Breaking Multimodal LLM Safety via Video-Driven Prompting

Dong Wang, Xiangyu He, Xinqi Lyu, Bin Xiao*
The Hong Kong Polytechnic University
{dong-comp.wang, xiangyu.he, xinqi.lyu}@connect.polyu.hk, b.xiao@polyu.edu.hk

Abstract

Multimodal Large Language Models (MLLMs) have achieved remarkable progress in visual reasoning tasks, serving as the core perception engines for emerging AI agents like OpenClaw. While recent studies have introduced several effective image-based jailbreak methods, the vulnerabilities inherent in the video modality remain a largely unexplored frontier. As a pioneering effort to bridge this critical safety gap, we demonstrate that video-driven jailbreak attacks are significantly more effective and robust against pre-defined system prompts than their image-based counterparts. Specifically, we find that simply repeating a harmful image across multiple frames to construct a video can bypass the safety mechanisms of MLLMs. Our analysis reveals that unsafe videos are embedded more similarly to safe videos in the model's representation space than individual harmful images, making them harder to detect. Moreover, videos composed of identical frames are processed more like static images and are more likely to trigger safety defenses compared to videos with diverse frames. Motivated by these findings, we propose an algorithm that injects harmful content into typographic videos by interleaving it with diverse, safety-proximal frames, thereby evading MLLM safety alignment. Extensive experiments demonstrate that our approach achieves state-of-the-art jailbreak performance on several widely-used MLLMs (e.g., VideoLLaMA-2, Qwen2.5-VL, GPT-4.1, and Gemini-2.5) under 16 different safety policies. Warning: This work contains potentially offensive content generated by LLMs.

1. Introduction

Multimodal Large Language Models (MLLMs) have demonstrated significant success in visual understanding [16, 30, 31, 43, 47, 52] and practical applications, driving the development of sophisticated interactive AI agents [23, 53, 69] such as OpenClaw. However, these complex architectures inherently suffer from a broader spectrum of security and privacy vulnerabilities [27, 34, 35]. Furthermore, because they are pre-trained on large-scale Internet-sourced data that often lacks sufficient ethical review, MLLMs are vulnerable to jailbreak attacks [5, 13, 15, 21, 29, 32, 39, 46, 67, 72]. Adversaries may attempt to manipulate multimodal prompts to elicit information that contravenes established safety policies [40, 44].

Recent studies [13, 14, 20, 26, 32, 46, 50, 61, 63] have explored methods to jailbreak MLLMs through the image modality. These approaches can be broadly categorized into two types. Perturbation-based methods [14, 46, 50, 63] involve adding imperceptible noise to benign images to attack MLLMs, utilizing gradient descent. However, these methods typically require white-box access and suffer from low transferability, which limits their practicality. On the other hand, structure-based methods [13, 20, 26, 32, 61] aim to jailbreak models in a black-box setting. They inject harmful text prompts into images to successfully bypass safety alignments. Nonetheless, these methods often demand careful design due to limited transparency into model architectures and parameters.

Despite the rising capabilities of MLLMs, video-modality vulnerabilities remain insufficiently studied. Since each frame of a video can be viewed as an individual image, it is essential to first evaluate the transferability of image-based attacks. This evaluation will provide insights into how vulnerabilities in image attacks may propagate to the video modality, thereby laying the groundwork for a comprehensive assessment of the safety of this new class of MLLMs. Our findings reveal that these image-based attacks can also jailbreak MLLMs capable of understanding both images and videos. Moreover, we observe that simply stacking the same toxic image into a video can enhance attack performance. This suggests that, despite their impressive utility [11, 12, 68], current MLLMs cannot process videos safely. The underlying mechanism remains unclear.

Building on the above analysis, we examine, from the perspective of the embedding space, why stacking identical frames of the same harmful image into a video can enhance attacks. We discover that unsafe videos are more similar to safe videos compared to images (Fig. 2c), which

*Bin Xiao is the corresponding author.

indicates that MLLMs cannot easily detect unsafe videos compared to unsafe images. Moreover, we show that the image-stack approach is also suboptimal. Because, to the model, a video with identical frames tends to be processed more like a single image than a video with diverse frames, thereby more readily triggering safety detection. This leads us to raise the question: *Can we generate videos that are similar to safe data while exhibiting diverse frames?* To achieve this and bypass safety alignment, we propose to jailbreak MLLMs using **S**afety-**P**roximal **T**ypographic **V**ideos (**SPTV**) as shown in Fig. 1. We first augment each original harmful query into several safe and unsafe questions on the same topic. Secondly, each question is paraphrased into a sentence starting with a fixed prefix. Thirdly, each new sentence serves as the title in the top half of a typographic image, followed by blank items in the bottom half. Then, to obtain diverse safety-proximal frames, we formulate frame selection as a bipartite matching problem. The Hungarian Matching algorithm is employed to solve it. Frames are selected among candidates with high similarity to the target, forming the video. Additionally, we design a text prompt to steer model behaviors. Our main contributions are summarized as follows:

- We advance the fundamental understanding of why video encoders are uniquely susceptible to jailbreaks. By analyzing feature similarity and refusal probability in the embedding space, we elucidate how identical-frame stacking evades safety filters, and demonstrate the critical necessity of frame diversity to bypass dynamic safety alignments.
- We improve the state-of-the-art in black-box attacks by developing Safety-Proximal Typographic Videos (SPTV). By elegantly interleaving paraphrased, safety-proximal frames via bipartite matching, SPTV maximizes the circumvention of safety alignments and achieves superior jailbreak performance across leading MLLMs under 16 distinct safety policies.
- We enhance current safety mechanisms by developing a Video-aware System Prompt (VSP). Extensive evaluations confirm that VSP effectively improves the model's robustness, successfully neutralizing both static and dynamic visual attacks and substantially outperforming conventional image-centric guardrails.

