Video Watermarking: Safeguarding Your Video from (Unauthorized) Annotations by Video-based LLMs

Jinmin Li [* 1] Kuofeng Gao [* 1] Yang Bai [2] Jingyun Zhang [3] Shu-Tao Xia [1 4]

Abstract

The advent of video-based Large Language Models (LLMs) has significantly enhanced video understanding. However, it has also raised some safety concerns regarding data protection, as videos can be more easily annotated, even without authorization. This paper introduces Video Watermarking, a novel technique to protect videos from unauthorized annotations by such video-based LLMs, especially concerning the video content and description, in response to specific queries. By imperceptibly embedding watermarks into key video frames with multi-modal flow-based losses, our method preserves the viewing experience while preventing misuse by video-based LLMs. Extensive experiments show that Video Watermarking significantly reduces the comprehensibility of videos with various video-based LLMs, demonstrating both stealth and robustness. In essence, our method provides a solution for securing video content, ensuring its integrity and confidentiality in the face of evolving video-based LLMs technologies.

1. Introduction

Recent developments in multi-modal understanding have been greatly enhanced by combining existing vision models with Large Language Models (LLMs) (Wang et al., 2024; Zhu et al., 2023; Liu et al., 2023; Deng et al., 2024). The combination of these models has led to impressive abilities in managing and interpreting video and language data. However, as video-based LLMs become more prevalent, worries about the safety and reliability of video data have grown, even when accessed without permission.

---

[*]Equal contribution [1]Tsinghua Shenzhen International Graduate School, Tsinghua University [2]Tencent Technology (Beijing) Co.Ltd [3]Tencent WeChat Pay Lab33 [4]Peng Cheng Laboratory. Correspondence to: Yang Bai <baiyang0522@gmail.com>.

Nowadays, in particular, Sora[1] has shown extraordinary performances in creating realistic and imaginative scenes from text instructions, demonstrating the significance of large multi-modal models, especially across video and language modalities. Among them, video-based LLMs (Li et al., 2023b; Zhang et al., 2023; Maaz et al., 2023) have significantly enhanced general video understanding in zero-shot settings and achieved exceptional performance in a wide range of video-related tasks (Tang & Li, 2004; Liu et al., 2006; Aafaq et al., 2019; Ma et al., 2024; 2023), such as video captioning (Venugopalan et al., 2015; Zhang et al., 2021; Seo et al., 2022), video retrieval (Luo et al., 2022; Gabeur et al., 2020; Dong et al., 2021; Ma et al., 2022b), and scene understanding (Cordts et al., 2016; Hu et al., 2021; Wu et al., 2022; Ma et al., 2022a). Yet, the very features that make these models powerful also render them susceptible to misuse and misidentification without proper safeguards.

Given the importance of video data in various applications, from content creation to surveillance, ensuring its protection from unauthorized annotation and misinterpretation of video content and description by video-based LLMs is imperative. To address this, we propose a flow-based multi-modal Video Watermarking to craft adversarial perturbations on video inputs for the first time. A flow-based temporal mask is introduced to select the most effective frames in the video, which is inspired by the video clipping adopted in video understanding tasks, especially for video-based LLMs (Chen et al., 2024; Xue et al., 2022). It can effectively improve the performance of video-based learning due to its focused annotation, reducing complexity, improved temporal understanding, and efficient processing. Motivated by these benefits, we utilize a light-weighted flow-based mechanism, to conduct a similar splitting and selection operation on video frames. Extensive experiments have demonstrated the effectiveness, efficiency, and imperceptibility of our Video Watermarking on four benchmark video-based LLMs and two datasets.

Our methodology harnesses the inherent capabilities of multimodal models to devise watermarks that act as a bulwark against unauthorized exploitation by video-based LLMs. Through an innovative flow-based mechanism, we embed

---

[1]https://openai.com/sora

watermarks into the key selected video frames with such finesse that they are imperceptible to the human eye yet robustly shield the content from LLMs. Our research zeroes in on the strategic deployment of Video Watermarking to bolster the security of video data. We illustrate that the strategic application of these watermarks substantially diminishes the likelihood of unauthorized access and misinterpretation of video content and description by video-based LLMs. This proactive measure not only safeguards the sanctity of the video content but also fortifies the privacy and security of the data, ensuring that it remains inviolable in the digital realm.

In summary, our contribution can be outlined as follows:

- We present the first in-depth study focused on safeguarding video data, specifically the content and description, from unauthorized use by video-based LLMs through our novel Video Watermarking approach. This pioneering work establishes a new paradigm for protecting video content integrity in the era of advanced multi-modal AI.

- Our findings significantly contribute to the understanding of how watermarks can be strategically employed to reinforce the robustness of multi-modal systems. The insights gained are invaluable for the development and deployment of secure large multi-modal models that respect data usage protocols and ethical standards.

- Our extensive experiments validate the efficacy of our Video Watermarking in preventing the misuse of video content and description. By integrating watermarks into a minimal portion of video frames—less than 20%—we effectively thwart unauthorized access and ensure that video data remains protected from arbitrary exploitation.

2. Related Work

2.1. Video-based Large Language Models

Video-based large language models (video-based LLMs) effectively integrate visual and temporal information from video data to gain significant achievement in multiple video-related tasks. Numerous approaches (Li et al., 2023b; Maaz et al., 2023; Zhang et al., 2023) have been proposed to address the challenges associated with video-based LLMs, such as incorporating different architectures and training processes to enhance the models' ability to capture and process complex video information. Concretely, Video-ChatGPT (Maaz et al., 2023) is based on the LLaVA framework and incorporates average pooling to improve the perception of temporal sequences. VideoChat (Li et al., 2023b) employs the QFormer to map visual representations
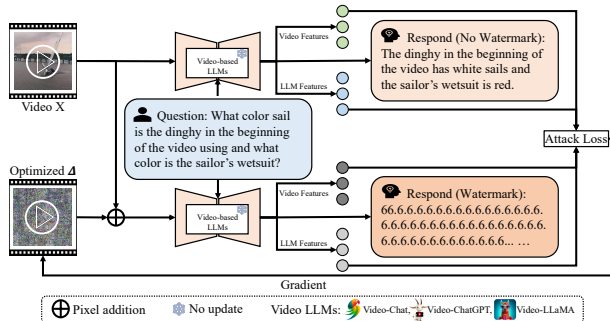

Figure 1. Schematics of our Video Watermarking.

to Vicuna, executing a two-stage training process. Video-LLaMA (Zhang et al., 2023) integrates a frame embedding layer and ImageBind to introduce temporal and audio information into the LLM backbone. This alignment between videos and LLMs facilitates visual context-aware interaction, surpassing the capabilities of LLMs. However, ensuring the secure and ethical use of these models necessitates the integration of protective measures such as watermarking.

