Transferability of Adversarial Attacks in Video-based MLLMs: A Cross-modal Image-to-Video Approach

Linhao Huang[1,2,3*], **Xue Jiang**[3,4*], **Zhiqiang Wang**[5*], **Wentao Mo**[1,3], **Xi Xiao**[1,2†], **Yong-Jie Yin**[6], **Bo Han**[4], **Feng Zheng**[3†]

[1]Shenzhen International Graduate School, Tsinghua University,
[2]Peng Cheng Laboratory, Shenzhen, Guangdong, China
[3]Southern University of Science and Technology,
[4]TMLR Group, Hong Kong Baptist University,
[5]Hong Kong University of Science and Technology,
[6]China Electronics Corporation
{hlh23, mow10}@mails.tsinghua.edu.cn, {csxjiang, bhanml}@comp.hkbu.edu.hk, zwangmk@connect.ust.hk, xiaox@sz.tsinghua.edu.cn, yinyongjie@mail.bnu.edu.cn, f.zheng@ieee.org

Abstract

Video-based multimodal large language models (V-MLLMs) have shown vulnerability to adversarial examples in video-text multimodal tasks. However, the transferability of adversarial videos to unseen models—a common and practical real-world scenario—remains unexplored. In this paper, we pioneer an investigation into the transferability of adversarial video samples across V-MLLMs. We find that existing adversarial attack methods face significant limitations when applied in black-box settings for V-MLLMs, which we attribute to the following shortcomings: (1) lacking generalization in perturbing video features, (2) focusing only on sparse key-frames, and (3) failing to integrate multimodal information. To address these limitations and deepen the understanding of V-MLLM vulnerabilities in black-box scenarios, we introduce the Image-to-Video MLLM (I2V-MLLM) attack. In I2V-MLLM, we utilize an image-based multimodal large language model (I-MLLM) as a surrogate model to craft adversarial video samples. Multimodal interactions and spatiotemporal information are integrated to disrupt video representations within the latent space, improving adversarial transferability. Additionally, a perturbation propagation technique is introduced to handle different unknown frame sampling strategies. Experimental results demonstrate that our method can generate adversarial examples that exhibit strong transferability across different V-MLLMs on multiple video-text multimodal tasks. Compared to white-box attacks on these models, our black-box attacks (using BLIP-2 as a surrogate model) achieve competitive performance, with average attack success rate (AASR) of 57.98% on MSVD-QA and 58.26% on MSRVTT-QA for Zero-Shot VideoQA tasks, respectively.

Extended version — https://arxiv.org/pdf/2501.01042v3

1 Introduction

Recent work has shown that video-based multimodal large language models (V-MLLMs) are vulnerable to adversarial video samples (Li et al. 2024a), even though they have achieved remarkable performance on a wide range of video-text multimodal tasks (Li et al. 2024b; Jin et al. 2024; Lin et al. 2024; Maaz et al. 2024; Dai et al. 2023; Zhang, Li, and Bing 2023). Existing work primarily focuses on white-box attacks, where information about the target model is accessible. However, the transferability of adversarial video samples across V-MLLMs remains unexplored, which is a more common and practical setting in real-world scenarios. It is still uncertain whether the adversarial videos generated on the source model can effectively attack other target models, posing significant security risks to the deployment of V-MLLMs in real-world applications.

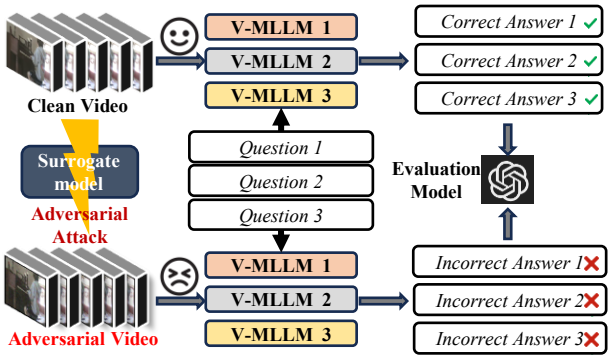

Figure 1: An example of transferable adversarial attack on different target V-MLLMs for Zero-Shot VideoQA task.

In this paper, we pioneer an investigation into the transferability of adversarial video samples across V-MLLMs. Through detailed analysis in Sec. 3.2, we think previous methods have these shortcomings: (1) lacking generalization in perturbing video features, (2) focusing only on sparse key-frames, and (3) failing to integrate multimodal information. FMM attack (Li et al. 2024a) is the first proposed white-box attack method targeting V-MLLMs. It utilizes flow-based temporal mask to select key-frames and applies perturbations to these frames. FMM attack performs well in

*Equal contribution.
†Corresponding author.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