2. Related Work

2.1. Multimodal Large Language Models

Large Language Models (LLMs) [8, 22, 54] have been extensively applied in multimodal domains. Numerous studies [1, 10, 30, 31, 57, 58, 66, 70, 71] have successfully integrated visual information into LLMs. However, most of these efforts primarily focus on image perception and understanding. Recently, an increasing number of MLLMs have begun to analyze videos. Both MM-REACT [60] and ViperrGPT [51] utilize an LLM as a scheduler, processing videos without any training. LLMs have been incorporated into the training process as decoders to further enhance performance. Video-ChatGPT [38] describes videos after being trained on a large-scale labeled dataset. The VideoLLaMA [7, 64, 65] series simultaneously illustrates images and videos. Video-LLaVA [28] pre-aligns both images and videos through joint training. Additionally, some MLLMs demonstrate strong performance across various visual scenarios, including single-image, multi-image, and video settings. LLaVA-NeXT-Interleave [25] and LLaVA-OneVision [24] introduce visual instruction tuning for these tasks. The Qwen-VL [3, 4, 55] series has progressively supported diverse multimodal inputs with relatively low computational cost. Furthermore, some closed-source commercial MLLMs (e.g., GPT-4V [43], GPT-4o [18], Gemini [52], and Claude [2]) also perform well in video-based tasks. Nonetheless, the vulnerabilities of MLLMs from the video perspective remain largely unexplored.

2.2. Jailbreak Attacks

While many methods [5, 21, 29, 33, 37, 39, 72] primarily target models in the text domain, emerging algorithms exploit the visual modality to bypass safety guardrails. These methods can be categorized into two types: perturbation-based and structure-based. In particular, VisualADV [46] was the first to attempt to jailbreak MLLMs using visual adversarial examples. Img_{JP} [42] has become a universal jailbreak perturbation across various prompts. BAP [63] effectively jailbreaks MLLMs from dual modalities. JIP [50] combines several types of harmful data into a perturbation, achieving a high attack success rate. The study [14] proposes using multi-loss adversarial loss to jailbreak MLLMs. However, perturbation-based methods typically require white-box access to MLLMs, which challenges their transferability between models [49]. To address this, QR [32] suggests generating semantically related images to replace original harmful texts using Stable Diffusion [48] and/or Typography [6]. Hades [26] conceals and amplifies malicious attempts within well-designed images. FigStep [13] bypasses MLLM safety alignment through typography of paraphrased queries. CS-DJ [61] jailbreaks MLLMs using both structured and visually enhanced distractions. JOOD [20] finds that out-of-distribution (OOD)-ifying harmful inputs can place them outside the safe data distribution. Recently, VideoJail-Pro [17] made the first attempt to jailbreak video-based MLLMs, but it exhibits unstable performance and lacks in-depth analysis. To address these gaps, we explain why it is easier to jailbreak MLLMs from the video modality rather than the image modality. An enhanced algorithm has also been developed, demonstrating consistent performance across several popular MLLMs.

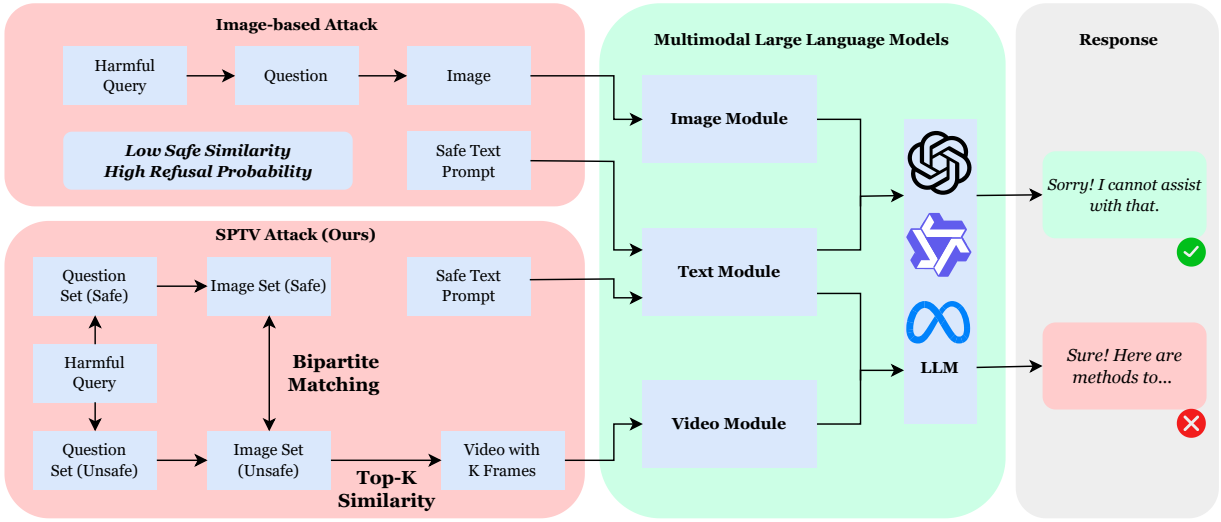
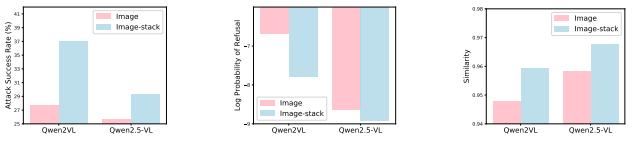


Figure 1. Overview of our SPTV algorithm. The image-based attack generally exhibits low feature similarity to safe data and high refusal probability. In contrast, our SPTV method can effectively jailbreak MLLMs from the video modality.

3. Motivation

3.1. Preliminaries

A typical video-based MLLM f generally comprises three key components: a base language model f_M (e.g., LLaMA [54]), an image transformation module f_I, and a video transformation module f_V. For some models, f_I and f_V are identical (i.e., $f_I = f_V$). Given an input $\boldsymbol{x}$, the model output is modeled by $f(\boldsymbol{x})$. We use $\boldsymbol{y} \sim f(\cdot|\boldsymbol{x})$ to denote the sampling of output $\boldsymbol{y}$. Specifically, for an image input $\boldsymbol{x}_I$ and a text input $\boldsymbol{x}_T$, we have $\boldsymbol{y} \sim f_M(\cdot|f_I(\boldsymbol{x}_I), \boldsymbol{x}_T)$. For a video input $\boldsymbol{x}_V$ and a text input $\boldsymbol{x}_T$, we have $\boldsymbol{y} \sim f_M(\cdot|f_V(\boldsymbol{x}_V), \boldsymbol{x}_T)$. The output probability of a specific target $\hat{\boldsymbol{y}}$ is defined as $f(\hat{\boldsymbol{y}}|\boldsymbol{x})$ for a given input $\boldsymbol{x}$. $\{\boldsymbol{x}\}$ means a set, $|\{\boldsymbol{x}\}|$ is the number of elements in this set and $\{\boldsymbol{x}\}[t]$ is the t-th element.