2.2. Adversarial Attack

While adversarial attacks (Wang et al., 2022; Wei et al., 2019; Bai et al., 2022; Wu et al., 2023; Zou et al., 2023; Zhao et al., 2024; Li et al., 2024a; Gao et al., 2024b;a;c; Fang et al., 2024a;b; Guo et al., 2024; Li et al., 2023a; Xiao et al., 2024; Li et al., 2023c; 2024b; Bai et al., 2019; 2024b; 2020; Bai et al.; Yang et al., 2024) have been widely studied in classification models, the focus of our research lies in the proactive enhancement of video-based LLMs through watermarking. Inspired by the potential vulnerabilities observed in vision tasks, we shift the paradigm to explore how watermarking can fortify large multi-modal models (Qi et al., 2023; Zhuang et al., 2023; Gong et al., 2023; Yang et al., 2024; Zong et al., 2024; Bai et al., 2024a; Liang et al., 2024) such as vision large language models (VLLMs) (Zhu et al., 2023; Wang et al., 2024; Liu et al., 2023; Radford et al., 2021b) and text-to-image diffusion models (Rombach et al., 2022). Unlike adversarial attacks designed to manipulate models into generating specific outputs, our watermarking approach is crafted to protect the integrity and security of video content within these models. In this paper, we introduce a pioneering watermarking strategy that is specifically tailored to safeguard video-based LLMs, ensuring their robustness against unauthorized use and potential data misuse.

3. Methodology

3.1. Threat model

Goals and capabilities. The goal is to craft an imperceptible adversarial perturbation for videos, which can induce

video-based LLMs to generate an incorrect sequence during the victim model's deployment. Following the most commonly used constraint for the involved perturbation, it is restricted within a predefined magnitude in the l_p norm, ensuring it is difficult to detect.

Knowledge and background. As suggested in (Bagdasaryan et al., 2023; Qi et al., 2023), we assume that the victim video-based LLMs can be accessed in full knowledge, including architectures and parameters. Additionally, we consider a more challenging scenario where the victim video-based LLMs are inaccessible, as detailed in the Appendix.

3.2. Preliminary: the Pipeline of Video-based LLMs

Let $\mathcal{F}_\theta(\cdot)$ represents a victim video-based large language model with parameters θ, composed of a video feature extractor $f_\phi(\cdot)$ and a large language model $g_\psi(\cdot)$. Consider a clean video $\mathbf{X} \in \mathbb{R}^{T \times C \times H \times W}$, where T denotes the number of frames, and C, H, and W represent the channel, height, and width of a specific frame, along with a corresponding user query Q_{text} for the video. To provide a response, video-based large language models $\mathcal{F}(\cdot)$ usually first extract a video feature $Q_{video} = f_\phi(\mathbf{X})$, and subsequently, generate predefined prompts based on a consistent template to concatenate both video features and text queries as follows:

$$\text{USER: } < Q_{text} > < Q_{video} > \text{Assistant:}$$

Then, the predefined prompts are processed by LLMs $g_\psi(\cdot)$ to generate a desired response $Y_{respond} = g_\psi(Q_{text}, Q_{video}) = g_\psi(Q_{text}, f_\phi(\mathbf{X}))$. It is important to mention that, to ensure the loss function remains minimal, we use the hidden state A_{hidden} before the final layer.

3.3. Problem Formulation

The goal of generating watermarking examples $\hat{\mathbf{X}}$ is to mislead video-based LLMs to produce incorrect responses while utilizing the most imperceptible adversarial perturbation Δ, where $\Delta = \hat{\mathbf{X}} - \mathbf{X}$. To balance these two objectives, we introduce a hyper-parameter λ, and formulate the overall objective function as follows:

$$\arg\min_\Delta \lambda\|\Delta\|_{2,1} - \ell(Y, \mathcal{F}_\theta(Q_{text}, \hat{\mathbf{X}})), \quad (1)$$

where Y is the ground truth answer corresponding to Q_{text} and $\mathbf{X}$, as well as $\ell(\cdot, \cdot)$ is the loss function used to measure the difference between the predicted and ground truth answers. Furthermore, in a more realistic scenario, the ground truth answer Y can not always be available. In such cases, we utilize the output approximation $\mathcal{F}_\theta(Q_{text}, \mathbf{X})$ instead of Y.

Therefore, the overall objective function in Eq. 1 can be

further formulated as follows:

$$\arg\min_\Delta \lambda\|\Delta\|_{2,1} - \ell(\mathcal{F}_\theta(Q_{text}, \mathbf{X}), \mathcal{F}_\theta(Q_{text}, \hat{\mathbf{X}})). \quad (2)$$

In particular, the $\ell_{2,1}$ norm (Wei et al., 2019) is employed to quantify the magnitude of the perturbation, which can be defined as follows:

$$\|\Delta\|_{2,1} = \sum_{i}^{T} \|\Delta_i\|_2, \quad (3)$$

where $\Delta_i \in \mathbb{R}^{C \times H \times W}$ represents the i-th frame in Δ. The $\ell_{2,1}$ norm applies the l_1 norm across frames, ensuring the sparsity of generated perturbations. Note that a smaller $\ell_{2,1}$ norm value corresponds to more perceptible perturbations, which are hard to notice by human inspectors.

3.4. Optimization Objective

Our proposed Video Watermarking is to induce video-based LLMs to generate *incorrect* responses with imperceptible adversarial perturbations. Two losses are proposed from the perspective of video features $\ell_{video}(Q_{video}, \hat{Q}_{video})$ in Eq. 4 and LLM features $\ell_{LLM}(A_{hidden}, \hat{A}_{hidden})$ in Eq. 5. Moreover, inspired by the idea that video clipping can improve video comprehension by selecting the most effective frames, a flow-based temporal mask $\mathbf{M}_f$ is proposed to carry out a similar selection process on video frames. By utilizing the flow of a video, the proposed flow-based temporal mask $\mathbf{M}_f$ can filter out similar frames, ensuring the effectiveness of imperceptible adversarial perturbations while achieving increased sparsity.

Video Features Loss. Video-based LLMs first use a video feature extractor $f_\phi(\cdot)$ to extract spatiotemporal video features $Q_{video} = f_\phi(\mathbf{X})$. We simply adopt MSE loss to measure the distance of video features between the clean video $\mathbf{X}$ and the adversarial video $\hat{\mathbf{X}}$. Hence, the video features loss can be formulated as:

$$\begin{aligned}
\ell_{video} &= \frac{1}{n} \sum_{i=1}^{n} (Q_{video_i} - \hat{Q}_{video_i})^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} (f_\phi(\mathbf{X})_i - f_\phi(\hat{\mathbf{X}})_i)^2,
\end{aligned} \quad (4)$$

where n is the total number of elements in the features, with Q_{video_i} and $\hat{Q}_{video_i}$ being the i-th elements of the clean and watermarking video features, respectively.

