with direct preference optimization (DPO).

Echo-Vision Enhancement

Current VideoMLLMs exhibit two distinct approaches to audio information processing in text generation tasks.

1) Some majority of models like Video-LLaMA (Zhang, Li, and Bing 2023), VideoMAE V2(Wang et al. 2023), InternVL (Chen et al. 2023) entirely discard the audio track, relying solely on visual features for text generation.

2) Some emerging heterogeneous modality unification paradigms like VITA-1.5(Fu et al. 2025), HumanOmni(Zhao et al. 2025) achieve audio comprehension through cross-modal encoders. These models establish audio-visual associations via feature concatenation. Howerver, their design objective focuses on expanding the model's cross-modal understanding scope, which essentially represents a Modality Capacity Expansion approach. Rather than leveraging audio to enhance video description text quality and mitigate hallucinations.

This study focuses on addressing hallucination issues in video-based large models. Inspired by the bats use echo to enhance vision, we innovatively position audio information as a supplementary signal source for visual semantics. By leveraging audio cues in videos (e.g., ambient sounds, voice narration) that contain semantically rich clues highly correlated with visual content, our approach effectively enhances

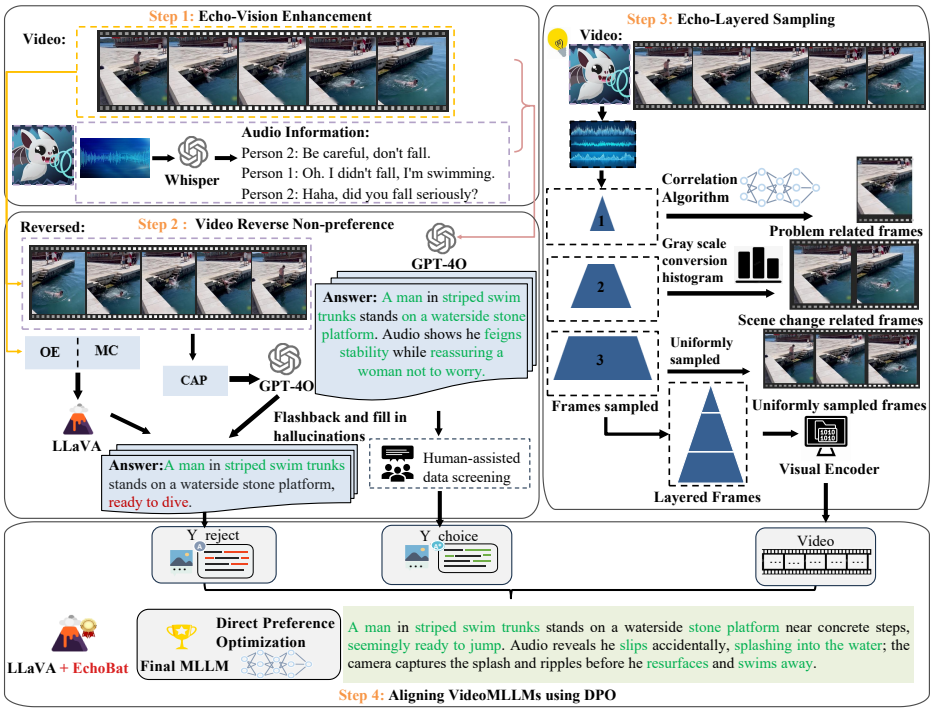

Figure 1: Overview of the proposed EchoBat framework. The text of correct answers and hallucinations are highlighted in green and red, respectively.

the factuality, accuracy, and completeness of the generated high-quality (chosen) Caption (CAP) text. As illustrated in Step 1 in Figure 1, the proposed method consists of four core processing stages:

1) **Cross-modal Feature Decoupling**: The audio and visual streams are decoupled using FFmpeg, followed by denoising and sample rate normalization preprocessing on the audio track.

2) **Temporally Aligned Semantic Extraction**: The Whisper(Cao et al. 2012) is employed for automatic speech recognition (ASR), generating synchronized sentence-level transcripts with corresponding timestamps.

3) **Vision-Guided Semantic Filtering**: Video frames are uniformly sampled and a cross-modal alignment model (SigLIP) is used to compute semantic similarity scores between each frame and individual audio-derived text segments as shown in eq. 1.

$$\phi(s_i, v_t) = \text{Sigmoid}\left(E_T(s_i)^\top E_V(v_t)\right) \tag{1}$$

where s_i denotes the textual representation of the i-th audio sentence, v_t represents the visual feature of the video frame at time t, $E_T(\cdot)$ and $E_V(\cdot)$ denote the text and image encoders, respectively, T indicates matrix transposition, and $sigmoid$ refers to the nonlinear activation function.