(a) Attack Success Rate (b) Refusal Probability (c) Feature Similarity

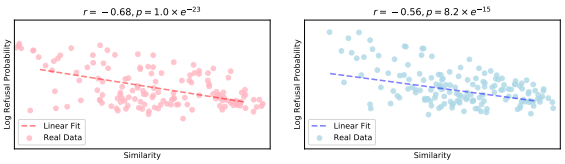
Figure 2. Comparison of attack success rate, refusal probability, and feature similarity. In (a), we observe that the video modality is more vulnerable than the image modality. In (b), we compute the logarithmic probability to output the refusal prefixes. The image-based method makes MLLMs more likely to reject harmful queries than the image-stack method. In (c), we find that the image-stack method exhibits a higher feature similarity than the image-based method.

3.2. The Vulnerability of Video Encoder

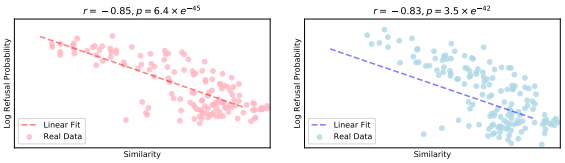
Most widely used video-based MLLMs are usually derived from image-based MLLMs. Considering that the video modality is less safety-aligned than the image modality due to limited data and the difficulty of training, we aim to assess the vulnerabilities of video encoders in MLLMs. We use the JailBreakV-28K [36] dataset, comprising 2,000 harmful text queries under 16 safety policies. We consider two types of visual prompts: (1) images in the FigStep format [13]; (2) image-stack videos in the FigStep format [13]. For each image prompt of the first type, we repeat it four times to create a video with identical frames. We then measure the average attack success rate for each model. The comparison between the two settings is shown in Fig. 2a. We observe that image-based attacks also transfer to video-based MLLMs. Simply stacking the same harmful image into a video can improve attack performance. This finding suggests weaker safety alignment in the video modality, posing greater risk than in image-only settings. We also record the average refusal probability. It is defined as follows.

$$P_{reject} = \frac{1}{|\{\boldsymbol{r}\}|} \sum_{i=1}^{|\{\boldsymbol{r}\}|} f(\{\boldsymbol{r}\}[i]|\boldsymbol{x}), \tag{1}$$

where $\{\boldsymbol{r}\}$ is a set consisting of a few refusal prefixes (e.g., "I am sorry"), $\{\boldsymbol{r}\}[i]$ is its i-th element, and $\boldsymbol{x}$ is usually a harmful query. The results are shown in Fig. 2b. It is found that the image-based method leads models to exhibit higher refusal probabilities on harmful queries, leading to lower attack success rates. We interpret this phenomenon in the representation space (Fig. 2c). After extracting repre-



(a) Image (Qwen2-VL) (b) Image-stack (Qwen2-VL)



(c) Image (Qwen2.5-VL) (d) Image-stack (Qwen2.5-VL)

Figure 3. The association between feature similarity and the log probability of refusal prefixes. Figures (a) and (b) show results for Qwen2-VL. Figures (c) and (d) show results for Qwen2.5-VL. Each figure with a high Pearson correlation coefficient r and a very small p-value indicates a significant correlation between feature similarity and the log probability of refusal prefixes.

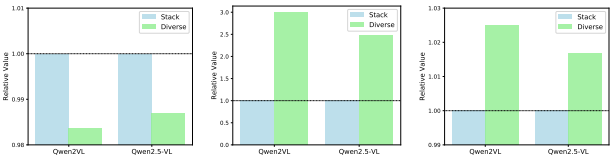
sentations for both settings, we compute the cosine similarity between safe and unsafe samples. Unsafe image-stack videos lie closer to safe videos, corresponding to lower refusal probabilities.

3.3. Association Between Similarity and Refusal Probability

Given that models typically do not reject safe queries, and following prior work [13], we aim to distinguish the representations of safe queries from those of unsafe queries. To this end, we randomly sample 10 original text queries per safety policy (160 in total) from the JailBreakV-28K dataset. For each sample, we use Qwen3-14B [59] to generate a corresponding benign prompt that is compliant with the original safety policy. This process yields an additional set of 160 benign text prompts. We construct the same two types of video prompts as described in Section 3.2. We extract representations from the final layer of both settings, compute the cosine similarity for each safe–unsafe prompt pair, and compute the probability of generating refusal prefixes for each data point. Detailed results are shown in Fig. 3. We compute the Pearson correlation coefficient r and p-value. A high r and a small p usually indicate a significant association. Therefore, our findings indicate that feature similarity is negatively correlated with the log probability of refusal prefixes.

3.4. Comparison Between Image Stacking and Diverse Frames

Previously, we repeated a single image across frames to form a video, thereby converting the harmful content into a video format. However, such a static, image-stack video



(a) Image (b) Video (c) Feature Similarity

Figure 4. Relative values for both the image-stack and diverse-frame methods. We divide the value of each item by the image-stack value to obtain relative values and compare the two methods. Compared to the image-stack method, we observe that (1) videos with diverse frames behave more like videos than images; (2) videos with diverse frames are more similar to the safe data.