LLM Features Loss. In addition to the deviation of the original feature space in video domains of video-based LLMs, we also consider that in textual domains to enhance the watermarking effect. Given a hidden state from the final layer of LLMs $A_{hidden} = g_\psi(Q_{text}, Q_{video}) =$

$g_\psi(Q_{text}, f_\phi(\mathbf{X}))$, the LLM features loss between the clean video $\mathbf{X}$ and the watermarking video $\hat{\mathbf{X}}$ can be formulated as follows:

$$
\begin{aligned}
\ell_{LLM} &= \frac{1}{n}\sum_{i=1}^{n}(A_{hidden_i} - \hat{A}_{hidden_i})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(g_\psi(Q_{text}, Q_{video})_i - g_\psi(Q_{text}, \hat{Q}_{video})_i)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(g_\psi(Q_{text}, f_\phi(\mathbf{X}))_i - g_\psi(Q_{text}, f_\phi(\hat{\mathbf{X}}))_i)^2,
\end{aligned}
\tag{5}
$$

where A_{hidden_i} and $\hat{A}_{hidden_i}$ are the i-th elements of the clean and watermarking LLM features respectively, and n is the total number of elements in the features.

Flow-based Temporal Mask. Video-based LLMs (Chen et al., 2024; Xue et al., 2022) adopt video clipping to split and select the most effective frames in a video, which can enhance video understanding. Inspired by these advantages, we propose a flow-based temporal mask (FTM), $\mathbf{M}_f$, to perform a similar selection process on video frames. This flow-based mask targets the top K frames with the most significant movement and changes. See Sec. 4.3 for a detailed discussion. Specifically, we initialize a binary mask $\mathbf{M}_f$ of the same length as the number of frames in the video. Then use LiteFlowNet (Hui et al., 2018) to compute the flow magnitude for each frame. Finally, this binary mask $\mathbf{M}_f$ assigns a value of 1 to the top K frames with the largest flow and 0 to the remaining frames. Combined with $\mathbf{M}_f$, our proposed FTM can achieve more sparse adversarial perturbations in both temporal and spatial domains. For temporal sparsity, a flow-based temporal mask $\mathbf{M}_f$ on the video is adopted to ensure that some frames remain unperturbed. For spatial sparsity, the $\ell_{2,1}$ norm of adversarial perturbations in Eq. 3 is employed to constrain the spatial perturbation magnitude in each frame.

Overall Optimization Objective. To sum up, combined flow-based temporal mask $\mathbf{M}_f$ with two proposed watermarking loss functions (ℓ_{video} and ℓ_{LLM}), the overall objective function in Eq. 1 can be further formalized as:

$$
\begin{aligned}
\arg\min_\Delta \lambda_1\|\mathbf{M}_f\Delta\| &- \lambda_2\ell_{video}(f_\phi(\mathbf{X}), f_\phi(\mathbf{X} + \mathbf{M}_f\Delta)) \\
-\lambda_3\ell_{LLM}&(g_\psi(Q_{text}, f_\phi(\mathbf{X})), g_\psi(Q_{text}, f_\phi(\mathbf{X} + \mathbf{M}_f\Delta))),
\end{aligned}
\tag{6}
$$

where $\mathbf{M}_f \in \{\mathbf{0}, \mathbf{1}\}^{T \times C \times H \times W}$ represents the flow-based temporal mask. $\lambda_1, \lambda_2, \lambda_3$ correspond to the three loss weights, which aim to balance them during the optimization. Our overall watermarking procedure is described in Algorithm 1 in the Appendix.

4. Experiments

4.1. Implementation Details

Models and datasets. We assess open-source and state-of-the-art video-based LLMs such as Video-ChatGPT and VideoChat (VideoChat is in the Appendix), ensuring reproducibility of our results. Concretely, we adopt Video-ChatGPT and VideoChat with a LLaMA-7B LLM (Touvron et al., 2023). In line with the Video-ChatGPT methodology, we curate a test set based on the ActivityNet-200 and MSVD-QA datasets. The questions and answers within our test set, derived from the ActivityNet-200 and MSVD-QA datasets, are meticulously hand-annotated to ensure accuracy and quality.

Baselines. For evaluation, we design three spatial baselines, including videos with random perturbations, black videos with all pixel values set to 0, and white videos with all pixel values set to 1. In addition, we compare our proposed flow-based temporal mask with two straightforward temporal mask methods, serving as temporal mask baselines: the sequence temporal mask and the random temporal mask. Specifically, the sequence temporal mask consists of a continuous sequence of frame indices, while the random temporal mask comprises a randomly chosen sequence of frame indices.

Metrics. We utilize a variety of evaluation metrics to assess the robustness of the models.

(a) CLIP Score. CLIP (Radford et al., 2021a) score characterizes the semantic similarity between the adversarial answer and the ground-truth answer.

(b) Image Captioning Metrics. Various metrics such as BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015) are used to evaluate the quality of the adversarial answer generated by the model.

(c) GPT Score. Following Video-ChatGPT (Maaz et al., 2023) and Video-LLaMA (Zhang et al., 2023). We also employ an evaluation pipeline using the GPT-3.5 and GPT-4 models.

(d) Sparsity. Sparsity refers to the ratio of frames without perturbations (clean frames) to the total number of frames in a specific video. The sparsity M_{spa} is calculated as $M_{spa} = 1 - K/T$, where K represents the number of watermarking frames, and T is the total number of frames in a video.

4.2. Main Results

Quantitative Evaluation. As depicted in the left of Table 1, the random, black, white, and flow-based watermarking all significantly decrease the clip score and image caption score compared to the original clean videos. Among these, our

Table 1. Video Watermarking against Video-ChatGPT on the ActivityNet-200 dataset and the MSVD-QA dataset: Comparison of CLIP score, image caption metrics and GPT score for different watermarking types. Random spatial watermarking denotes random perturbations added to video frames, and Black spatial watermarking and White spatial watermarking denote video frames being all 0 and all 1, respectively. The sparsity of the temporal mask is set to 0. Δ: the mean of the modified pixels.

Dataset	Type	Δ	Clip Score ↓		Image Caption ↓		GPT-3.5 ↓		GPT-4 ↓	
			RN50	RN101	BLEU	ROUGE-L	Accurate	Score	Accurate	Score
ActivityNet	Clean	0	0.7817	0.7827	0.2029	0.4820	0.50	3.20	0.33	2.10
	Random	8	0.7637	0.7681	0.1986	0.4793	0.45	3.10	0.33	2.03
	Black	100	0.7661	0.7676	0.1691	0.4570	0.30	2.50	0.17	0.96
	White	148	0.7564	0.7534	0.1689	0.4545	0.31	2.60	0.19	1.24
	Ours	8	**0.6211**	**0.6274**	**0.1336**	**0.3694**	**0.20**	**1.64**	**0.13**	**0.88**
MSVD-QA	Clean	0	0.8322	0.8180	0.3864	0.6843	0.62	3.84	0.60	3.12
	Random	8	0.8249	0.8141	0.4107	0.7042	0.58	3.72	0.60	3.08
	Black	110	0.8145	0.7902	0.3548	0.6478	0.46	3.26	0.40	2.12
	White	142	0.8057	0.8090	0.3969	0.6736	0.48	3.36	0.44	2.28
	Ours	8	**0.7337**	**0.7181**	**0.3240**	**0.5746**	**0.36**	**2.92**	**0.34**	**1.84**
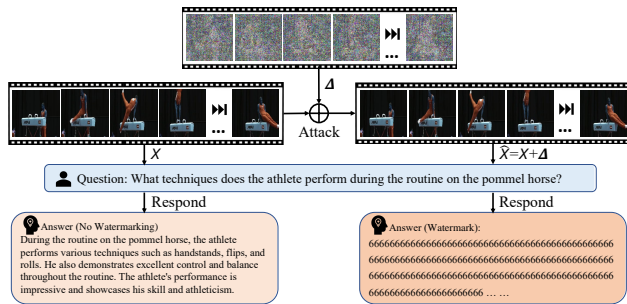
Figure 2. Watermarking videos generated for Video-ChatGPT.