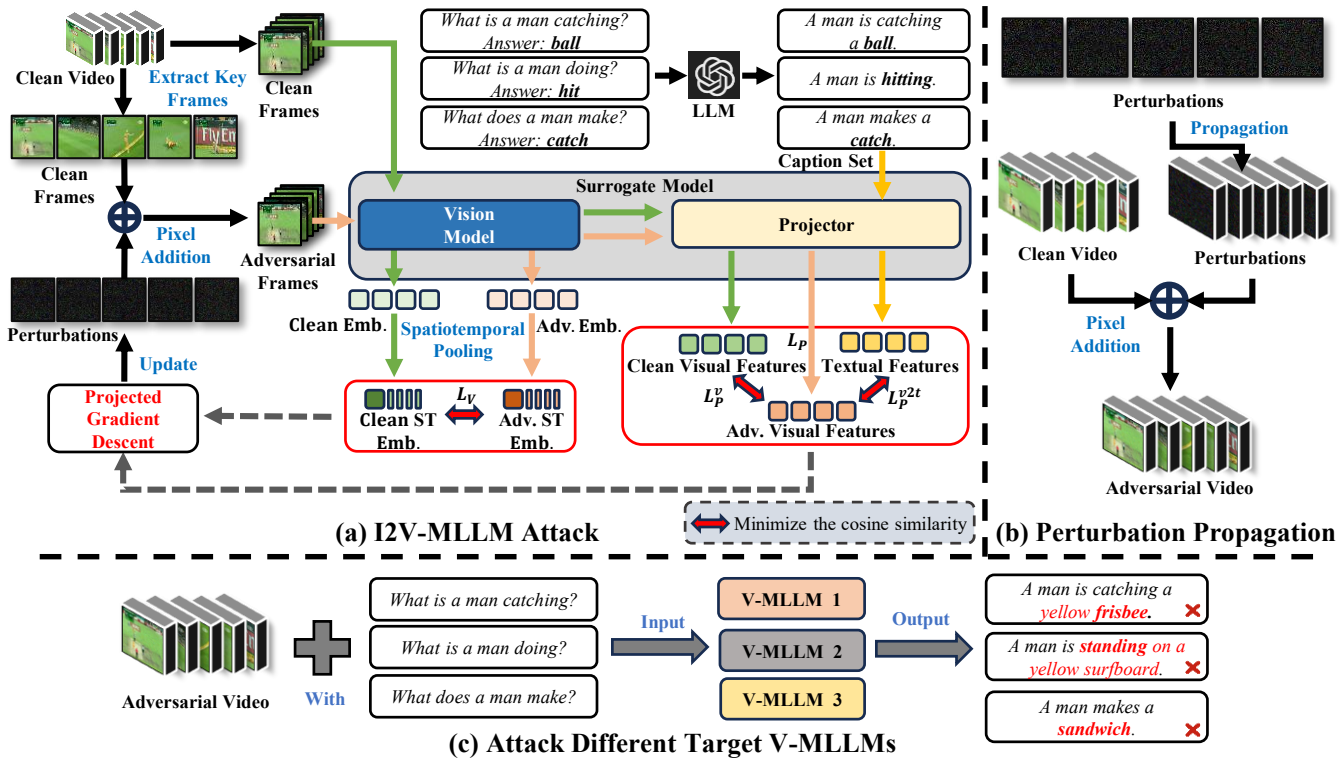

Figure 2: Overview of our proposed method. (a) **I2V-MLLM Attack:** The clean video is divided into K clips. Key frames are extracted from these clips to form the clean frames X, which is then fed into the vision model to extract clean frame-level embeddings $F_V(X)$. These embeddings are subsequently aggregated via spatiotemporal pooling to obtain clean spatiotemporal embeddings $F_V^{st}(X)$. Perturbations are initialized and added to clean frames X to generate adversarial frames X_{adv}. The same process is applied to extract $F_V(X_{adv})$ and $F_V^{st}(X_{adv})$. An LLM reformulates the QA pairs into a caption set T. $F_V(X)$, $F_V(X_{adv})$, and T are then passed through the projector to extract visual features $F_P^v(X)$, adversarial visual features $F_P^v(X_{adv})$, and textual features $F_P^t(T)$, respectively. Perturbations are updated via the PGD method by minimizing three cosine similarity-based losses: L_V, L_P^v, and L_P^{v2t}. (b) **Perturbation Propagation:** The final perturbations applied to key-frames are propagated back to their corresponding video clips to construct the adversarial video. (c) **Attack Different Target V-MLLMs.**

the white-box setting but has limited transferability in the black-box setting. FMM attack heavily relies on the video features, which causes the generated perturbations to overfit to the video features extracted by the surrogate model, thereby reducing their generalizability. Additionally, since FMM attack applies perturbations only to key-frames, it cannot ensure that all frames sampled by the target model are perturbed. Taking low-level image features into account can help with improving transferability of adversarial samples. Previous image-to-video cross-modal attacks (Wei et al. 2021; Wang, Guo, and Wang 2023; Kim et al. 2023) demonstrate the possibility of using image models as surrogates to attack video models in the black-box setting. However, these traditional attack methods typically focus on the video classification tasks with vision-only models, failing to integrate multimodal information.

To address these limitations, we propose a highly transferable attack method, named as Image To Video MLLM (I2V-MLLM) attack. In I2V-MLLM, we utilize an image-based multimodal large language model (I-MLLM) as a surrogate model to craft adversarial video samples without accessing the internals of target V-MLLMs. Specifically, we extract key-frames from videos and send them into an I-MLLM to obtain adversarial perturbations. Multimodal interactions and spatiotemporal information are integrated to disrupt video representations within the latent space, improving adversarial transferability. Additionally, perturbation propagation technique is introduced to handle different unknown frame sampling strategies used by V-MLLMs.

We conduct various experiments on three well-established datasets, MSVD-QA (Xu et al. 2017) and MSRVTT-QA (Xu et al. 2017) to evaluate the performance of our proposed I2V-MLLM attack in multiple video-text multimodal tasks. The experimental results demonstrate that our method can generate adversarial videos with strong transferability across different V-MLLMs (Chat-Univi (Jin et al. 2024), LLava-Next-Video (Zhang et al. 2024b), VideoChat, Video-LLaMA (Zhang, Li, and Bing 2023)), and achieve competitive performance with white-box attacks against V-MLLMs. Our main contributions are summarized as follows:

- We explore the transferable adversarial attack on four V-MLLMs and analyze the reasons for the low transferabil-

ity when using existing methods to generate adversarial video samples. To the best of our knowledge, this is the first work to explore black-box attacks on V-MLLMs.

- We propose a highly transferable attack method, named I2V-MLLM, for V-MLLMs using I-MLLMs to generate adversarial video samples. The adversarial videos generated by this method can effectively disrupt different V-MLLMs, significantly degrading their performance on multiple video-text multimodal tasks.
- We conduct extensive experiments on four different V-MLLMs using MSVD-QA, MSRVTT-QA. The results demonstrate that our proposed attack method has strong transferability across V-MLLMs.

2 Related Work

2.1 Multimodal Large Language Models

MLLMs comprise a vision model, pretrained LLM, and projector converting key visual information into LLM-processable textual representations. MLLMs fall into image-based and video-based types. I-MLLMs (Dai et al. 2023; Liu et al. 2023; Zhu et al. 2023; Alayrac et al. 2022; Awadalla et al. 2023; Hu et al. 2023; Bai et al. 2023) handle image-text inputs, excelling at visual question answering and image captioning. V-MLLMs extend I-MLLMs with core temporal modules enabling processing of video input, supporting tasks like VideoQA, spatiotemporal localization, and video captioning. For example, Chat-UniVi (Jin et al. 2024) extracts key frames and uses DPC-KNN (Du, Ding, and Jia 2016) to group them into events, Video-LLaMA (Zhang, Li, and Bing 2023) employs sequential encoding to model frame temporal relationships, and VideoChatGPT (Maaz et al. 2024) applies temporal pooling on video features. These methods capture fine-grained temporal dynamics for more comprehensive video understanding.