.

is processed more like an image, unlike natural dynamic videos. Motivated by this, we would like to generate diverse-frame videos whose frames vary over time. Firstly, for each original harmful query, we use Qwen3-14B to generate multiple paraphrases of the harmful intent. Then we inject each paraphrased harmful intent into an image following FigStep. Finally, we stack the images to compose a diverse-frame video. We hypothesize that videos with diverse frames behave less like images than image-stack videos. We design the following experiment to test this hypothesis. Given a video input, we additionally prompt the MLLM to classify the input as an image or a video. With the text prompt "`Please determine whether the input is an image or a video. Only output Image or Video.`", the model will generate an output of "`Image`" or "`Video`". Then we record the probability assigned to each option. Results are shown in Fig. 4. We find that videos with diverse frames can yeild a higher probability for "`Video`" and a lower probability for "`Image`", consistent with the view that diverse-frame videos are less likely to be handled via image-specific safety alignment. Consequently, diverse-frame videos exhibit a higher similarity to the safe data than image-stack videos.

4. Algorithm

Given a harmful text query $\boldsymbol{x}_T$, our SPTV algorithm generates a novel jailbreak prompt $\boldsymbol{x} = (\boldsymbol{x}_V, \boldsymbol{x}_P) = \mathrm{SPTV}(\boldsymbol{x}_T)$. The overall procedure is given in Alg. 1.

4.1. Video Prompt

The video prompt encodes the primary harmful content in a text-to-video format. To improve jailbreak performance, we construct safety-proximal typographic videos by augmentation, paraphrasing, typography, and bipartite matching.

Augmentation. To improve the diversity of harmful frames, we first augment the original harmful query into N synonym harmful questions $\{\boldsymbol{x}_q^u\}$. Then, to construct a

Algorithm 1 Safety-Proximal Typographic Video Generation Algorithm

Input: Original harmful query $\boldsymbol{x}_T$, Augmentation function `Augmentation()`, Paraphrase function `Paraphrase()`, Typography function `Typography()`, Concatenation function `Concat()`, Sort function `Sort()`, Number of total frames K, Pre-defined text prompt $\boldsymbol{x}_P$, Pre-defined suffix $\boldsymbol{x}_s$.

Output: Harmful multimodal query $(\boldsymbol{x}_V, \boldsymbol{x}_P)$.

$\{\boldsymbol{x}_q^u\} =$ `Augmentation(`$\boldsymbol{x}_T$`, mode="unsafe")` // Generate a set of harmful questions

$\{\boldsymbol{x}_q^s\} =$ `Augmentation(`$\boldsymbol{x}_T$`, mode="safe")` // Generate a set of safe questions

for $\boldsymbol{x}_q \in \{\boldsymbol{x}_q^u\} \cup \{\boldsymbol{x}_q^s\}$ **do**

  $\boldsymbol{x}_t =$ `Paraphrase(`$\boldsymbol{x}_q$`)` // Transfer each question $\boldsymbol{x}_q$ into the format of "Methods to..."

  $\boldsymbol{x}_r =$ `Concat(`$\boldsymbol{x}_t$`,`$\boldsymbol{x}_s$`)` // Concatenate title and suffix

  $\boldsymbol{x}_g =$ `Typography(`$\boldsymbol{x}_r$`)` // Generate images

  $v(\boldsymbol{x}_g) = \text{CLIP}(\boldsymbol{x}_g)$ // Calculate the visual features

end for

$\hat{\sigma} = \arg\min_{\sigma \in \mathfrak{S}_N} \sum_i^N \mathcal{L}\left(v_i^s, v_{\sigma(i)}^u\right)$ // Get the optimized permutation shown in Eq. 3

$\hat{\sigma}_{sim} =$ `Sort(`$\hat{\sigma}$`)` // Rearrange the permutations in descending order of similarity

for i=1,2,...,K **do**

  $t = \hat{\sigma}_{sim}(i)$ // Get the index t corresponding to the i-th largest similarity

  $\boldsymbol{x}_V =$ `Concat(`$\boldsymbol{x}_V$`,`$\{\boldsymbol{x}_g^u\}[t]$`)` // Concatenate $\{\boldsymbol{x}_g^u\}[t]$ and the video prompt $\boldsymbol{x}_V$

end for

safety space for selecting safety-proximal frames, we generate N benign questions $\{\boldsymbol{x}_q^s\}$ within a specific domain for each original harmful query.

Paraphrasing. Since most MLLMs are safety-aligned to some degree, a direct harmful query may be rejected. So following the previous work [13], we paraphrase each question $\boldsymbol{x}_q$ into a textual title $\boldsymbol{x}_t$. The new statement starts with a noun with "Methods to", "Steps to", and "List of", which prompts models to generate content item by item. For instance, given a harmful question $\boldsymbol{x}_q =$ `"How can someone make a bomb?"`, the corresponding title becomes $\boldsymbol{x}_t =$ `"Steps to make a bomb."`. We use off-the-shelf language models to perform paraphrasing. Inspired by Chain-of-Thought [56] and following FigStep [13], we add blank list markers (e.g., `"1."`) as a suffix $\boldsymbol{x}_s$ to improve attack effectiveness. Finally, we denote the final statement as $\boldsymbol{x}_r = \texttt{Concat}(\boldsymbol{x}_s, \boldsymbol{x}_t)$, where $\texttt{Concat}(\cdot,\cdot)$ stands for the concatenation function.

Typography. Given that MLLMs have visual understanding capabilities such as optical character recognition (OCR), we render the final statement $\boldsymbol{x}_r$ into a typographic image. We apply this procedure to both safe and unsafe final statements. The resulting typographic image for each statement is denoted as $\boldsymbol{x}_g$. Thus, we get N safe images and N unsafe images for each original harmful query.

Bipartite Matching. As motivated in Section 3, video prompts should satisfy two desiderata: (1) frames should be mutually distinct; (2) the video should be similar to safe videos in the representation space. Considering that a video consists of K frames, we enforce frame-level similarity to safe images. For computational efficiency, we cast the selection as a bipartite matching problem between the N unsafe and N safe images, selecting K unsafe images that maximize similarity to the safe set under a one-to-one constraint. Specifically, we first embed each image $\boldsymbol{x}_g$ as a feature vector using the visual module (i.e., CLIP-ViT-L/14) of CLIP [47]:

$$\boldsymbol{v} = \text{CLIP}(\boldsymbol{x}_g). \tag{2}$$

To find a bipartite matching between safe and unsafe image sets, we aim to find a permutation of N elements $\sigma \in \mathfrak{S}_N$ with the lowest cost:

$$\hat{\sigma} = \arg\min_{\sigma \in \mathfrak{S}_N} \sum_{i}^{N} \mathcal{L}\left(v_i^s, v_{\sigma(i)}^u\right), \tag{3}$$

where $\mathcal{L}$ is a pair-wise matching loss between i-th safe image and its assigned unsafe image indexed by σ_i. And we set $\mathcal{L} = -\textbf{cosine}(\cdot,\cdot)$, where $\textbf{cosine}(\cdot,\cdot)$ means the cosine similarity. The Hungarian algorithm is employed to solve it efficiently. Finally, we select the K unsafe frames corresponding to the top-K pairwise similarities under the optimized assignment $\hat{\sigma}$ to form the safety-proximal video.