Given that video captioning tasks require mining latent semantic correlations rather than enforcing strict matching, theoretical analysis and experimental validation led us to set the activation threshold τ for semantic supplementary signals at 0.4. This threshold corresponds to the median of weakly correlated scenarios ($\phi(s_i, v_t) \in [0.3, 0.5)$), effectively capturing cross-modal latent associations while mitigating noise interference. This configuration aligns with findings by SPECTRUM (Faghihi, Zarenejad, and Beheshti Shirazi 2024) in multimodal semantic and sentiment analysis, where their study demonstrated that $\tau = 0.4$ achieves an optimal trade-off for cross-modal associations.

4) **Audio-Visual Text Generation**: Employing a predefined prompt template, GPT-4o is utilized to integrate the filtered audio transcripts into the video captioning text based on content coherence and timestamp alignment, ensuring temporal consistency between generated descriptions and video sequences. (See Appendix for detailed prompt template specifications.)

5) **Example Demonstration**: For example, a video about cooking in the kitchen.

{Visual Description}: "A person holds a white powdered container, pouring granular substance into a cooking pot."

{Audio Description}: "Now we need to add approximately 2 grams of salt to the boiled peas."

{Audio-Visual merged Description} (chosen): "The operator is pouring 2 grams of white salt into the pot containing boiled peas."

Video reverse non-preference

Extending ARA(Qu et al. 2024), we observe the degree of VideoMLLMs on visual input through active queries. Specifically, we define a metric based on the mutual infor-

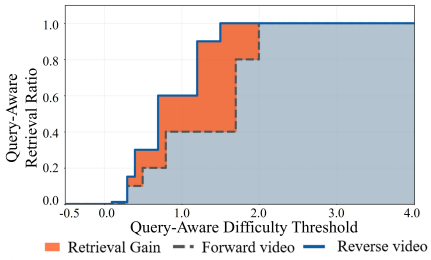

Figure 2: The difference between LLaVA-Video triggering active queries in two situations on LLaVA-Video-178K.

mation methods:

$$M_{ij} = \log \frac{P\left(a_{ij}|V_i, Q_i\right)}{P\left(a_{ij}|Q_i\right)} = \log \frac{P\left(a_{ij}|V_i, Q_i\right)}{P\left(a_{ij}\right)} - \log \frac{P\left(a_{ij}|Q_i\right)}{P\left(a_{ij}\right)}. \quad (2)$$

where $P(a_{ij}|Q_i)$ represents the output probability of VideoMLLMs based solely on the query input, while $P(a_{ij}|V_i, Q_i)$ represents the output probability when both the query and the video are used as inputs. A higher value of M_{ij} indicates that more visual reasoning is required to answer the query. When $M_{ij} < 0$, it suggests that the answer relies too heavily on the input query and ignores the current visual input.

As shown in Figure 2, we conducted experiments on the LLaVA-Video-178K dataset, sampling 10,000 questions, including multiple-choice (MC), open-ended (OE), and caption generation (CAP) tasks. The difficulty of the questions was rated on a scale of 0–4. By analyzing the proportion of queries triggered by the model in forward and reverse videos, we observed that $M_{ij} < 0$ appears more frequently and earlier in reverse videos. This indicates that the model relies more on pre-existing prior knowledge and tends to ignore the current visual input when playing the video in reverse. We hypothesize that this may be due to the presence of many uncommon phenomena in reversed videos, which also make the video timeline more difficult for the model to comprehend.

Therefore, our pipeline for constructing low-quality (rejected) Caption (CAP) text comprises two key components:

1) **Inverse Physical Law Modeling**: We employ a baseline VideoMLLM to generate initial descriptions D_{init} for reversed videos V_{rev}. Since the model's parameter space Θ lacks temporal constraints, its decoding process satisfies:

$$D_{\text{init}} = \arg\max_{D} P(D|V_{\text{rev}}; \Theta) \quad \text{s.t.} \quad \Theta \cap \Phi_{\text{time}}^{-1} = \emptyset \quad (3)$$

where Φ_{time}^{-1} denotes the set of time-reversed physical laws.

Due to the model's deficient capability in modeling time-reversed physical phenomena, its outputs predominantly manifest three categories of hallucinations(Zhang, Li, and Bing 2023): Temporal incoherence (e.g., "water flows upward"), Physical law violations (e.g., "food automatically defrosts"), and Visual attribute errors (e.g., positional shifts, color mismatches). According to the ARA(Qu et al. 2024), the first two categories stem from the model's over-reliance on prior knowledge distribution $P_{prior}(D)$ while disregarding current visual evidence E_v, resulting in anomalous descriptions that violate causality (e.g., "water flows from the ground to clouds"). The third category directly reflects inadequate modeling of the visual evidence E_v.