2.2 Adversarial Attacks on MLLMs

Despite strong performance, MLLMs are highly susceptible to adversarial attacks (Zhao et al. 2023; Luo et al. 2024; Cui et al. 2023; Tu et al. 2023; Zhang et al. 2024a; Bailey et al. 2023; Lu et al. 2023). For I-MLLMs, studies have explored their vulnerabilities: (Fu et al. 2023) introduces Trojan-like images compelling models to invoke attacker-specified malicious tools/APIs; (Dong et al. 2023) uses open-source MLLMs to generate transferable examples attacking closed-source models (Bard (Google 2024), Bing Chat (Microsoft 2024), GPT-4V (OpenAI 2024b)), demonstrating strong cross-MLLM transferability; (Gu et al. 2025) proposes DynVLA Attack, applying dynamic perturbations to vision-language connectors to enhance cross-alignment generalization. While I-MLLM attacks are well-explored, V-MLLMs remain underexplored. (Li et al. 2024a) proposes a flow-based white-box attack on V-MLLMs, but real-world V-MLLMs often have inaccessible internal architectures/parameters. To address this, we explore black-box adversarial attacks on V-MLLMs.

2.3 Adversarial Attacks on Video Models

Video models are widely used in autonomous vehicles, video verification, security, and related areas, but remain vulnerable to adversarial attacks (Li et al. 2018; Xie et al. 2022; Jiang et al. 2019; Wei et al. 2020; Cao et al. 2024). For example, U3D (Xie et al. 2022) deceives classifiers via universal video perturbations, while StyleFool (Cao et al. 2023) uses unrestricted style transfer perturbations to attack video classifiers. Recent work explores cross-modal image-to-video attacks (Wei et al. 2021; Kim et al. 2023; Wang, Guo, and Wang 2023) with promising results. However, these attacks target video classification, ignoring visual-textual interactions. V-MLLMs integrate visual-textual information, making these methods incompatible. To address this, our method incorporates multimodal interactions in adversarial video crafting, aligning with V-MLLM operations.

3 Methodology

3.1 Preliminary

Given a video sample $V \in \mathcal{V}$ with M associated QA pairs $\{(q_m, a_m)\}_{m=1}^{M}$, where q_m is the m-th question and a_m is the corresponding answer. Let F denote the I-MLLM (e.g., BLIP-2 (Li et al. 2023), MiniGPT-4 (Zhu et al. 2023)) and G denote the V-MLLM (e.g., Video-LLaMA (Zhang, Li, and Bing 2023), Chat-UniVi. We use $G(V, q)$ to denote the answer generated by the V-MLLM for the given video V and question q. The goal of our proposed attack is to generate an adversarial example $V_{adv} = V + \delta'$ using F, which can cause G to produce an answer $G(V_{adv}, q_i)$ that differs significantly from the correct answer a_i, without accessing the parameters or structure of G, where δ' denotes the adversarial perturbations specifically tailored for V. To ensure that the adversarial perturbation δ' is imperceptible, we restrict it by $||\delta'||_\infty \leq \epsilon$, where $||\cdot||_\infty$ denotes the L_∞ norm, and ϵ is a constant for the norm constraint. We utilize the evaluation model E (i.e., GPT-4o-mini (OpenAI 2024a)) to assess whether the generated answer aligns with the reference answer. We aim to find imperceptible adversarial perturbations that minimize the number of correct responses, formulated as follows:

$$\underset{\delta'}{\operatorname{argmin}} \frac{1}{M} \sum_{i=1}^{M} E(G(V + \delta', q_i), a_i), \text{s.t. } ||\delta'||_\infty \leq \epsilon, \quad (1)$$

where $E(\cdot, \cdot)$ is the evaluation model's judgment function, which outputs 1 if they match, and 0 otherwise.

3.2 Motivation

To explore the transferability of adversarial videos across V-MLLMs, we first conduct an investigation of existing attack methods. Based on the experimental results (in Tab. 1), we attribute their poor transferability to the following limitations: (1) focusing only on sparse key-frames, (2) lacking generalization in perturbing video features, and (3) failing to integrate multimodal information.

Focusing only on sparse key-frames. FMM attack exhibits limited performance when the key-frame ratio is low.

This is because FMM selects key-frames based on optical flow and only perturbs these frames, while V-MLLMs typically sample frames sequentially, making it difficult to ensure that all frames extracted by the target model are perturbed. To address this, we first modify the FMM attack by replacing the sparse spatial perturbation with full perturbation on the key-frames sampled by V-MLLMs, which we call the Vanilla attack. While this adjustment improves white-box performance, the transferability still remains constrained. To further enhance transferability, we propagate perturbations from key-frames across the entire video, as shown in rows 1, 2, 4, and 5 of Tab. 1.

Lacking generalization in perturbing video features. Adversarial perturbations generated based on certain V-MLLM can overfit to specific video module, limiting their generalization to other V-MLLMs. To improve transferability, we focus on lower-level image features. The I2V attack (Wei et al. 2021), which perturbs each video frame to disrupt image features, demonstrates improved transferability when using image models as surrogates to craft adversarial video samples, as shown in rows 3, 4, and 5 of Tab. 1.

Failing to integrate multimodal information. The I2V attack shows a limited improvement in transferability, as it was originally designed for video classification and does not account for the multimodal interactions, which is essential for V-MLLMs. Therefore, we propose using an I-MLLM as a surrogate, integrating multimodal interaction information into the process of generating adversarial video samples, which leads to a significant improvement in transferability, as demonstrated in rows 3 and 6 of Tab. 1.

In summary, we propose using I-MLLMs as surrogates to generate adversarial videos that incorporate multimodal interactions. In addition, we introduce a perturbation propagation method to handle different unknown frame sampling strategies. The results in Tab. 1 demonstrate the strong transferability of I2V-MLLM across different V-MLLMs. The following sections detail the I2V-MLLM attack.