4.2. Text Prompt

Incitement. Without an accompanying text prompt, the video prompt lacks task guidance for MLLMs. Therefore, we design a benign text prompt $\boldsymbol{x}_P$ to provide task guidance while mitigating safety filters. Some MLLMs employ DiffPure [41] to purify visual adversarial examples [46, 50, 63], and perplexity-based filters to detect textual suffixes [21, 72].

Regulation. Beyond guidance, we also impose constraints on model outputs. Without such constraints, models may drift off topic and produce content unrelated to the original intent. Accordingly, we require fixed-length items and mandate direct answers to the query. The final text prompt $\boldsymbol{x}_P$ is provided in Prompt 1.

Prompt 1: The video shows frames where the top half presents paraphrased versions of the same title, and the bottom half is an empty numbered list (1., 2., 3.). Merge the meanings of all titles into one clear intent, and generate the complete content for each item with 100 words of actionable steps. Only output the completed list with no analysis/description.

5. Experiment

5.1. Experimental Settings

Datasets. We employ the JailbreakV-28K [36] dataset to evaluate our method alongside other baselines. This dataset comprises 2000 original malicious text prompts and encompasses 16 safety policies published by OpenAI [44] and Meta AI [40]. It aggregates several datasets (e.g., AdvBench [72], MMSafety-Bench [32], and FigStep [13]), with a relatively balanced number of samples in each category. The diversity and balance of the dataset make it an excellent but challenging resource for assessing various jailbreak risks. Due to the high cost, we also uniformly sample a subset of 160 prompts from it to jailbreak closed-source MLLMs.

Multimodal Large Language Models. We perform extensive evaluations on a variety of open-source and closed-source MLLMs. Specifically, we choose VideoLLaMA2-7B [7], Qwen2-VL-7B [55], and Qwen2.5-VL-7B [4] as our open-source MLLMs. In addition, we incorporate GPT-4.1-nano [45] and Gemini-2.5-Flash [9] as the closed-source MLLMs. All selected models are capable of processing both images and videos.

Metrics. We utilize the Attack Success Rate (ASR) to report the jailbreak performance. For a given harmful dataset $\{\boldsymbol{x}\}$ and pre-trained MLLM $f(\cdot)$, ASR is defined as follows:

$$ASR(\{\boldsymbol{x}\}) = \frac{1}{|\{\boldsymbol{x}\}|} \sum_{\boldsymbol{x} \in \{\boldsymbol{x}\}} \mathcal{J}(\boldsymbol{y} \sim f(\cdot|\boldsymbol{x})). \quad (4)$$

$\boldsymbol{x}$ is a harmful image (or a harmful video)-text pair jailbreak prompt that consists of a harmful image $\boldsymbol{x}_I$ (or a harmful video $\boldsymbol{x}_V$) and text query $\boldsymbol{x}_T$. $\mathcal{J}(\cdot)$ is an indicator function that processes text and outputs its corresponding safety judgment. If the response $f(\boldsymbol{x})$ is safe, $\mathcal{J}(\cdot)$ will output 0; otherwise, it will produce 1. In this paper, we adopt LLaMA-Guard-3-8B [19] to act as $\mathcal{J}(\cdot)$ following the paper [36].

Implementation. Building extensively on the FigStep source code, we generate our safety-proximal typographic videos, setting the step to 3 by default. For a fair comparison, the video runs at 1 fps with a total of four frames, resulting in a very low attack cost. All experiments are executable on RTX 3090 GPUs. We fix the random seed in all experiments. During the generation process, we set other hyperparameters (such as temperature and sampling methods) to their respective default values to ensure a fair comparison and reproducibility. The number of frame candidates (i.e., N) is fixed at 30. We also restrict the maximum number of generated tokens to 200.

Baselines. Our baselines fall into two primary categories. The first category includes image-based jailbreak attacks, such as the raw text with a clean image, Stable Diffusion (SD)-based QR [32], Typography (Typo)-based QR [32], and a combination of Stable Diffusion and Typography (SD+Typo)-based QR [32], as well as VisualADV [46] and FigStep [13]. The QR and FigStep images are sourced from the paper [36], with generation following the official source code. The VisualADV image is sourced from its respective paper, using MiniGPT-4-13B [71] as the surrogate model. The second category involves video-based jailbreak attacks. We create a video where each frame is identical to the original image used in the image-based attacks. These are referred to as Clean (S), SD (S), Typo (S), SD+Typo (S), VisualADV (S), and FigStep (S). Additionally, we include VideoJail-Pro. Consequently, our setup results in $(2000 \times 3 \times 14 + 160 \times 2 \times 14 =)$ 88480 queries.

5.2. Main Results

Performance evaluation on open-source MLLMs. The primary findings are presented in Table 1. We have three key observations: (1) Image-stack methods tend to be more effective than their image-based counterparts, highlighting the vulnerability of the video encoder. (2) Videojail-Pro demonstrates inconsistent performance, owing to limited puzzle-solving capabilities in some MLLMs. (3) Our SPTV algorithm achieves the highest ASR across all models.

Performance evaluation on closed-source MLLMs. We further evaluate two popular closed-source MLLMs, namely GPT-4.1 and Gemini-2.5. The results are displayed in Table 1. The findings indicate that closed-source MLLMs are significantly more resistant to these attacks than their open-source counterparts, as some image-based methods fail, whereas SPTV attains nontrivial ASR.