(a) Similarity of adjacent frames

(b) Similarity of answer and current frame

Figure 3. Relationship between optical flow and key frames. 'Clip Score of Adjacent Frames' describes the similarity between the current frame and its adjacent frames, the smaller this score is the more different the current frame is. 'Clip Score of Answer and Current Frame' indicates the similarity between the current frame and the answer corresponding to the user's input question, the larger the score indicates that the current frame contains more information about the answer. **The frames selected by flow-based masks in our Video Watermarking are key frames in the video.**

proposed Video Watermarking yields the most significant results. As demonstrated in the right of Table 1, the random, black, white, and flow-based watermarking all significantly decrease the GPT score and accuracy. Among them, the Video Watermarking achieves the most substantial results. Importantly, our flow-based watermarking is paramount for data protection. It completely conceals the data information, ensuring that it is effectively shielded from unauthorized access and misuse. This robust approach to data obfuscation maintains the confidentiality of the content while preserving the viewing experience for legitimate users.

Qualitative Evaluation. We also present qualitative examples (see Fig. 2) of the watermarking videos, in which the model produces garbled responses without any meaningful content. Fig. 2 vividly illustrates the chaos induced in the model's responses by our subtle and imperceptible watermarking.

4.3. Discussions

Essence of Flow-based Masks. We address the essence of flow-based temporal masks in our Video Watermarking. The flow-based masks by selecting key frames method is a powerful tool for video understanding and manipulation,
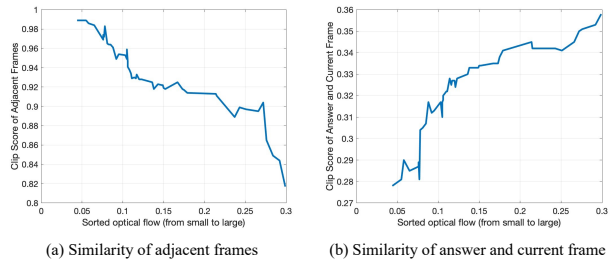
which allows for precise control over specific elements in a video sequence, making it easier to edit and manipulate the video in a variety of ways. We use the clip image-image score of adjacent frames and clip text-image score between the answer and current frame to assess the importance and non-fungibility of our selected frames in Video Watermarking, where a smaller clip image-image score suggests less similarity between a frame and its adjacent frames, and a bigger clip text-image score suggests more similarity between the current frame and the answer of the user input. As depicted in Fig. 3, a larger optical flow corresponds to a higher inconsistency between the current frame and its neighboring frames, while containing more information about the answer. This observation suggests that the frames selected using our Video Watermarking are crucial frames in the video, and watermarking them will be more effective.

Transfer-based Black-box Watermarking

5

Table 2. Black-box watermarking against VideoChat (Li et al., 2023b) on the ActivityNet-200 (Caba Heilbron et al., 2015) dataset: Comparison of image caption metrics and GPT score for different watermarking types. The sparsity of the temporal mask is set to 0. Δ: the mean of the modified pixels. We apply the watermarking video on Video-ChatGPT (Maaz et al., 2023) and directly transfer it to VideoChat (Li et al., 2023b).

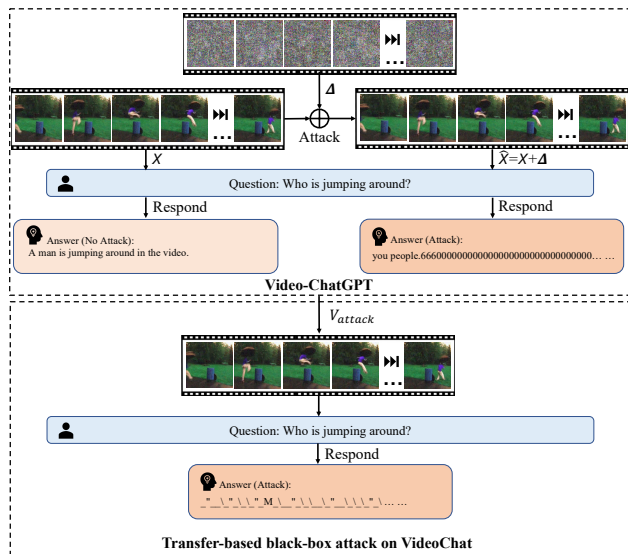| Type	Δ	Image Caption ↓			GPT-3.5 ↓		GPT-4 ↓	
		BLEU	ROUGE	CIDEr	Accurate	Score	Accurate	Score
Clean	0	0.0765	0.3358	0.3379	0.35	2.87	0.28	1.79
Transfer-based watermarking	2	**0.0638**	**0.2492**	**0.2870**	**0.08**	**1.56**	**0.10**	**1.32**

Figure 4. Transfer-based black-box watermarking on VideoChat.

In addition to white-box watermarking, we have also investigated the transferability of these watermarking. We conduct a black-box watermarking on VideoChat (Li et al., 2023b). Specifically, we employ the Video Watermarking method to perform a white-box watermarking on Video-ChatGPT (Maaz et al., 2023), resulting in the watermarking video V_{attack}. This video V_{attack} is then directly used as input for VideoChat (Li et al., 2023b), with the experimental results displayed in Table 2. It is evident that the model's answer accuracy decreases significantly. Even without obtaining the gradient of VideoChat (Li et al., 2023b), the watermarking is successful, and there are instances of garbled text. The visualization results is shown in Fig. 4.

4.4. Ablation Studies

Loss of Different Modalities. In Table 4, the video + LLM watermarking outperforms the individual video and LLM watermarking across all metrics, demonstrating the superior performance of the combined approach. This can be attributed to the complementary nature of video and LLM features, which, when targeted simultaneously, leads to a more potent Video Watermarking that effectively disrupts

the model's output, resulting in lower scores.

4.5. Limitation

Our Video Watermarking is primarily concentrated on the digital world, operating under the assumption that input videos are fed directly into the models. However, as technology advances, we anticipate that video-based LLMs will be increasingly deployed in more complex, real-world scenarios. These scenarios could include autonomous driving, where input videos are not pre-recorded but rather captured in real-time from physical environments via cameras. Future research should explore the execution and impact of watermarking in the physical world. This would provide a more comprehensive evaluation of the security of video-based LLMs, contributing to the development of more robust and reliable systems for real-world deployment.