Attack	Target Model				AASR
	C-U	L-N-V	V-C	V-L	
FMM	8.70	18.76	13.84	27.93*	17.31
Vanilla	11.62	25.31	15.10	64.14*	29.04
I2V	32.17	33.39	41.13	36.51	35.80
FMM w/ Prop.	14.54	25.31	14.62	27.93*	20.60
Vanilla w/ Prop.	14.94	32.24	17.50	64.14*	32.21
I2V-MLLM	**48.39**	**45.54**	**63.09**	**74.91**	**57.98**

Table 1: Attack success rates (ASR, %) on the MSVD-QA validation set for Zero-Shot VideoQA tasks. **FMM** and **I2V** denote attack methods from (Li et al. 2024a) and (Wei et al. 2021), respectively. **Vanilla** attack applies full perturbations on all key-frames sampled by V-MLLMs. **Prop.** denotes perturbation propagation. * indicates white-box attacks. AASR represents the average ASR across all target models for each surrogate model. A higher AASR indicates better adversarial transferability. **Note:** C-U: Chat-UniVi, L-N-V: LLaVA-NeXT-Video, V-C: VideoChat, V-L: Video-LLaMA.

3.3 I2V-MLLM Attack

The proposed I2V-MLLM attack utilizes an I-MLLM to produce adversarial video samples, targeting image-to-video cross-modal black-box attacks on V-MLLMs with significant transferability. By manipulating the intermediate features of vision models and projectors of I-MLLMs, our approach generates adversarial video samples that interfere with the intermediate features of black-box V-MLLMs. The I2V-MLLM attack consists of three components: vision model attack, projector attack, and perturbation propagation.

Vision Model Attack To enhance generalization in perturbing video features, I2V-MLLM disrupts both image features and spatiotemporal information extracted by the vision model. We first split the video V into K clips: $V = \{v^1, v^2, \ldots, v^K\}$, where K = total number of frames $\times$ key-frame ratio β. We select the first frame x^k from each clip v^k as the key-frame, resulting in K key-frames, $X = \{x^1, x^2, \ldots, x^K\}$, each capturing the essential information of their respective clips. We extract spatiotemporal embeddings of X using the vision model. This model independently encodes the K frames, producing frame-level embeddings $F_V(X) \in \mathbb{R}^{K\times N\times D_1}$, where $F_V(\cdot)$ denotes the encoder of the vision model, N is the number of patches per frame, and D_1 is the dimension of the embeddings. Frame-level embeddings are average-pooled along the temporal dimension to obtain temporal embeddings $F_V^t(X) \in \mathbb{R}^{N\times D_1}$, which implicitly incorporates temporal information of K frames. Similarly, the frame-level embeddings are average-pooled along the spatial dimension to obtain spatial embeddings $F_V^s(X) \in \mathbb{R}^{K\times D_1}$, which incorporate the spatial information of K frames. The temporal and spatial embeddings are concatenated to obtain the original spatiotemporal embeddings$F_V^{ts}(X) = [F_V^t(X), F_V^s(X)] \in \mathbb{R}^{(N+K)\times D_1}$. For the adversary $X_{adv} = \{x^1+\delta^1, x^2+\delta^2, \ldots, x^K+\delta^K\}$, we can similarly obtain the adversarial spatiotemporal embeddings $F_V^{ts}(X_{adv})$. To disrupt the spatiotemporal features, I2V-MLLM optimizes the adversarial perturbations by minimizing the cosine similarity between the original and the adversarial spatiotemporal embeddings:

$$\mathcal{L}_V = \sum_{i=1}^{N+K} \frac{Cos(F_V^{ts}(X)_i, F_V^{ts}(X_{adv})_i)}{N+K}, \qquad (2)$$

where $F_V^{ts}(X)_i$ and $F_V^{ts}(X_{adv})_i$ represent the i-th elements in the spatiotemporal embeddings of the original and the adversarial video frames, respectively.

Projector Attack To further disrupt V-MLLMs' capacity for video-text multimodal tasks, I2V-MLLM interferes with the intermediate feature of the projector (e.g. Q-Former (Li et al. 2023)), which plays an essential role in aligning visual and textual representations. We feed the projector with the original frame-level embeddings $F_V(X)$, the adversarial frame-level embeddings $F_V(X_{adv})$ from the vision model, and the caption set $T = \{t_1, t_2, \ldots, t_M\}$. After multimodal alignment, they are transformed into the original visual features $F_P^v(X) \in \mathbb{R}^{N_1\times D_2}$, the adversarial vi-

sual features $F_P^v(X_{adv}) \in \mathbb{R}^{N_1 \times D_2}$, and the textual features $F_P^t(T) \in \mathbb{R}^{N_2 \times D_2}$. Here, N_1 and N_2 represent the number of visual features and the textual features, respectively. And D_2 denotes the dimension of these features. The captions are complete sentences generated based on the question q and the answer a using GPT-4o-mini (OpenAI 2024a). For example, given the question q: 'What is the man doing?' and the answer a: 'eat', the corresponding caption t would be: 'The man is eating.' To perturb the image features aligned with the text, I2V-MLLM optimizes the adversarial perturbations by minimizing the cosine similarity between the original and the adversarial visual features:

$$\mathcal{L}_{P_v} = \sum_{n_1=1}^{N_1} \frac{Cos(F_P^v(X)_{n_1}, F_P^v(X_{adv})_{n_1})}{N_1}, \quad (3)$$

where $F_P^v(X)_{n_1}$ and $F_P^v(X_{adv})_{n_1}$ are the n_1-th visual feature of the original and the adversarial video frames, respectively. To disrupt multimodal interactions between adversarial frames and text, I2V-MLLM optimizes the adversarial perturbations by minimizing the cosine similarity between the adversarial visual features and the textual features:

$$\mathcal{L}_{P_{v2t}} = \sum_{n_1=1}^{N_1} \sum_{n_2=1}^{N_2} \frac{Cos(F_P^v(X_{adv})_{n_1}, F_P^t(T)_{n_2})}{N_1 N_2}, \quad (4)$$

where $F_P^t(T)_{n_2}$ is the n_2-th textual feature of T. The total loss function for projector is

$$\mathcal{L}_P = \mathcal{L}_{P_v} + \mathcal{L}_{P_{v2t}}. \quad (5)$$

Optimization and Perturbation Propagation To maximize the efficacy of the adversarial attack, we combine the losses $\mathcal{L}_V$ and $\mathcal{L}_P$ into a unified objective. This combined loss ensures that both the vision model and the projector's intermediate features are significantly perturbed. The unified loss is formulated as:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_V + \lambda_2 \mathcal{L}_P, \quad (6)$$

where λ_1 and λ_2 correspond to the two loss weights, which aim to balance them during the optimization.