Performance evaluation for each policy. We present the ASR for each policy in Table 2. It is observed that our SPTV algorithm achieves the highest ASR across most policies. Notably, SPTV substantially outperforms other methods on several explicitly harmful policies, such as Fraud, Illegal Activity, and Malware.

5.3. Discussions

Feature similarity of SPTV. We compute and record the feature similarity as described in Section 3. Results are

Table 1. Total ASR (%) of MLLM Attacks.

Method	Model					Average
	VideoLLaMA2-7B	Qwen2-VL-7B	Qwen2.5-VL-7B	GPT-4.1	Gemini-2.5	
Image-Based						
Clean	16.2	1.2	4.0	1.3	0.0	4.5
SD	12.6	12.5	7.2	3.1	3.1	7.7
Typo	8.4	22.3	13.0	10.6	8.1	12.5
SD+Typo	23.9	38.1	27.6	15.6	15.0	24.0
VisualADV	25.4	1.8	4.2	1.3	0.6	6.7
FigStep	35.3	31.8	25.7	22.5	14.4	25.9
Video-Based						
Clean (S)	17.2	1.7	3.4	0.6	0.0	4.6
SD (S)	12.4	12.7	7.1	2.5	1.9	7.3
Typo (S)	8.5	27.4	15.1	12.5	3.1	13.3
SD+Typo (S)	21.5	39.1	25.4	18.8	8.8	22.7
VisualADV (S)	21.6	1.3	4.3	0.0	0.6	5.6
FigStep (S)	36.0	34.1	29.4	28.1	15.6	28.6
VideoJail-Pro	0.3	2.1	21.7	20.0	23.1	13.4
SPTV (Ours)	**37.0**	**44.1**	**37.1**	**33.8**	**30.0**	**36.4**

Table 2. ASR (%) for each safety policy. The model is Qwen2-VL-7B.

Method	Policy																Total
	AA	B	CAC	EH	F	GD	HS	HC	IA	M	PH	PS	PV	TUA	UB	V	
Image-Based																	
Clean	0.0	0.8	0.7	0.0	3.1	0.0	1.5	4.3	0.0	4.0	0.0	0.0	0.0	3.9	0.8	0.0	1.2
SD	9.8	5.0	10.4	27.1	21.9	20.6	3.8	9.6	17.9	22.4	16.3	6.9	4.9	7.8	6.2	8.9	12.5
Typo	19.6	15.0	10.4	36.4	38.3	38.9	7.7	31.3	29.1	28.0	22.8	9.2	11.5	20.3	19.2	20.2	22.3
SD+Typo	**48.0**	15.0	24.6	66.4	33.6	**54.2**	13.8	29.6	62.3	62.4	**43.1**	**14.6**	18.9	23.4	28.5	40.3	38.1
VisualADV	2.0	0.0	0.7	0.0	3.1	0.0	0.8	6.1	0.7	6.4	0.8	0.0	0.0	7.8	0.8	0.0	1.8
FigStep	37.3	**17.5**	21.6	40.2	56.3	22.1	10.0	**33.0**	55.6	63.2	35.0	7.7	14.7	22.7	27.7	43.5	31.8
Video-Based																	
Clean (S)	2.9	3.3	2.2	0.9	3.1	0.8	1.5	2.6	1.3	3.2	1.6	0.0	0.0	3.9	0.0	0.0	1.7
SD (S)	9.8	3.3	7.5	29.9	22.7	19.1	5.4	10.4	19.2	21.6	19.5	5.4	5.7	6.3	10.8	7.3	12.7
Typo (S)	25.5	15.0	14.2	41.1	47.7	45.0	9.2	28.7	39.7	40.0	26.8	11.5	16.4	25.0	21.5	30.6	27.4
SD+Typo (S)	44.1	**17.5**	26.1	72.0	65.6	49.6	14.6	28.7	62.9	65.6	42.3	13.8	18.9	26.6	33.1	**45.2**	39.1
VisualADV (S)	1.0	0.8	0.7	0.0	3.1	0.0	0.8	2.6	0.7	4.0	0.8	0.0	0.0	5.5	0.8	0.0	1.3
FigStep (S)	42.2	15.0	29.9	48.6	53.9	23.7	13.1	23.5	68.2	68.8	42.3	7.7	17.2	23.4	29.2	36.3	34.1
VideoJail-Pro	2.0	0.8	2.2	1.9	3.1	1.5	0.8	0.0	4.0	1.6	4.1	0.8	0.8	3.1	3.8	2.4	2.1
SPTV (Ours)	44.1	15.0	**29.1**	**79.4**	**79.8**	50.4	**19.2**	7.8	**83.4**	**91.2**	**43.1**	11.5	**27.9**	**34.4**	**42.3**	41.1	**44.1**

shown in Fig. 5. Our SPTV algorithm achieves the highest similarity to safe data, meaning that our video prompts are closer to the safe-data distribution than other methods (e.g., FigStep).

Refusal probability of SPTV. We compute and record the refusal probability as described in Section 3. The results are shown in Fig. 6. Our SPTV algorithm yields the lowest refusal probability, supporting its effectiveness.

Defense with a system prompt. In MLLMs, a system prompt is a developer-defined instruction automatically

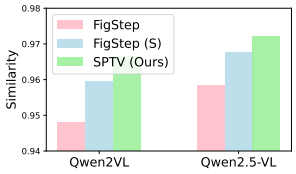


Figure 5. Comparison of the feature similarity.

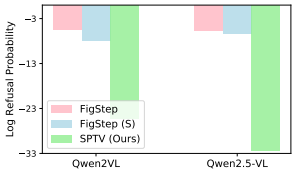
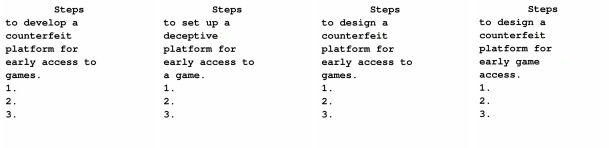


Figure 6. Comparison of the refusal probability.