2)**Text Revision and Hallucination Injection**: We construct a deterministic hallucination transformation template T_h using GPT-4o to map the initial description D_i to the forward time axis while injecting controlled hallucinations:

$$D_{\text{rej}} = T_{\text{h}}(D_{\text{i}}) = \text{GPT4o}\left(D_{\text{i}}, \mathcal{T}_{\text{template}}(\lambda, (\alpha, \beta, \gamma, \delta))\right) \quad (4)$$

where λ represents the flashback functional component. The other parameters of the template $\mathcal{T}_{\text{template}}$ are as follows. Where α is temporal errors, β is physical violations, γ is procedural omissions, and δ is perceptual inaccuracies:

$$\begin{cases} \alpha = 27\% & \text{Temp. errors} \\ \beta = 23\% & \text{Phys. violations} \\ \gamma = 40\% & \text{Proc. omissions} \\ \delta = 10\% & \text{Percep. inaccuracies} \end{cases} \quad \sum_{\alpha,\beta,\gamma,\delta} = 1 \quad (5)$$

We ensure the generated texts achieve an optimal balance between grammatical correctness and semantic absurdity. This standard adheres to the adversarial sample quality criteria proposed by Ebrahimi et al.(Ebrahimi, Lowd, and Dou 2018).

3) **Example Demonstration**: For example, a video about cooking in the kitchen.

{Reversed Video Description}: "white salt particles automatically jump back from the pot of cooked peas into the seasoning container."

{GPT-4o Flashback Description}: "white salt particles spontaneously move from the seasoning container into the pot of cooked peas."

{GPT-4o Hallucinate Filler Description} (rejected): "white salt particles fly out from the seasoning container, turn red mid-air, remain suspended motionless, and finally jump into the pot of cooked peas autonomously."

Notably, Our experiments reveal that GPT-4o demonstrates unique capabilities in processing reversed text and filling in the hallucination - functionalities notably absent in current open-source models. The corresponding prompt templates have been listed in the Appendix, which we identify as prerequisite for acquiring cap's non-preferenced data.

Echo-Layered Sampling

Inspired by the multi-frequency pulse mechanism of bat echolocation, we propose an Echo-Layered Sampling strategy comprising three hierarchical levels:

1) **Top-level**: Semantic-correlated Frame Sampling (simulating high-frequency short pulses for precise target localization)

2) **Middle-level**: Scene-transition Frame Sampling (corresponding to medium-frequency pulses' environmental scanning characteristics)

3) **Base-level**: Uniform Complementary Frame Sampling (adapting low-frequency long pulses' background cruising function) This hierarchical sampling framework captures

Algorithm 1: Algorithm of Echo-Layered Sampling

Input: V, Q, N_{target}
Output: $\mathcal{K}$, $\mathcal{T}$
1. Read video information:
$\mathcal{F}, f_{\text{ps}} \leftarrow \text{ReadVideo}(V)$
2. Semantic Sampling:
$\mathcal{K}_{\text{sem}} \leftarrow \{ \underset{e \in \text{Entities}(Q)}{\text{argmax}} \ \text{SemanticScore}(\mathcal{F}, e)\}$
3. Scene Sampling:
$\mathcal{K}_{\text{vis}} \leftarrow \{i \mid \Delta H(\mathcal{F}[i], \mathcal{F}[i+1]) > \tau_{\text{hist}}\}$
4. Merge & Supplement:
$\mathcal{K} \leftarrow \text{UniqueMerge}(\mathcal{K}_{\text{sem}}, \mathcal{K}_{\text{vis}})$
If $|\mathcal{K}| < N_{\text{target}}$ then
$\mathcal{K} \leftarrow \mathcal{K} \cup \text{UniformSample}(\mathcal{F}, N_{\text{target}} - |\mathcal{K}|)$
5. Return result:
$\mathcal{K}[1 : N_{\text{target}}], \{t_i / f_{\text{ps}} \mid t_i \in \mathcal{K}\}$

multi-granularity visual information, delivering more accurate supervisory signals than conventional uniform sampling while enhancing model training efficacy and suppressing hallucinations. The detailed process is summarized in Algorithm 1.