We optimize δ_k according to the following expression:

$$\delta^k = \arg\min_{\delta^k}(\mathcal{L}_{total}), s.t. ||\delta^k||_\infty \leq \epsilon, k = 1, \ldots, K. \quad (7)$$

Finally, we replicate δ^k to match the length of its corresponding video clip v^k, obtaining perturbed clip δ'^k. The adversarial video is then constructed by applying pixel-wise addition to combine these perturbed clips with the original ones: $V_{adv} = V + \delta' = \{v^1 + \delta'^1, v^2 + \delta'^2, \ldots, v^K + \delta'^K\}$. We term this approach Direct Propagation (DP). Due to the high similarity between consecutive frames, DP proves to be a simple yet effective method.

4 Experiment

4.1 Experimental Setting

In this section, we present the experimental setting, including datasets, models, attack setting and metrics.

Datasets and models. Referring to the quantitative benchmarking framework proposed in (Maaz et al. 2024), we evaluate our I2V-MLLM attack on Zero-Shot VideoQA tasks using the validation set of MSVD-QA and MSRVTT-QA. We perform the proposed method on three I-MLLMs: BLIP-2, InstructBLIP (Dai et al. 2023), and MiniGPT-4. Our method is evaluated on four different V-MLLMs: Chat-UniVi, LLaVA-NeXT-Video, VideoChat, and Video-LLaMA (Zhang, Li, and Bing 2023), each with a Vicuna-7B (Chiang et al. 2023) as the LLM.

Attack setting. In I2V-MLLM, we employ the projected gradient descent (PGD) (Madry et al. 2019) with a perturbation bound of $\epsilon = 16$, an iteration number of $I = 50$, and a step size of $\alpha = 1$ for the attack process. The parameters λ_1 and λ_2 are both set to 1, and the key-frame ratio β is set to 30%. I2V attack, utilizing CLIP-L/14 (Radford et al. 2021) as the surrogate model, applies tailored perturbations to each frame of the video. For a fair comparison, the PGD parameters ($\epsilon = 16$, $I = 50$ and $\alpha = 1$) in FMM, Vanilla, and I2V attacks maintain the same for our method. Additionally, in the FMM setup, the key-frame ratio β is also set to 30%. All the experiments are conducted on an NVIDIA-A6000 GPU.

Metrics. We use Attack Success Rate (ASR) to evaluate the effectiveness of adversarial examples on Zero-Shot VideoQA tasks. It measures the percentage of successful attacks on questions the model answered correctly for clean videos. Answer correctness is evaluated using GPT-4o-mini, which checks whether the model's prediction semantically aligns with the ground truth. We also provide the average ASR (AASR) across all evaluated V-MLLMs. A higher ASR or AASR indicates better adversarial transferability. To evaluate the model's overall performance when encountering adversarial videos, we further employ GPT-assisted methods (Maaz et al. 2024) to assess Accuracy (Acc.) and GPT-Score. Specifically, accuracy (Acc.) refers to the model's prediction accuracy, while the GPT score (Score) assesses the quality of the model's predictions, assigning a relative score on a scale from 0 to 5. GPT-4o-mini is used for evaluation due to its strong text understanding and cost efficiency.

4.2 Attack Performance

In this section, we compare our proposed I2V-MLLM attack with the FMM, Vanilla, and I2V attacks. The results, summarized in Tab. 2 and Tab. 3, provide a quantitative comparison of the ASR, AASR, Acc., and GPT Score for the MSVD-QA and MSRVTT-QA datasets, respectively.

Evaluation of ASR. As shown in Tab. 2 and Tab. 3, I2V-MLLM achieves the highest AASR compared to previous attack methods, achieving AASR of 57.98%, 53.63%, and 53.88% for MSVD-QA and 60.76%, 56.04%, and 56.44% for MSRVTT-QA when taking BLIP-2, InstructBLIP, MiniGPT-4 as surrogate models, respectively, significantly outperforming previous attack methods. I2V-MLLM (BLIP-2) achieves the best attack performance on Video-LLaMA and near-best attack performance on other target models. It achieves ASRs of 48.39%, 45.54%, 63.09%, and 74.91% on MSVD-QA, and 47.93%, 53.78%, 62.38%, and 78.95% on MSRVTT-QA, respectively, outperforming both the FMM and I2V attacks while achieving performance

Attack	Surrogate Model	Chat-UniVi			LLaVA-NeXT-Video			VideoChat			Video-LLaMA			AASR
		ASR ↑	Acc. ↓	Score ↓	ASR ↑	Acc. ↓	Score ↓	ASR ↑	Acc. ↓	Score ↓	ASR ↑	Acc. ↓	Score ↓	
Clean	/	/	60.89	3.34	/	48.95	2.90	/	60.24	3.42	/	53.81	3.09	/
FMM	Chat-UniVi	16.00*	57.41*	3.18*	16.33	50.38	2.93	13.21	60.75	3.39	21.47	53.31	3.06	16.76
	LLaVA-NeXT-Video	9.22	60.65	3.34	20.48*	47.84*	2.83*	13.49	60.30	3.38	21.32	53.43	3.05	16.13
	VideoChat	8.12	61.81	3.38	15.38	51.30	2.98	14.62*	59.91*	3.35*	20.74	54.08	3.09	14.72
	Video-LLaMA	8.70	61.53	3.36	18.76	49.40	2.89	13.84	60.20	3.38	27.93*	48.39*	2.84*	17.31
Vanilla	Chat-UniVi	56.34*	30.40*	1.88*	12.64	48.60	2.87	6.00	59.05	3.35	21.78	53.53	3.06	24.19
	LLaVA-NeXT-Video	9.33	59.05	3.30	52.45*	24.96*	1.53*	5.15	59.97	3.37	24.94	50.82	2.95	22.97
	VideoChat	7.35	59.61	3.38	7.93	49.90	2.91	68.90*	23.81*	1.67*	20.74	54.15	3.09	26.23
	Video-LLaMA	11.62	58.88	3.26	25.31	41.01	2.52	15.10	59.33	3.34	64.14*	23.88*	1.72*	29.04
I2V	CLIP-L/14	32.17	51.53	2.92	33.39	43.63	2.60	41.13	49.57	2.91	36.51	46.49	2.71	35.80
I2V-MLLM	BLIP-2	**48.39**	**34.72**	**2.03**	45.54	29.33	1.94	**63.09**	**26.08**	**1.82**	**74.91**	**17.07**	**1.39**	**57.98**
	InstructBLIP	45.74	35.10	2.16	44.61	30.64	2.13	54.26	31.99	2.10	69.90	20.58	1.58	53.63
	MiniGPT-4	43.58	37.02	2.21	**46.50**	**27.98**	**1.76**	56.51	30.49	2.06	68.92	21.37	1.60	53.88