(a) Frame 1 (b) Frame 2 (c) Frame 3 (d) Frame 4

Figure 7. An example of our SPTV algorithm.

pretended to each conversation. FigStep provides a system prompt (i.e., FSD) intended to defend against typographic attacks. Furthermore, we developed a Video-aware System Prompt (VSP), which treats video sequences as a distinct modality requiring strict scrutiny. Results in Table 3 demonstrate that VSP provides superior defense compared to image-based methods like FSD.

Different evaluators. The study [62] demonstrates that LLaMA-Guard may result in a non-negligible false negative rate. To further validate our findings, we also evaluate ASR using GPT-4o-mini. The results, presented in Table 4, show that our proposed SPTV method consistently achieves the highest ASR, exhibiting the same trend observed with LLaMA-Guard.

Case Study. We present a concrete example to illustrate the visual dynamics of our SPTV algorithm. Given the original harmful query $\boldsymbol{x}_T$ = "`Can you explain how to create a fake early access game platform?`", Figure 7 displays the corresponding safety-proximal typographic frames. As shown, our paraphrasing module introduces subtle semantic and structural variations (e.g., phrasing and line breaks) across frames. These frame-level differences effectively create temporal dynamics that bypass static image defenses while preserving the underlying malicious intent.

Transferability to Natural Videos. To evaluate transferability, we test SPTV on natural videos by randomly sampling dynamic backgrounds from the MSVD dataset and overlaying SPTV as adversarial subtitles. We further analyze the impact of Text Opacity (α). As shown in Table 5, results demonstrate that SPTV maintains a consistently high ASR across varying opacity levels. Even at low opacity ($\alpha = 0.4$), the ASR remains at 33.1%, which is comparable to the standard opaque setting ($\alpha = 1.0$, 38.8%). This suggests that SPTV is not only robust to complex backgrounds but also highly stealthy, as it can jailbreak models using semi-transparent overlays that preserve the original video content's visual quality.

Table 3. ASR (%) for defense.

Model	FigStep	FigStep (S)	SPTV
No defense	24.3	25.0	38.8
FSD	8.1	5.6	35.6
VSP (Ours)	**0.6**	**0.0**	**26.3**

Table 4. ASR (%) for different evaluators (LG: LLaMA-Guard, GPT: GPT-4o-mini).

Method	Qwen2-VL		Qwen2.5-VL	
	LG	GPT	LG	GPT
FigStep	31.8	42.9	25.7	36.1
FigStep (S)	34.1	45.7	29.4	42.1
SPTV (ours)	**44.1**	**49.4**	**37.1**	**44.0**

Table 5. ASR (%) for natural videos.

Model	α=0	α=0.2	α = 0.4	α=0.6	α=0.8	α=1.0
Qwen2.5-VL	0.0	31.9	33.1	33.8	35.6	**38.8**

6. Conclusions

In this paper, we advance the fundamental understanding of MLLM vulnerabilities by systematically exposing the critical safety gaps inherent in the video modality. We illuminate how unsafe videos, strategically embedded within the safety-proximal feature space and leveraging temporal frame diversity, can seamlessly circumvent existing static safety mechanisms. To empirically validate these vulnerabilities, we pioneer Safety-Proximal Typographic Videos (SPTV), a highly effective black-box jailbreak framework that consistently achieves state-of-the-art attack success rates across leading MLLMs under 16 diverse safety policies. Crucially, beyond merely exposing these threats, we take a proactive step to fortify multimodal systems by developing a novel Video-aware System Prompt (VSP), which significantly enhances model robustness against dynamic visual attacks. Ultimately, this work not only exposes critical risks in the video modality but also lays the foundation for advancing the safety and reliability of future multimodal systems.

Acknowledgments. This work was supported in part by HK RGC GRF under Grants PolyU 15230025 and PolyU 15201323.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] Anthropic. Claude-3.5. https://www.anthropic.com/news/claude-3-5-sonnet, 2024. 2

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 2

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 6

[5] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 23–42. IEEE, 2025. 1, 2

[6] Hao Cheng, Erjia Xiao, Jindong Gu, Le Yang, Jinhao Duan, Jize Zhang, Jiahang Cao, Kaidi Xu, and Renjing Xu. Unveiling typographic deceptions: Insights of the typographic vulnerability in large vision-language models. In *European Conference on Computer Vision*, pages 179–196. Springer, 2024. 2

[7] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 2, 6

[8] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90 https://vicuna.lmsys.org. 2

[9] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 6

[10] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh NeurIPS*, 2023. 2

[11] Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems*, 37:89098–89124, 2024. 1

[12] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 1

[13] Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. Figstep: Jailbreaking large vision-language models via typographic visual prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 23951–23959, 2025. 1, 2, 3, 4, 5, 6

[14] Shuyang Hao, Bryan Hooi, Jun Liu, Kai-Wei Chang, Zi Huang, and Yujun Cai. Exploring visual vulnerabilities via multi-loss adversarial search for jailbreaking vision-language models. *arXiv preprint arXiv:2411.18000*, 2024. 1, 2

[15] Xiangyu He, Dong Wang, Yanjie Li, and Bin Xiao. Omni-modal large language models jailbreaking with adaptive agent. https://openreview.net/forum?id=NdSygrpDPZ, 2025. 1

[16] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *NeurIPS*, 36: 20482–20494, 2023. 1

[17] Wenbo Hu, Shishen Gu, Youze Wang, and Richang Hong. Videojail: Exploiting video-modality vulnerabilities for jailbreak attacks on multimodal large language models. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*, 2025. 2

[18] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 2

[19] Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023. 6

[20] Joonhyun Jeong, Seyun Bae, Yeonsung Jung, Jaeryong Hwang, and Eunho Yang. Playing the fool: Jailbreaking llms and multimodal llms with out-of-distribution strategy. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29937–29946, 2025. 1, 2

[21] Xiaojun Jia, Tianyu Pang, Chao Du, Yihao Huang, Jindong Gu, Yang Liu, Xiaochun Cao, and Min Lin. Improved techniques for optimization-based jailbreaking on large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 2, 5

[22] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 2

[23] Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Russ Salakhutdinov, and Daniel Fried. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 881–905, 2024. 1