5. Conclusion

In this paper, we present Flow-based Video Watermarking, pioneering a new frontier in safeguarding video content against unauthorized exploitation by video-based LLMs. Our experiments demonstrate that with minimal watermarking on less than 20% of video frames, we can significantly protect video data from misinterpretation and misuse, highlighting the efficacy of our method. Our work also sheds light on the broader implications for multi-modal model security. The introduction of our watermarking technique is a testament to the proactive steps necessary to ensure the ethical and secure application of AI technologies in handling sensitive video content.

Impact Statement

The social impact of our work is significant as it aims to increase public awareness about the concerns associated with the availability of video-based LLMs. The misuse and misinterpretation of video data can lead to serious privacy leakage and misinformation. By highlighting the responsible use and handling of such video information, this work could potentially lead to the development of more secure systems and policies, thereby ensuring the privacy and safety of individuals and communities.

References

Aafaq, N., Akhtar, N., Liu, W., Gilani, S. Z., and Mian, A. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12487–12496, 2019.

Bagdasaryan, E., Hsieh, T.-Y., Nassi, B., and Shmatikov, V. (ab) using images and sounds for indirect instruction injection in multi-modal llms. *arXiv preprint arXiv:2307.10490*, 2023.

Bai, J., Chen, B., Gao, K., Wang, X., and Xia, S.-T. Practical protection against video data leakage via universal adversarial head. *Pattern Recognition*, 131:108834, 2022.

Bai, J., Gao, K., Min, S., Xia, S.-T., Li, Z., and Liu, W. Badclip: Trigger-aware prompt learning for backdoor attacks on clip. In *CVPR*, 2024a.

Bai, Y., Zeng, Y., Jiang, Y., Xia, S.-T., Ma, X., and Wang, Y. Improving adversarial robustness via channel-wise activation suppressing. In *International Conference on Learning Representations*.

Bai, Y., Feng, Y., Wang, Y., Dai, T., Xia, S.-T., and Jiang, Y. Hilbert-based generative defense for adversarial examples. In *Proceedings of the IEEE/CVF International conference on computer vision*, pp. 4784–4793, 2019.

Bai, Y., Zeng, Y., Jiang, Y., Wang, Y., Xia, S.-T., and Guo, W. Improving query efficiency of black-box adversarial attack. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pp. 101–116. Springer, 2020.

Bai, Y., Pei, G., Gu, J., Yang, Y., and Ma, X. Special characters attack: Toward scalable training data extraction from large language models. *arXiv preprint arXiv:2405.05990*, 2024b.

Caba Heilbron, F., Escorcia, V., Ghanem, B., and Carlos Niebles, J. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 961–970, 2015.

Chen, T.-S., Siarohin, A., Menapace, W., Deyneka, E., Chao, H.-w., Jeon, B. E., Fang, Y., Lee, H.-Y., Ren, J., Yang, M.-H., et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. *arXiv preprint arXiv:2402.19479*, 2024.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Deng, C., Duan, Y., Jin, X., Chang, H., Tian, Y., Liu, H., Zou, H. P., Jin, Y., Xiao, Y., Wang, Y., et al. Deconstructing the ethics of large language models from long-standing issues to new-emerging dilemmas. *arXiv preprint arXiv:2406.05392*, 2024.

Dong, J., Li, X., Xu, C., Yang, X., Yang, G., Wang, X., and Wang, M. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4065–4080, 2021.

Fang, H., Kong, J., Yu, W., Chen, B., Li, J., Xia, S., and Xu, K. One perturbation is enough: On generating universal adversarial perturbations against vision-language pre-training models. *arXiv preprint arXiv:2406.05491*, 2024a.

Fang, H., Qiu, Y., Yu, H., Yu, W., Kong, J., Chong, B., Chen, B., Wang, X., and Xia, S.-T. Privacy leakage on dnns: A survey of model inversion attacks and defenses. *arXiv preprint arXiv:2402.04013*, 2024b.

Gabeur, V., Sun, C., Alahari, K., and Schmid, C. Multimodal transformer for video retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pp. 214–229. Springer, 2020.

Gao, K., Bai, Y., Bai, J., Yang, Y., and Xia, S.-T. Adversarial robustness for visual grounding of multimodal large language models. In *ICLR Workshop*, 2024a.

Gao, K., Bai, Y., Gu, J., Xia, S.-T., Torr, P., Li, Z., and Liu, W. Inducing high energy-latency of large vision-language models with verbose images. In *ICLR*, 2024b.

Gao, K., Gu, J., Bai, Y., Xia, S.-T., Torr, P., Liu, W., and Li, Z. Energy-latency manipulation of multi-modal large language models via verbose samples. *arXiv preprint arXiv:2404.16557*, 2024c.

Gong, Y., Ran, D., Liu, J., Wang, C., Cong, T., Wang, A., Duan, S., and Wang, X. Figstep: Jailbreaking large vision-language models via typographic visual prompts. *arXiv preprint arXiv:2311.05608*, 2023.

Guo, H., Li, J., Dai, T., Ouyang, Z., Ren, X., and Xia, S.-T. Mambair: A simple baseline for image restoration with state-space model. *arXiv preprint arXiv:2402.15648*, 2024.

Hu, W., Zhao, H., Jiang, L., Jia, J., and Wong, T.-T. Bidirectional projection network for cross dimension scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14373–14382, 2021.

Hui, T.-W., Tang, X., and Loy, C. C. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8981–8989, 2018.

Li, J., Dai, T., Zhu, M., Chen, B., Wang, Z., and Xia, S.-T. Fsr: A general frequency-oriented framework to accelerate image super-resolution networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1343–1350, 2023a.

Li, J., Gao, K., Bai, Y., Zhang, J., Xia, S.-t., and Wang, Y. Fmm-attack: A flow-based multi-modal adversarial attack on video-based llms. *arXiv preprint arXiv:2403.13507*, 2024a.

Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., and Qiao, Y. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023b.

Li, R., Li, Q., Zhang, Y., Zhao, D., Jiang, Y., and Yang, Y. Interpreting unsupervised anomaly detection in security via rule extraction. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *NeurIPS*, 2023c.

Li, R., Li, Q., Zhang, Y., Zhao, D., Xiao, X., and Jiang, Y. Genos: General in-network unsupervised intrusion detection by rule extraction. *CoRR*, abs/2403.19248, 2024b.

Liang, S., Zhu, M., Liu, A., Wu, B., Cao, X., and Chang, E.-C. Badclip: Dual-embedding guided backdoor attack on multimodal contrastive learning. In *CVPR*, 2024.

Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023.

Liu, W., Li, Z., and Tang, X. Spatio-temporal embedding for statistical face recognition from video. In *ECCV*, 2006.

Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., and Li, T. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022.

Ma, Y., Wang, Y., Wu, Y., Lyu, Z., Chen, S., Li, X., and Qiao, Y. Visual knowledge graph for human action reasoning in videos. In *ACM MM*, 2022a.

Ma, Y., Yang, T., Shan, Y., and Li, X. Simvtp: Simple video text pre-training with masked autoencoders. *arXiv preprint arXiv:2212.03490*, 2022b.