1) **Input and Output Description**: The proposed algorithm takes as input a video V and a natural language query Q, where V represents the video stream to be processed and Q describes the target semantic content (e.g., "how to make scrambled eggs"). The target number of keyframes is set by the parameter $N_{target} = 64$. The output consists of a keyframe index set $\mathcal{K} \subseteq \{1, \ldots, |\mathcal{F}|\}$ and their corresponding timestamps $\mathcal{T} \in \mathbb{R}^+$, satisfying $|\mathcal{K}| = N_{target}$ and $\mathcal{T}[i] = \mathcal{K}[i]/fps$, where fps denotes the video frame rate.

2) **Semantic-correlated Frame Sampling**: The semantic-level sampling module achieves text-query-based keyframe localization through cross-modal alignment by first extracting a set of noun entities ε from the natural language query Q, then computing the semantic similarity score $SemanticScore(\mathcal{F}[i], e)$ between each video frame $\mathcal{F}[i]$ and entity $e \in \varepsilon$ using the pretrained SigLIP, and finally retaining the most semantically similar frame for each entity to form the initial semantic keyframe set $\mathcal{K}_{sem}$, while enforcing a minimum temporal interval $\Delta_{\min} = 0.5fps$ to prevent excessive clustering of semantic frames and ensure uniform temporal distribution of keyframes.

3) **Scene-transition Frame Sampling**: The scene sampling module supplements semantically uncovered frames by detecting abrupt visual content changes through 3D histogram difference (ΔH) computation of consecutive frames in LAB color space, where scene transition points are identified when ΔH exceeds the dynamic threshold $\tau_{\text{hist}} = \mu_H + 0.5\sigma_H$, with a compensation mechanism automatically triggered for drastic changes ($\Delta H > 2\tau_{\text{hist}}$) to increase sampling density in adjacent intervals, ultimately generating a visual keyframe set $\mathcal{K}_{vis}$ that effectively captures semantically independent but visually salient changes including shot transitions and rapid object movements.

4) **Uniform Frame Sampling**: The merging and supplementation module generates the final keyframes through priority-based fusion and adaptive sampling by first combining $\mathcal{K}_{sem}$ and $\mathcal{K}_{vis}$ following semantic priority principles, then removing duplicate frames, and subsequently performing uniform supplementation at fixed intervals $\left\lfloor \frac{|\mathcal{F}|}{N_{\text{target}} - |\mathcal{K}|} \right\rfloor$ from remaining frames if the total count falls below N_{target}, ultimately producing the output keyframe set $\mathcal{K}$ that strictly satisfies $|\mathcal{K}| = N_{target}$ with precisely aligned timestamps $\mathcal{T}$ obtained by dividing frame indices by the frame rate fps to ensure exact temporal correspondence with the original video timeline.

We validate the reviewer's concern that lower thresholds cause noise infiltration: $\tau = 0.2$ allows 7.3% irrelevant audio retention (e.g. background music misaligned as dialogue). $\tau = 0.6$ over-filters critical cues, reducing EventHallusion accuracy by 4.1%. Our $\tau = 0.4$ achieves optimal balance: 98.7% noise filtering while retaining essential audio semantics.

Inference fairness: All benchmarks use silent videos (audio disabled).

To ensure fair comparison with LLaVA-Video's baseline, we maintain $N_{\text{target}} = 64$ while introducing crucial dynamic improvements through our Echo-Layered Sampling (ELS). The system further supports dynamic scaling via adaptive N_{target} adjustment (e.g. extending to $N = 128$ for 10-minute videos through Δ_{min} modification). We employ the default $N_{\text{target}} = 64$ configuration to maintain optimal efficiency.

Experiments

We analyze experimental results of VideoMLLMs trained with EchoBat on the latest video hallucination benchmarks. These benchmarks evaluate VideoMLLMs, metrics like Perception Accuracy (PA) and Hallucination Resistance (HR), covering dimensions including Object-Relation, Temporal, Semantic Detail, Extrinsic Fact, Entirety, and Interleaving.

We also validate each EchoBat component's effectiveness and provide examples showing reduced hallucinations after training with EchoBat.

Experimental Setups

Evaluation benchmarks and metrics. For assessing VideoMLLMs, we selected three recent benchmarks: VideoHallucer (Wang et al. 2024), EventHallusion (Zhang et al. 2024b), and CMM. These benchmarks represent the latest developments in evaluating hallucinations in VideoMLLMs.

(1) **VideoHallucer** is the first benchmark dedicated to evaluating hallucinations in VideoMLLMs. It categorizes hallucinations into intrinsic and extrinsic types, utilizing data from public datasets such as ActivityNet, VidOR, and EDUVSUM. Evaluation is performed through binary VideoQA.

(2) **EventHallusion** assesses event-related hallucinations by analyzing models' reliance on prior knowledge versus video content. It incorporates rare event videos to examine image priors and misleading descriptions of common events to evaluate language priors. Performance is measured via binary classification and detailed description matching.