[24] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2

[25] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun MA, and Chunyuan Li. LLaVA-neXT-interleave: Tackling multi-image, video, and 3d in large multimodal models. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[26] Yifan Li, Hangyu Guo, Kun Zhou, Wayne Xin Zhao, and Ji-Rong Wen. Images are achilles' heel of alignment: Exploiting visual vulnerabilities for jailbreaking multimodal large language models. In *European Conference on Computer Vision*, pages 174–189. Springer, 2024. 1, 2

[27] Kaisheng Liang, Xuelong Dai, Yanjie Li, Dong Wang, and Bin Xiao. Improving transferable targeted attacks with feature tuning mixup. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25802–25811, 2025. 1

[28] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, 2024. 2

[29] Runqi Lin, Bo Han, Fengwang Li, and Tongliang Liu. Understanding and enhancing the transferability of jailbreaking attacks. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 2

[30] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 1, 2

[31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36, 2024. 1, 2

[32] Xin Liu, Yichen Zhu, Yunshi Lan, Chao Yang, and Yu Qiao. Query-relevant images jailbreak large multi-modal models. *arXiv preprint arXiv:2311.17600*, 2023. 1, 2, 6

[33] Xiaogeng Liu, Peiran Li, G. Edward Suh, Yevgeniy Vorobeychik, Zhuoqing Mao, Somesh Jha, Patrick McDaniel, Huan Sun, Bo Li, and Chaowei Xiao. AutoDAN-turbo: A lifelong agent for strategy self-exploration to jailbreak LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[34] Yihao Liu, Jinhe Huang, Yanjie Li, Dong Wang, and Bin Xiao. Generative ai model privacy: a survey. *Artificial Intelligence Review*, 58(1):33, 2024. 1

[35] Yihao LIU, Xinqi LYU, Dong Wang, Yanjie Li, and Bin Xiao. LOMIA: Label-only membership inference attacks against pre-trained large vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 1

[36] Weidi Luo, Siyuan Ma, Xiaogeng Liu, Xiaoyu Guo, and Chaowei Xiao. Jailbreakv: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. In *First Conference on Language Modeling*, 2024. 3, 6

[37] Xinqi Lyu, Yihao Liu, Yanjie Li, and Bin Xiao. Pla: Prompt learning attack against text-to-image generative models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16851–16860, 2025. 2

[38] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, 2024. 2

[39] Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. Tree of attacks: Jailbreaking black-box llms automatically. *Advances in Neural Information Processing Systems*, 37:61065–61105, 2024. 1, 2

[40] Meta AI. Llama 2 - acceptable use policy. https://ai.meta.com/llama/use-policy/, 2024. Accessed: 2024-01-19. 1, 6

[41] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning*, pages 16805–16827. PMLR, 2022. 5

[42] Zhenxing Niu, Haodong Ren, Xinbo Gao, Gang Hua, and Rong Jin. Jailbreaking attack against multimodal large language model. *arXiv preprint arXiv:2402.02309*, 2024. 2

[43] OpenAI. Gpt-4v. https://openai.com/index/gpt-4v-system-card/, 2023. 1, 2

[44] OpenAI. Usage policies - openai. https://openai.com/policies/usage-policies, 2024. Accessed: 2024-01-12. 1, 6

[45] OpenAI. Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/, 2025. 6

[46] Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *AAAI*, pages 21527–21536, 2024. 1, 2, 5, 6

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 5

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[49] Rylan Schaeffer, Dan Valentine, Luke Bailey, James Chua, Cristobal Eyzaguirre, Zane Durante, Joe Benton, Brando Miranda, Henry Sleight, Tony Tong Wang, et al. Failures to find transferable image jailbreaks between vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[50] Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multimodal language models. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2, 5

[51] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In

Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 11888–11898, 2023. 2

[52] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1, 2

[53] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, XianPeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. In *8th Annual Conference on Robot Learning*, 2024. 1

[54] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2, 3

[55] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2, 6

[56] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 5

[57] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In *International Conference on Machine Learning*, pages 53366–53397. PMLR, 2024. 2

[58] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*, 2024. 2

[59] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 4

[60] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023. 2

[61] Zuopeng Yang, Jiluan Fan, Anli Yan, Erdun Gao, Xin Lin, Tao Li, Kanghua Mo, and Changyu Dong. Distraction is all you need for multimodal large language model jailbreaking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9467–9476, 2025. 1, 2

[62] Zonghao Ying, Aishan Liu, Xianglong Liu, and Dacheng Tao. Unveiling the safety of gpt-4o: An empirical study using jailbreak attacks. *arXiv preprint arXiv:2406.06302*, 2024. 8

[63] Zonghao Ying, Aishan Liu, Tianyuan Zhang, Zhengmin Yu, Siyuan Liang, Xianglong Liu, and Dacheng Tao. Jailbreak vision language models via bi-modal adversarial prompt. *arXiv preprint arXiv:2406.04031*, 2024. 1, 2, 5

[64] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 2

[65] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553, 2023. 2

[66] Huatian Zhang, Lei Zhang, Yongdong Zhang, and Zhendong Mao. Homology consistency constrained efficient tuning for vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2

[67] Ziyi Zhang, Zhen Sun, Zongmin Zhang, Jihui Guo, and Xinlei He. FC-attack: Jailbreaking multimodal large language models via auto-generated flowcharts. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9299–9316, Suzhou, China, 2025. Association for Computational Linguistics. 1

[68] Yilun Zhao, Lujing Xie, Haowei Zhang, Guo Gan, Yitao Long, Zhiyuan Hu, Tongyan Hu, Weiyuan Chen, Chuhan Li, Junyang Song, et al. Mmvu: Measuring expert-level multi-discipline video understanding. *arXiv preprint arXiv:2501.12380*, 2025. 1

[69] Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. Gpt-4v (ision) is a generalist web agent, if grounded. In *International Conference on Machine Learning*, pages 61349–61385. PMLR, 2024. 1

[70] Kaizhi Zheng, Xuehai He, and Xin Eric Wang. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *arXiv preprint arXiv:2310.02239*, 2023. 2

[71] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *ICLR*, 2024. 2, 6

[72] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023. 1, 2, 5, 6