Table 2: The results on the **MSVD-QA** for Zero-Shot VideoQA tasks. ASR (%) indicates attack success rate. Acc.(%) denotes the accuracy of the model's predictions, while the Score represents GPT Score, which assesses the model and assigns a relative score to the predictions on a scale of 0 to 5. AASR represents the average ASR across all target models for each surrogate model. A higher AASR indicates better adversarial transferability. The best ASR for each target model under **black-box attacks** is highlighted in **bold**. * indicates white-box attacks for reference.

comparable to the white-box Vanilla attack. We also incorporate Acc. and GPT Score as metrics to better analyze the impact of adversarial videos on V-MLLM performance.

Evaluation of the quality of generated answers. We also incorporate Acc. and GPT Score as metrics to better analyze the impact of adversarial videos on V-MLLM performance. As shown in Tab. 2 and Tab. 3, the proposed I2V-MLLM significantly reduces both Acc. and Scores across all target models, particularly for VideoChat and Video-LLaMA. On the MSVD-QA dataset, Acc. drops to 26.08% and 17.07%, while Scores fall to 1.82 and 1.39. On the MSRVTT-QA dataset, Acc. further declines to 18.72% and 10.68%, with Scores of 1.57 and 1.17, respectively. Significant effects are also observed on Chat-UniVi and LLaVA-NeXT-Video. These significant performance drops highlight the I2V-MLLM attack's destructive power, transferability, and effectiveness across multiple V-MLLMs, while exposing existing models' vulnerability in black-box scenarios.

4.3 Ablation Study

In this section, we provide ablation studies on the loss functions, key-frame ratio β, perturbation propagation and different projectors in I2V-MLLM attack. Experiments are conducted on the MSVD-QA for Zero-Shot VideoQA tasks.

Influence of loss functions. In Fig. 3, we provide ablation study on the components of the loss functions used in our I2V-MLLM. The surrogate model is BLIP-2, and the generated adversarial videos are evaluated across four V-MLLMs. It can be observed that using either $\mathcal{L}_V$ or $\mathcal{L}_P$ alone achieves satisfactory attack performance. Combining both, which simultaneously disrupts low-level image features and the alignment between visual and textual modalities, further enhances the attack performance.

Influence of key-frame ratio and propagation method.

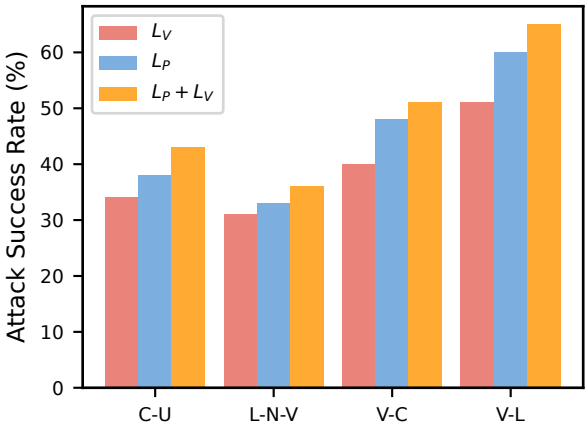

Figure 3: ASR (%) of the I2V-MLLM attack with different loss functions. **Note**: C-U: Chat-UniVi, L-N-V: LLaVA-NeXT-Video, V-C: VideoChat, V-L: Video-LLaMA.

The number of key-frames used to optimize the perturbation, as well as the decision to propagate these perturbations across the entire video, significantly affects the transferability of adversarial video samples. Fig. 4 illustrates the results obtained with various key-frame ratios, comparing scenarios with and without perturbation propagation. It can be observed that as the key frame ratio increases, the generated adversarial samples show improved transferability. Perturbation propagation substantially improves AASR by ensuring that all the frames extracted by unseen V-MLLMs are perturbed. As illustrated by the gain curve in the Fig. 4, the improvement from perturbation propagation initially rises

Attack	Surrogate Model	Chat-UniVi			LLaVA-NeXT-Video			VideoChat			Video-LLaMA			AASR
		ASR ↑	Acc. ↓	Score ↓	ASR ↑	Acc. ↓	Score ↓	ASR ↑	Acc. ↓	Score ↓	ASR ↑	Acc. ↓	Score ↓	
Clean	/	/	39.62	2.51	/	29.17	2.06	/	38.92	2.50	/	31.42	2.17	/
FMM	Chat-UniVi	23.39*	36.85*	2.36*	24.79	31.60	2.17	9.04	39.44	2.53	32.50	32.03	2.20	22.43
	LLaVA-NeXT-Video	13.20	40.01	2.52	28.62*	29.90*	2.09*	8.52	39.24	2.51	32.27	31.94	2.19	20.65
	VideoChat	12.83	40.52	2.54	25.29	31.10	2.15	15.15*	37.99*	2.46*	30.48	32.56	2.21	20.94
	Video-LLaMA	12.72	40.80	2.55	27.92	30.25	2.12	8.16	39.71	2.53	37.38*	29.60*	2.07*	21.55
Vanilla	Chat-UniVi	52.19*	23.10*	1.68*	25.85	30.09	2.10	9.58	39.52	2.52	32.32	32.24	2.21	29.99
	LLaVA-NeXT-Video	13.07	41.08	2.56	65.90*	15.06*	1.46*	8.50	39.56	2.53	35.23	30.41	2.13	30.68
	VideoChat	12.64	41.35	2.57	25.27	30.72	2.14	63.46*	16.69*	1.42*	30.84	32.58	2.21	33.05
	Video-LLaMA	14.33	40.78	2.55	37.91	26.14	1.91	8.83	39.53	2.52	64.11*	18.07*	1.53*	31.29
I2V	CLIP-L/14	30.05	34.53	2.28	35.62	26.96	1.96	18.83	38.59	2.50	36.16	30.16	2.12	30.17
I2V-MLLM	BLIP-2	**47.93**	**28.42**	**2.00**	**53.78**	**19.34**	**1.58**	**62.38**	**18.72**	**1.57**	**78.95**	**10.68**	**1.17**	**60.76**
	InstructBLIP	45.37	31.88	2.14	50.96	21.72	1.70	54.78	22.66	1.76	73.04	13.52	1.34	56.04
	MiniGPT-4	46.60	30.94	2.11	49.47	21.12	1.67	56.41	21.95	1.73	73.28	13.63	1.32	56.44

Table 3: The results on the **MSRVTT-QA**. The corresponding metrics and settings are consistent with those in Tab. 2.