Ma, Y., Cun, X., He, Y., Qi, C., Wang, X., Shan, Y., Li, X., and Chen, Q. Magicstick: Controllable video editing via control handle transformations. *arXiv preprint arXiv:2312.03047*, 2023.

Ma, Y., He, Y., Cun, X., Wang, X., Chen, S., Li, X., and Chen, Q. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *AAAI*, 2024.

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

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Qi, X., Huang, K., Panda, A., Wang, M., and Mittal, P. Visual adversarial examples jailbreak large language models. *arXiv preprint arXiv:2306.13213*, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021a.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021b.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Seo, P. H., Nagrani, A., Arnab, A., and Schmid, C. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17959–17968, 2022.

Tang, X. and Li, Z. Video based face recognition using multiple classifiers. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, pp. 345–349. IEEE, 2004.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., and Saenko, K. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pp. 4534–4542, 2015.

Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36, 2024.

Wang, X., Zhang, Z., Tong, K., Gong, D., He, K., Li, Z., and Liu, W. Triangle attack: A query-efficient decision-based adversarial attack. In *ECCV*, 2022.

Wei, X., Zhu, J., Yuan, S., and Su, H. Sparse adversarial perturbations for videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 8973–8980, 2019.

Wu, B., Wei, S., Zhu, M., Zheng, M., Zhu, Z., Zhang, M., Chen, H., Yuan, D., Liu, L., and Liu, Q. Defenses in adversarial machine learning: A survey. *arXiv preprint arXiv:2312.08890*, 2023.

Wu, Y.-H., Liu, Y., Zhan, X., and Cheng, M.-M. P2t: Pyramid pooling transformer for scene understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Xiao, J., Xu, Z., Zou, Q., Li, Q., Zhao, D., Fang, D., Li, R., Tang, W., Li, K., Zuo, X., Hu, P., Jiang, Y., Weng, Z., and Lyv.R, M. Make your home safe: Time-aware unsupervised user behavior anomaly detection in smart homes via loss-guided mask. *arXiv preprint arXiv:2406.10928*, 2024.

Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., and Zhuang, Y. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1645–1653, 2017.

Xue, H., Hang, T., Zeng, Y., Sun, Y., Liu, B., Yang, H., Fu, J., and Guo, B. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5036–5045, 2022.

Yang, D., Bai, Y., Jia, X., Liu, Y., Cao, X., and Yu, W. Cheating suffix: Targeted attack to text-to-image diffusion models with multi-modal priors. *arXiv preprint arXiv:2402.01369*, 2024.

Zhang, H., Li, X., and Bing, L. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

Zhang, Z., Qi, Z., Yuan, C., Shan, Y., Li, B., Deng, Y., and Hu, W. Open-book video captioning with retrieve-copy-generate network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9837–9846, 2021.

Zhao, Y., Pang, T., Du, C., Yang, X., Li, C., Cheung, N.-M. M., and Lin, M. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Zhuang, H., Zhang, Y., and Liu, S. A pilot study of query-free adversarial attack against stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2384–2391, 2023.

Zong, Y., Bohdal, O., Yu, T., Yang, Y., and Hospedales, T. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. *arXiv preprint arXiv:2402.02207*, 2024.

Zou, Q., Li, Q., Li, R., Huang, Y., Tyson, G., Xiao, J., and Jiang, Y. Iotbeholder: A privacy snooping attack on user habitual behaviors from smart home wi-fi traffic. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(1):1–26, 2023.

A. Algorithm details

Our overall watermarking procedure is described in Algorithm 1.

Algorithm 1 Video Watermarking on Video-based LLMs Using PGD Optimization

---

Require: Clean video $\mathbf{X}$, user input text Q_{text}, sparsity M_{spa}, video feature extractor $f_\phi(\cdot)$, LLM $g_\psi(\cdot)$, step size α, iterations T

Ensure: Adversarial video $\hat{\mathbf{X}}$

1: Compute video optical flow and obtain flow-based temporal mask $\mathbf{M}_f$ with M_{spa}
2: Initialize perturbation $\Delta \leftarrow 0$
3: **while** $t < T$ **do**
4:     Calculate video features loss $\ell_{video}(Q_{video}, \hat{Q}_{video})$ using Eq. 4
5:     Calculate LLM features loss $\ell_{LLM}(A_{hidden}, \hat{A}_{hidden})$ using Eq. 5
6:     Update perturbation Δ using Eq. 6 with step size α
7: **end while**
8: Compute adversarial video $\hat{\mathbf{X}} \leftarrow \mathbf{X} + \mathbf{M}_f \cdot \Delta$
9: **Return:** Adversarial video $\hat{\mathbf{X}}$

B. Implementation Details

Models. We assess open-source and state-of-the-art video-based LLMs such as Video-ChatGPT (Maaz et al., 2023) and VideoChat (Li et al., 2023b), ensuring reproducibility of our results. Video-ChatGPT is a multi-modal model that seamlessly combines a video-adapted visual encoder (CLIP (Radford et al., 2021a)) with a LLM, which is proficient in comprehending and generating intricate conversations related to videos. VideoChat integrates video foundation models and large language models via a learnable neural interface.
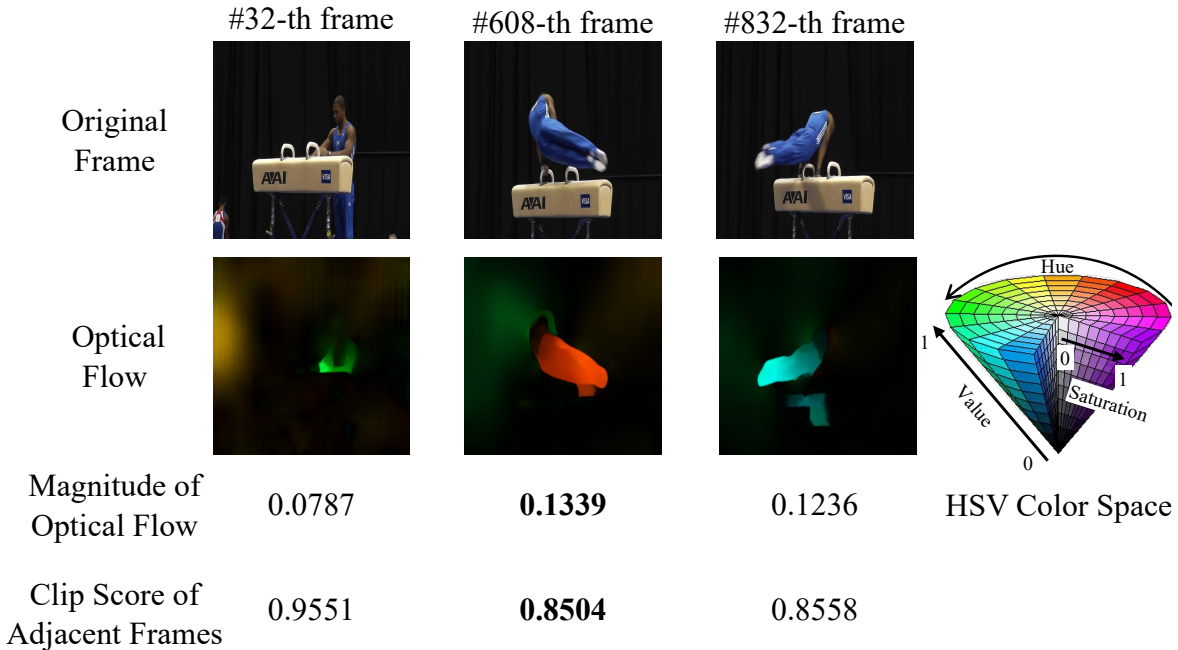