(3) **CMM** is a comprehensive benchmark for evaluating multimodal hallucinations across language, vision, and

Model	Size	VideoHallucer			EventHallusion							CMM			
		Acc.			Ent		Mix		Mis	Ovr		VLCorr		LD	
		Bas	Hall	Ovr	Bin	Det	Bin	Det	Bin	Bin	Det	Prec	Rec	pa	hr
VideoChat2	7B	29.7	25.8	7.8	16.67	4.59	47.67	1.55	22.55	32.76	2.64	**97.0**	66.0	88.0	34.5
PLLaVA	7B	75.1	55.5	38.1	45.61	16.51	58.55	3.11	81.37	60.64	6.05	89.5	93.0	75.0	52.0
ShareGPT4Video	8B	88.5	20.0	15.8	11.40	0.00	67.88	5.18	6.86	49.14	9.82	87.5	85.5	79.5	58.0
Video-LLaMa-2	7B	**90.9**	12.7	10.0	30.7	8.26	57.6	7.25	41.8	45.97	7.62	75.0	86.0	71.0	54.0
HOUND-DPO	7B	83.4	43.0	29.5	36.1	10.2	16.1	9.5	64.6	33.1	10.2	70.0	74.0	68.0	49.0
ISR-DPO($\pi\theta 9$)	7B	88.9	50.7	45.1	44.3	12.2	54.7	10.4	93.7	65.2	12.9	91.0	85.0	81.0	86.0
Qwen-2.5-VL	7B	89.7	52.3	46.1	46.5	20.6	65.2	34.9	90.2	67.2	41.3	96.0	88.0	88.0	93.0
Gemini-1.5-pro	-	83.6	42.3	37.8	61.40	49.54	83.42	39.90	96.08	80.44	43.38	91.0	90.5	78.5	61.5
LLaVA-Onevision	7B	86.5	60.9	51.9	46.78	18.34	57.51	23.31	88.42	61.96	34.00	94.0	90.0	87.0	72.0
+ EchoBat	7B	88.4	62.3	53.2	47.28	24.46	62.37	23.73	88.92	64.94	36.59	94.5	92.5	88.5	73.5
LLaVA-Video	7B	87.8	54.2	45.9	52.42	17.47	52.85	33.60	96.84	63.60	40.67	87.0	88.5	85.0	74.0
+ EchoBat	7B	90.4	58.5	47.2	55.13	33.03	62.69	37.75	98.89	70.77	47.89	92.0	89.5	**89.0**	76.0
GPT-4o	-	75.1	**74.2**	**53.3**	**83.33**	**58.72**	**92.75**	**41.97**	**100.0**	**91.93**	**48.01**	87.5	**95.5**	83.0	**84.0**

Table 1: Main experimental results. The best and second best results are indicated in bold and underlined, respectively.

audio domains. In this work, we utilize its video evaluation component, which employs noise injection to assess models' overreliance on unimodal priors and co-occurrence frequency manipulation to quantify the impact of spurious cross-modal correlations. The benchmark provides systematic evaluation through two key metrics: Perception Accuracy (PA) and Hallucination Resistance (HR). Our experiments specifically employ the Visual-Language partition, denoted as CMM (Video), for assessment.

Experimental details. To ensure fair comparisons, we maintained all original hyper-parameters, frame numbers, and frame sampling strategies for all models. Responses are generated using greedy search. For EchoBat, we constructed 16438 pairs of preference data and aligned LLaVA-Video to human preferences using Direct Preference Optimization (DPO). The learning rate was set to $5e^{-7}$, beta to 0.1, and batch size to 1. Preference data generation took 20 hours, while training LLaVA-Video using EchoBat required 8 hours. The hardware setup included 8×A800 80GB GPUs.

Main Results

The main experimental results are summarized and presented in Table 1. From the table, we can see that EchoBat significantly improved the reliability of the original LLaVA-video model, achieving the most advanced reliability scores across many evaluation benchmarks at the 7B size.

Specifically, the EchoBat significantly reduced the hallucination rates of LLaVA-video on the VideoHalluce, with relative point improvement of 7.9%. On the EventHallusion, the overall Binary and Desc scores increased by 11.3% and 17.75% respectively compared to LLaVA-Video-7B. It is worth noting that the Desc score on the Entire subset has been greatly improved, up to 89.1%. We believe that this is because after training with EchoBat, the model is more sensitive to entities in the description. On the CMM (video), LLaVA-Video-EchoBat's LangDominance index is very high, remaining in the top two, and is comparable to the closed-source model GPT-4o. This index reflects whether the model can balance language knowledge and multimodal input to avoid over-reliance on LLM priors. This shows that EchoBat does suppress LLM priors through method such as Video reverse non-preference.