Propagation Method	/	OFP	BP	DP
Chat-UniVi	37.69	34.03	44.63	**48.39**
LLaVA-NEXT-Video	23.34	32.07	39.72	**45.54**
VideoChat	23.67	42.26	54.60	**63.09**
Video-LLaMA	45.89	62.40	71.23	**74.91**
AASR	32.65	42.69	52.55	**57.98**

Table 4: Attack success rates (ASR, %) of the I2V-MLLM attack with different perturbation propagation method. '/' represents no perturbation propagation. A higher AASR indicates better adversarial transferability.

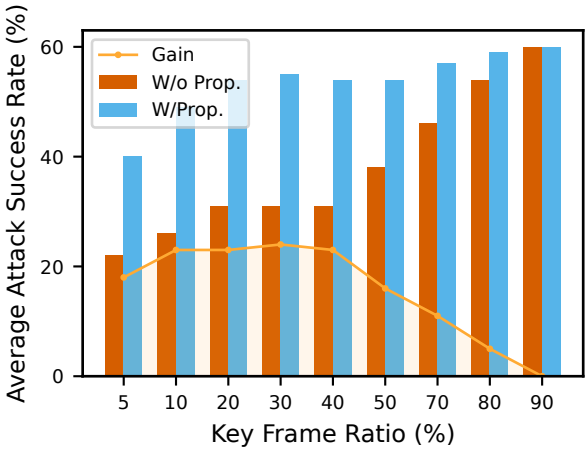

Figure 4: AASR (%) of the I2V-MLLM attack with various key-frame ratios, comparing scenarios with and without perturbation propagation. 'Prop.' represents 'Propagation'.

with the key-frame ratio but then diminishes, reaching its maximum at 30%. With an AASR already high at a 30% key-frame ratio, further increases yield minimal gains, and perturbation propagation reaches its maximal benefit at this point. Therefore, we adopt a key-frame ratio of $\beta = 30\%$.

Influence of propagation method. Different perturbation propagation methods may yield varying results. We set the key-frame ratio $\beta = 30\%$ and test three distinct perturbation propagation methods: Direct Propagation (DP), Optical Flow-based Propagation (Dosovitskiy et al. 2015) (OFP), and Bidirectional Linear Interpolation Propagation (Dai et al. 2017) (BP). As shown in Tab. 4, DP achieves the most significant improvement in AASR. Due to the high similarity between consecutive frames, DP proves to be a simple yet effective method. OFP suffers from added complexity and may distort the perturbation, resulting in lower effectiveness. BP's slightly lower performance stems from its reliance on interpolation, which may dilute perturbation intensity compared to DP's direct application. Therefore, we adopt DP in I2V-MLLM.

Other Ablation Studies. See the extended version.

5 Conclusion

In this paper, we are the first to systematically explore practical black-box transferable attacks on V-MLLMs. We conduct a thorough investigation of the limitations of existing adversarial methods, revealing that they exhibit notably lower transferability despite their impressive white-box performance. Our findings underscore the critical need for specially designed transferable attacks tailored to V-MLLMs. We propose the I2V-MLLM attack, a highly transferable cross-modal attack that leverages discriminative intermediate features of I-MLLMs and perturbation propagation to enhance cross-model transferability against V-MLLMs, while significantly reducing computational overhead. We hope our work will inspire further research on comprehensive evaluation and improving adversarial robustness of V-MLLMs.

Acknowledgments

Feng Zheng was supported by the National Natural Science Foundation of China (Project K23271004). Xi Xiao and Linhao Huang were supported by the Natural Science Foundation of Guangdong Province (Project 2025A1515011946) and the Major Key Project of PCL (Project PCL2023A06-4). XJ and BH were supported by the NSFC General Program (Project No. 62376235) and the Guangdong Basic and Applied Basic Research Foundation (Project Nos. 2022A1515011652 and 2024A151501239).

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; Ring, R.; Rutherford, E.; Cabi, S.; Han, T.; Gong, Z.; Samangooei, S.; Monteiro, M.; Menick, J.; Borgeaud, S.; Brock, A.; Nematzadeh, A.; Sharifzadeh, S.; Binkowski, M.; Barreira, R.; Vinyals, O.; Zisserman, A.; and Simonyan, K. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. arXiv:2204.14198.

Awadalla, A.; Gao, I.; Gardner, J.; Hessel, J.; Hanafy, Y.; Zhu, W.; Marathe, K.; Bitton, Y.; Gadre, S.; Sagawa, S.; Jitsev, J.; Kornblith, S.; Koh, P. W.; Ilharco, G.; Wortsman, M.; and Schmidt, L. 2023. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. arXiv:2308.01390.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv:2308.12966.

Bailey, L.; Ong, E.; Russell, S.; and Emmons, S. 2023. Image hijacks: Adversarial images can control generative models at runtime. *arXiv preprint arXiv:2309.00236*.

Cao, Y.; Xiao, X.; Sun, R.; Wang, D.; Xue, M.; and Wen, S. 2023. Stylefool: Fooling video classification systems via style transfer. In *2023 IEEE symposium on security and privacy (SP)*, 1631–1648. IEEE.

Cao, Y.; Zhao, Z.; Xiao, X.; Wang, D.; Xue, M.; and Lu, J. 2024. LogoStyleFool: Vitiating Video Recognition Systems via Logo Style Transfer. arXiv:2312.09935.

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.