	#32-th frame	#608-th frame	#832-th frame
Magnitude of Optical Flow	0.0787	**0.1339**	0.1236	HSV Color Space
Clip Score of Adjacent Frames	0.9551	**0.8504**	0.8558

Figure 5. Relationship between optical flow and key frames. 'Clip Score of Adjacent Frames' describes the similarity between the current frame and its adjacent frames, the smaller this score is the more different the current frame is. **The frames selected by flow-based masks in our Video Watermarking are key frames in the video.**

Flow-based Temporal Mask. In addition to the statistical analyses presented in the main manuscript, we visualize flow-based methods to enhance interpretability. As depicted in Fig. 5, the brightness represents the magnitude of the optical flow,

and the color indicates the motion direction. The motion flow magnitude varies across different frames, with a larger optical flow signifying more significant motion. Our FMM method tends to select the video frames with the largest optical flow as the key frame of the video. In other words, we tend to prioritize frames with substantial motion changes. Furthermore, the frames we select exhibit low similarity with their neighboring frames, indicating their importance.

Algorithm 2 delineates the process of generating the selected set U from the total set S. The optical flow is computed using a pre-trained liteflownet (Hui et al., 2018). For the random temporal mask, we generate the selected set U by randomly selecting K elements from the total set S. Here, S represents the set of frame indices, $S = \{1, 2, \ldots, T\}$, where T denotes the total number of frames in the video. For the sequence temporal mask, we construct the selected set U by sequentially selecting K elements from the total set S. In this scenario, U comprises a sequence of frames from total frames S, such as $\{1, 2, \ldots, K\}$ or $\{T - K + 1, T - K + 2, \ldots, T\}$, depending on the selected starting point within S.

Algorithm 2 Select Top K Frames with Maximum Flow

Require: video_frames $\in T \times C \times H \times W$, $S = \{1, 2, \ldots, T\}$
Ensure: U, a subset with K elements within S
 1: **for** each frame in video_frames **do**
 2:     Compute optical flow between adjacent frames
 3: **end for**
 4: **for** each optical flow **do**
 5:     Convert optical flow to color and magnitude components
 6:     Normalize the magnitude component
 7: **end for**
 8: Sort the frames based on the average value of the magnitude component
 9: Select the top K frame indices with the highest flow values as the set U

Datasets. In line with the Video-ChatGPT (Maaz et al., 2023) methodology, we curate a test set based on the ActivityNet-200 (Caba Heilbron et al., 2015) and MSVD-QA (Xu et al., 2017) datasets, featuring videos with rich, detailed descriptive captions and associated question-answer pairs obtained from human annotations. Utilizing this test set to generate adversarial examples, we effectively and quantitatively assess the adversarial robustness of video-based LLMs.

Experimental setups. For evaluation, we design three spatial baselines, including videos with random perturbations, black videos with all pixel values set to 0, and white videos with all pixel values set to 1. In addition, we compare our proposed flow-based temporal mask with two straightforward temporal mask methods, serving as temporal mask baselines: the sequence temporal mask and the random temporal mask. Specifically, the sequence temporal mask consists of a continuous sequence of frame indices, while the random temporal mask comprises a randomly chosen sequence of frame indices. We utilize a variety of evaluation metrics to assess the robustness of the models. CLIP (Radford et al., 2021a) score characterizes the semantic similarity between the adversarial answer and the ground-truth answer. A lower CLIP score signifies a lower semantic correlation between the adversarial answer and the ground-truth answer, indicating a more effective attack. Various Image Captioning metrics such as BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015) are used to evaluate the quality of the adversarial answer generated by the model. A lower score corresponds to a more effective attack. BLEU measures the overlap of n-grams between the generated and reference captions. ROUGE-L computes the longest common subsequence between them, reflecting their sentence-level similarity. CIDEr emphasizes the importance of semantically meaningful words in the captions. Following Video-ChatGPT (Maaz et al., 2023) and Video-LLaMA (Zhang et al., 2023). We also employ an evaluation pipeline using the GPT-3.5 and GPT-4 models. The pipeline employs GPT to assign a score from 1 to 5, evaluating the similarity between the output sentence and the ground truth, and a binary score (0 or 1) to measure its accuracy.

C. Additional Experimental Results

Garbling Effect. Intriguingly, our proposed Video Watermarking induces garbling in the model output, while the other three watermarking methods do not cause such distortion. This suggests that the Video Watermarking not only diminishes the model's cue information but also prompts the model to hallucinate. Furthermore, as shown in Fig. 6, we analyse the number of successfully attacked Video-ChatGPT in ActivityNet-200 (Caba Heilbron et al., 2015), and find that video loss is more effective in inducing garbled contents.
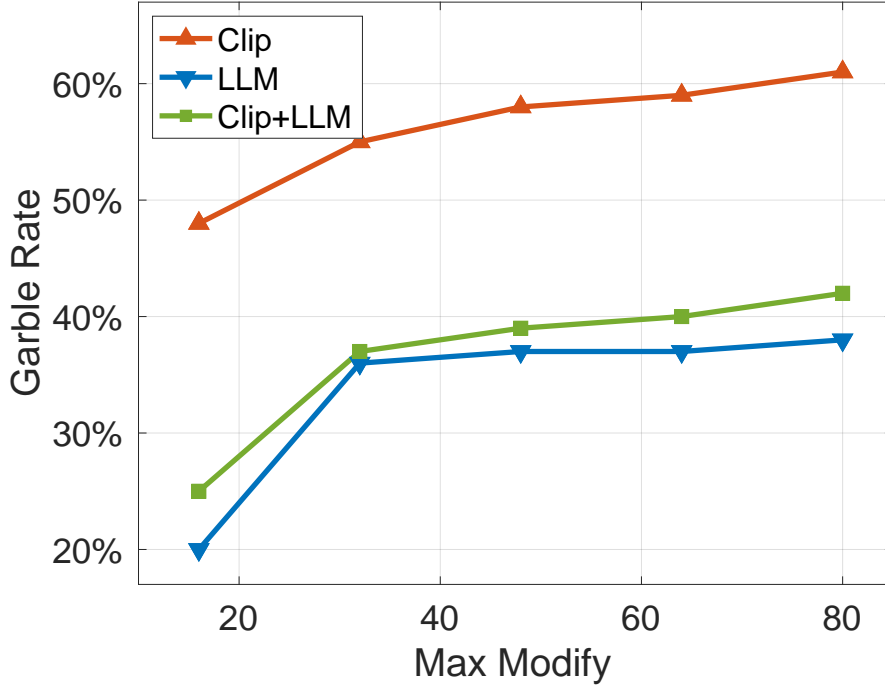
Figure 6. Comparison of different types of watermarking on the garbling rate. Max Modify denotes the maximum pixel value that can be modified, while the Garble Rate represents the percentage of responses that are garbled.