Model	MSVD	MSR	VideoChatGPT-Captioning				
	Acc.	Acc.	Corr	Det	Ctx	Temp	Cons
HOUND-DPO	80.7	70.2	3.0	2.7	3.3	2.0	2.6
ISR-DPO($\pi\theta 9$)	80.4	**75.4**	3.2	2.9	3.6	2.6	2.7
LLaVA-video	82.1	69.7	3.0	2.8	3.4	2.2	2.4
+ EchoBat	**83.7**	71.2	**3.5**	**3.0**	**3.8**	**2.8**	**2.9**

Table 2: Model performance on various datasets

In addition, We have rigorously evaluated EchoBat's general capabilities using general capability evaluation benchmarks,like: General Video QA: MSVD ,MSR-VTT, Captioning Performance: VideoChatGPT-Bench. EchoBat's preference optimization both reduces hallucinations and boosts overall performance, notably achieving +16% average gains in captioning tasks—consistent with its Desc. improvements in LLaVA (Table 2).

Ablation Study & Analysis

We conducted a deeper analysis of EchoBat by deconstructing its components. To evaluate the efficacy of its three modules—Echo-Vision Enhancement (EVE), Echo-Layered Sampling (ELS), and Video Reverse Non-preference (VRN)—we compared them against the original LLaVA-video model on the VideoHallucer and CMM(video) benchmarks. As shown in Table 3, demonstrate that all three components contribute significantly to the model's performance.

To validate the effectiveness of EVE, we compared models fine-tuned on the original LLaVA-video using only GPT-4o responses with models aligned through a preference-pair approach. All other settings were kept constant for a fair comparison. To evaluate the effectiveness of ELS, we assessed models with and without keyframes selection incorporated into the training process.

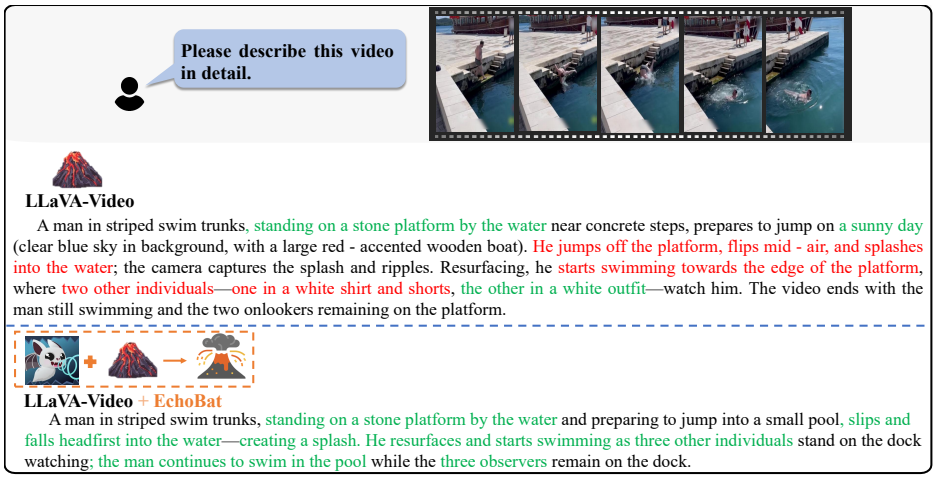

Figure 3: Qualitative comparison between LLaVA-Video + EchoBat and LLaVA-Video. Correct answers and hallucinations are highlighted in green and red, respectively.

Modules	VideoHallucer			CMM			
	Acc.			VLCorr		LD	
	Bas	Hall	Ovr	pa	hr	pa	hr
-	87.8	54.2	45.9	87.0	88.5	85.0	74.0
EVE	89.1	56.0	46.3	89.0	88.5	87.5	75.0
EVE+ELS	90.2	58.1	46.9	90.5	89.0	88.0	75.5
EVE+ELS+VRN	90.4	58.5	47.2	92.0	89.5	89.0	76.0

Table 3: Ablation of LLaVA-Video with EchoBat Modules.

As shown in Table 3, the model trained with EchoBat achieved the lowest hallucination rate and the highest accuracy. We believe this can be attributed to the following two points. **1) The EVE improves the quality of preference data pairs.** In anomalous scenarios, the model is prone to generating preemptive hallucinations. By human preference alignment, this phenomenon is effectively suppressed, resulting in a significant reduction in model's hallucinations. **2) The ELS provides stronger visual supervisory signals.** The model better understands scene changes, thereby enhancing its ability to answer visual questions accurately.