Cui, X.; Aparcedo, A.; Jang, Y. K.; and Lim, S.-N. 2023. On the Robustness of Large Multimodal Models Against Image Adversarial Attacks. arXiv:2312.03777.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 764–773.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500.

Dong, Y.; Chen, H.; Chen, J.; Fang, Z.; Yang, X.; Zhang, Y.; Tian, Y.; Su, H.; and Zhu, J. 2023. How Robust is Google's Bard to Adversarial Image Attacks? arXiv:2309.11751.

Dosovitskiy, A.; Fischer, P.; Ilg, E.; Hausser, P.; Hazirbas, C.; Golkov, V.; Van Der Smagt, P.; Cremers, D.; and Brox, T. 2015. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 2758–2766.

Du, M.; Ding, S.; and Jia, H. 2016. Study on density peaks clustering based on k-nearest neighbors and principal component analysis. *Knowledge-Based Systems*, 135–145.

Fu, X.; Wang, Z.; Li, S.; Gupta, R. K.; Mireshghallah, N.; Berg-Kirkpatrick, T.; and Fernandes, E. 2023. Misusing Tools in Large Language Models With Visual Adversarial Examples. arXiv:2310.03185.

Google. 2024. Gemini.

Gu, C.; Gu, J.; Hua, A.; and Qin, Y. 2025. Improving Adversarial Transferability in MLLMs via Dynamic Vision-Language Alignment Attack. *arXiv preprint arXiv:2502.19672*.

Hu, W.; Xu, Y.; Li, Y.; Li, W.; Chen, Z.; and Tu, Z. 2023. BLIVA: A Simple Multimodal LLM for Better Handling of Text-Rich Visual Questions. arXiv:2308.09936.

Jiang, L.; Ma, X.; Chen, S.; Bailey, J.; and Jiang, Y.-G. 2019. Black-box adversarial attacks on video recognition models. In *Proceedings of the 27th ACM International Conference on Multimedia*, 864–872.

Jin, P.; Takanobu, R.; Zhang, W.; Cao, X.; and Yuan, L. 2024. Chat-UniVi: Unified Visual Representation Empowers Large Language Models with Image and Video Understanding. arXiv:2311.08046.

Kim, H.-S.; Son, M.; Kim, M.; Kwon, M.-J.; and Kim, C. 2023. Breaking Temporal Consistency: Generating Video Universal Adversarial Perturbations Using Image Models. arXiv:2311.10366.

Li, J.; Gao, K.; Bai, Y.; Zhang, J.; tao Xia, S.; and Wang, Y. 2024a. FMM-Attack: A Flow-based Multi-modal Adversarial Attack on Video-based LLMs. arXiv:2403.13507.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597.

Li, K.; He, Y.; Wang, Y.; Li, Y.; Wang, W.; Luo, P.; Wang, Y.; Wang, L.; and Qiao, Y. 2024b. VideoChat: Chat-Centric Video Understanding. arXiv:2305.06355.

Li, S.; Neupane, A.; Paul, S.; Song, C.; Krishnamurthy, S. V.; Chowdhury, A. K. R.; and Swami, A. 2018. Adversarial perturbations against real-time video classification systems. *arXiv preprint arXiv:1807.00458*.

Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; and Yuan, L. 2024. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. arXiv:2311.10122.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. arXiv:2304.08485.

Lu, D.; Wang, Z.; Wang, T.; Guan, W.; Gao, H.; and Zheng, F. 2023. Set-level guidance attack: Boosting adversarial

transferability of vision-language pre-training models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 102–111.

Luo, H.; Gu, J.; Liu, F.; and Torr, P. 2024. An Image Is Worth 1000 Lies: Adversarial Transferability across Prompts on Vision-Language Models. arXiv:2403.09766.

Maaz, M.; Rasheed, H.; Khan, S.; and Khan, F. S. 2024. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. arXiv:2306.05424.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2019. Towards Deep Learning Models Resistant to Adversarial Attacks. arXiv:1706.06083.

Microsoft. 2024. Bing Chat.

OpenAI. 2024a. GPT-4o-Mini: Advancing Cost-Efficient Intelligence.

OpenAI. 2024b. GPT-4V(ision) System Card.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Tu, H.; Cui, C.; Wang, Z.; Zhou, Y.; Zhao, B.; Han, J.; Zhou, W.; Yao, H.; and Xie, C. 2023. How many unicorns are in this image? a safety evaluation benchmark for vision llms. *arXiv preprint arXiv:2311.16101*.

Wang, R.; Guo, Y.; and Wang, Y. 2023. Global-local characteristic excited cross-modal attacks from images to videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2635–2643.

Wei, Z.; Chen, J.; Wei, X.; Jiang, L.; Chua, T.-S.; Zhou, F.; and Jiang, Y.-G. 2020. Heuristic black-box adversarial attacks on video recognition models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12338–12345.

Wei, Z.; Chen, J.; Wu, Z.; and Jiang, Y.-G. 2021. Cross-Modal Transferable Adversarial Attacks from Images to Videos. arXiv:2112.05379.

Xie, S.; Wang, H.; Kong, Y.; and Hong, Y. 2022. Universal 3-dimensional perturbations for black-box attacks on video recognition systems. In *2022 IEEE Symposium on Security and Privacy (SP)*, 1390–1407. IEEE.

Xu, D.; Zhao, Z.; Xiao, J.; Wu, F.; Zhang, H.; He, X.; and Zhuang, Y. 2017. Video Question Answering via Gradually Refined Attention over Appearance and Motion. In *ACM Multimedia*.

Zhang, H.; Li, X.; and Bing, L. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. arXiv:2306.02858.

Zhang, H.; Shao, W.; Liu, H.; Ma, Y.; Luo, P.; Qiao, Y.; and Zhang, K. 2024a. Avibench: Towards evaluating the robustness of large vision-language model on adversarial visual-instructions. *arXiv preprint arXiv:2403.09346*.

Zhang, Y.; Wu, J.; Li, W.; Li, B.; Ma, Z.; Liu, Z.; and Li, C. 2024b. Video Instruction Tuning With Synthetic Data. arXiv:2410.02713.

Zhao, Y.; Pang, T.; Du, C.; Yang, X.; Li, C.; Cheung, N.-M.; and Lin, M. 2023. On Evaluating Adversarial Robustness of Large Vision-Language Models. arXiv:2305.16934.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. arXiv:2304.10592.