Table 3. Video watermarking against Video-ChatGPT on the ActivityNet-200 dataset. seq: sequence temporal mask. random: random temporal mask. Ours: our flow-based temporal mask. M_{spa}: Sparsity of temporal mask.

Metrics	$M_{spa} = 20\%$			$M_{spa} = 40\%$			$M_{spa} = 60\%$			$M_{spa} = 80\%$		
	seq	random	Ours	seq	random	Ours	seq	random	Ours	seq	random	Ours
RN50	0.7500	0.7142	0.6821	0.7578	0.7559	0.7460	0.7583	0.7671	0.7510	0.7715	0.7813	0.7349
RN101	0.7568	0.7251	0.7085	0.7500	0.7588	0.7559	0.7764	0.7769	0.7500	0.7788	0.7827	0.7637
BLEU	0.1572	0.1590	0.1527	0.1606	0.1751	0.1485	0.1900	0.1894	0.1830	0.2000	0.1957	0.1940
ROUGE-L	0.4073	0.3996	0.4109	0.4323	0.4458	0.4053	0.4767	0.4727	0.4713	0.4779	0.4658	0.4731
GPT3.5	2.02	2.01	2.00	2.45	2.24	2.22	2.64	2.63	2.62	2.76	2.78	2.75
GPT4	1.09	1.08	1.07	1.35	1.34	1.33	1.63	1.63	1.61	1.99	1.58	1.54

Perturbation Budget Δ. We compared the effects of different Δ. It's important to note that the mean of Δ, constrained by the sparsity loss, remains consistent across various levels of maximum modification Δ_{max}. As in Fig. 7 (a), $\Delta = 32$ performs best. On the one hand, the larger Δ is, the greater the potential modification of individual pixels, but on the other hand, due to the sparse loss, the amount of pixels that can be modified with a larger Δ becomes smaller. Therefore, $\Delta = 32$ seems to be a better trade-off. **Percentage of Selected Frames.** In some instances, while we launch watermarking on all video frames, video-based LLMs only randomly sample a subset of the video frames. As demonstrated in Fig. 7 (b), the potency of the watermarking escalates with an increasing number of sampled video frames. Interestingly, even when a minor fraction (40%, 20%) of video frames are sampled, we observe a substantial decline in the Clip score relative to the baseline. Surprisingly, a 20% sampling rate seems to yield superior results than a 40% rate. We speculate this unexpected outcome could be due to inherent fluctuations when a limited number of video frames are sampled, coupled with our stream-based approach that assures a certain minimum watermarking effectiveness. To make this engagement more compelling, we intend to delve deeper into this phenomenon with additional experiments in future studies.

Table 3 compares different mask ratios (sparsity) and temporal mask approaches. Our Video Watermarking outperforms the other two approaches across various sparsity levels, indicating the effectiveness of our proposed flow-based temporal mask, which leverages the concept of maximum flow prioritization. It demonstrates its potency in the realm of video-based LLM and its ability to maximize the extraction of video information. In the bottom of Table 3, Video Watermarking consistently outperforms the other two approaches across various sparsity levels, resulting in a more significant reduction of GPT scores and accuracies.
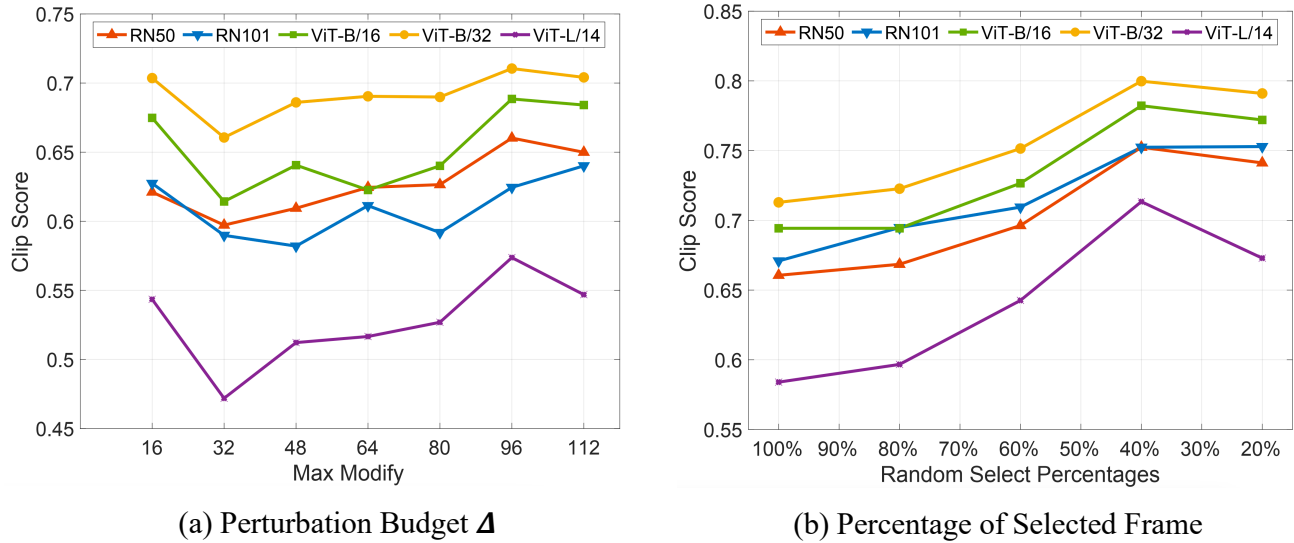
(a) Perturbation Budget Δ (b) Percentage of Selected Frame

Figure 7. Ablation Studies of different watermarking settings. (a) Comparison of different max modify pixels. (b) Comparison of different random select percentages.

Table 4. Comparison of different watermarking types on Clip Score and Image Captioning Metrics. video: represents watermarking targeting video features, LLM: represents watermarking targeting LLM features, and video + LLM: represents combined watermarking on both video and LLM features. Lower scores indicate better watermarking performance.

Type	Clip Score ↓					Image Captioning ↓		
	RN50	RN101	ViT-B/16	ViT-B/32	ViT-L/14	BLEU	ROUGE-L	CIDEr
Clean	0.7817	0.7827	0.8096	0.8115	0.7231	0.2029	0.4820	1.6364
video	0.7403	0.7524	0.7690	0.7744	0.6811	0.1975	0.4345	1.5862
LLM	0.7153	0.6904	0.7334	0.7515	0.6387	0.1042	0.3226	0.8350
video + LLM	**0.6491**	**0.6060**	**0.6836**	**0.6968**	**0.5566**	**0.0230**	**0.1585**	**0.1601**