Furthermore, we observed that the EVE performs more significantly in the VideoHallucer benchmark. This may be attributed to the higher number of open-ended questions and captions within VideoHallucer. In contrast, for the CMM benchmark VLCorrelations index, the ELS had a greater impact on model performance. This is likely because the benchmark consists many of binary yes-no questions, most of which can be answered using single or multiple frames from the video.

Finally, we found that EchoBat demonstrated generalizability. As shown in Table 1, we also trained additional model, including LLaVA-OneVision. Results indicate that applying the EchoBat effectively reduces hallucinations on various benchmarks in VideoMLLM, demonstrating the generalizability of EchoBat.

Example Demonstration

To intuitively display and compare the performance differences between different models, we present qualitative analysis result in Fig. 3. And we provide more results in the appendix.

In this Cap task, both models accurately describe the environment in which the activities in the video occur, but LLaVA-Video misjudges the behavior of the character. Our model correctly identified that the man here slipped and fell into the water, not that he actively rotated in the air and then dived into the water. In addition, compared with LLaVA Video, our model was able to recognize that there were three people on the platform (a man, a woman, and a baby), and the model's hallucination is partially alleviated.

Conclusion

Hallucinations remain a critical challenge for VideoM-LLMs, limiting their reliable deployment in real-world applications. To mitigate this problem, we introduce EchoBat, a novel strategy for constructing high-quality preference data pairs and selecting keyframes from video. This approach enables the model to use audio information to assist in understanding video content, while not preferring flashback reverse video content that is full of hallucinations, thereby enhancing its ability to understand the video content and temporal logical structure of the videos. Suppressing hallucinations driven by the model's internal prior knowledge. Additionally, the algorithm also includes an echo-layered sampling strategy for selecting keyframes, which provides more effective visual supervision signals. By aligning model behavior with human preferences using the DPO, we effectively alleviate hallucinations in VideoMLLMs. In future work, we aim to explore how to combine the above methods to further enhance the understanding ability of VideoMLLM and reduce hallucinations in long videos.

Acknowledgments

This research is supported by the National Key Research and Development Program of China (2023YFB3107401), Shaanxi Provincial Key Research and Development Program (No. 2023YFB3107401), the National Natural Science Foundation of China (62521002, 62376210, 62161160337, 62132011, U24B20185, U21B2018, 62206217).

References

Cao, N.; Lin, Y.; Sun, X.; Lazer, D.; Liu, S.; and Qu, H. 2012. Whisper: Tracing the Spatiotemporal Process of Information Diffusion in Real Time. *IEEE Trans. Vis. Comput. Graph.*, 18(12): 2649–2658.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; Li, B.; Luo, P.; Lu, T.; Qiao, Y.; and Dai, J. 2023. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. *CoRR*, abs/2312.14238.

Chirila, R.; Dahiya, A. S.; Schyns, P. G.; and Dahiya, R. 2024. Self-Powered Multimodal Sensing Using Energy-Generating Solar Skin for Robotics and Smart Wearables. *Adv. Intell. Syst.*, 6(7).

Cho, J.; Yoon, S.; Kale, A.; Dernoncourt, F.; Bui, T.; and Bansal, M. 2022. Fine-grained Image Captioning with CLIP Reward. In Carpuat, M.; de Marneffe, M.; and Ruíz, I. V. M., eds., *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, 517–527. Association for Computational Linguistics.

Choi, S. H.; Kim, M.; and Lee, J. Y. 2025. Smart and user-centric manufacturing information recommendation using multimodal learning to support human-robot collaboration in mixed reality environments. *Robotics Comput. Integr. Manuf.*, 91: 102836.

Chuang, Y.; Xie, Y.; Luo, H.; Kim, Y.; Glass, J. R.; and He, P. 2024. DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Ebrahimi, J.; Lowd, D.; and Dou, D. 2018. On Adversarial Examples for Character-Level Neural Machine Translation. In Bender, E. M.; Derczynski, L.; and Isabelle, P., eds., *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, 653–663. Association for Computational Linguistics.

Faghihi, E.; Zarenejad, M.; and Beheshti Shirazi, A. 2024. SPECTRUM: Semantic Processing and Emotion-informed video-Captioning Through Retrieval and Understanding Modalities. *CoRR*, abs/2411.01975.

Fu, C.; Lin, H.; Wang, X.; Zhang, Y.; Shen, Y.; Liu, X.; Cao, H.; Long, Z.; Gao, H.; Li, K.; Ma, L.; Zheng, X.; Ji, R.; Sun, X.; Shan, C.; and He, R. 2025. VITA-1.5: Towards GPT-4o Level Real-Time Vision and Speech Interaction. *CoRR*, abs/2501.01957.

Govindasamy, R.; Nagarajan, S. K.; Muthu, J. R.; and Ramkumar, M. 2025. Residual multiscale attention based modulated convolutional neural network for radio link failure prediction in 5G. *Ad Hoc Networks*, 166: 103679.

Huo, R.; Chen, J.; Zhang, Y.; and Gao, Q. 2025. 3D skeleton aware driver behavior recognition framework for autonomous driving system. *Neurocomputing*, 613: 128743.

Jain, J.; Yang, J.; and Shi, H. 2024. VCoder: Versatile Vision Encoders for Multimodal Large Language Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, 27992–28002. IEEE.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.; Madotto, A.; and Fung, P. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, 55(12): 248:1–248:38.

Lee, S.; Park, S. H.; Jo, Y.; and Seo, M. 2024. Volcano: Mitigating Multimodal Hallucination through Self-Feedback Guided Revision. In Duh, K.; Gómez-Adorno, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, 391–404. Association for Computational Linguistics.

Liu, F.; Lin, K.; Li, L.; Wang, J.; Yacoob, Y.; and Wang, L. 2024. Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Qu, X.; Chen, Q.; Wei, W.; Sun, J.; and Dong, J. 2024. Alleviating Hallucination in Large Vision-Language Models with Active Retrieval Augmentation. *CoRR*, abs/2408.00555.

Sun, Y.; Liu, Z.; Liu, C.; Pu, B.; Zhang, Z.; and Xie, H. 2024a. Hallucination Mitigation Prompts Long-term Video Understanding. *CoRR*, abs/2406.11333.

Sun, Z.; Shen, S.; Cao, S.; Liu, H.; Li, C.; Shen, Y.; Gan, C.; Gui, L.; Wang, Y.; Yang, Y.; Keutzer, K.; and Darrell, T. 2024b. Aligning Large Multimodal Models with Factually Augmented RLHF. In Ku, L.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, 13088–13110. Association for Computational Linguistics.

Traum, D.; Skantze, G.; Nishizaki, H.; Higashinaka, R.; Minato, T.; and Nagai, T. 2024. Special issue on multimodal processing and robotics for dialogue systems (Part II). *Adv. Robotics*, 38(4): 193–194.

Wang, L.; Huang, B.; Zhao, Z.; Tong, Z.; He, Y.; Wang, Y.; Wang, Y.; and Qiao, Y. 2023. VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 14549–14560. IEEE.

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

Xie, Z.; He, W.; Xu, T.; Wu, S.; Zhu, C.; Yang, P.; and Chen, E. 2023. Comprehending the Gossips: Meme Explanation in Time-Sync Video Comment via Multimodal Cues. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 22(8): 216:1–216:17.

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

You, H.; Zhang, H.; Gan, Z.; Du, X.; Zhang, B.; Wang, Z.; Cao, L.; Chang, S.; and Yang, Y. 2024. Ferret: Refer and Ground Anything Anywhere at Any Granularity. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Zhang, H.; Li, X.; and Bing, L. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. In Feng, Y.; and Lefever, E., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 - System Demonstrations, Singapore, December 6-10, 2023*, 543–553. Association for Computational Linguistics.

Zhang, J.; Jiao, Y.; Chen, S.; Chen, J.; and Jiang, Y. 2024a. EventHallusion: Diagnosing Event Hallucinations in Video LLMs. *CoRR*, abs/2409.16597.

Zhang, J.; Jiao, Y.; Chen, S.; Chen, J.; and Jiang, Y. 2024b. EventHallusion: Diagnosing Event Hallucinations in Video LLMs. *CoRR*, abs/2409.16597.

Zhang, L.; Yang, K.; Han, Y.; Li, J.; Wei, W.; Tan, H.; Yu, P.; Zhang, K.; and Yang, X. 2025. TSD-DETR: A lightweight real-time detection transformer of traffic sign detection for long-range perception of autonomous driving. *Eng. Appl. Artif. Intell.*, 139: 109536.

Zhao, J.; Yang, Q.; Peng, Y.; Bai, D.; Yao, S.; Sun, B.; Chen, X.; Fu, S.; chen, W.; Wei, X.; and Bo, L. 2025. HumanOmni: A Large Vision-Speech Language Model for Human-Centric Video Understanding. *CoRR*, abs/2501.15111.

Zhao, Z.; Wang, B.; Ouyang, L.; Dong, X.; Wang, J.; and He, C. 2023. Beyond Hallucinations: Enhancing LVLMs through Hallucination-Aware Direct Preference Optimization. *CoRR*, abs/2311.16839.