Protecting Your Video Content: Disrupting Automated Video-based LLM Annotations

Supplementary Material

A. Overview

The algorithm details are shown in Appendix B. Detailed implementation information, including model and method setups, can be found in Appendix C. In Appendix D.1, we analyze the length distribution of Mute-S to demonstrate its effectiveness in producing short sentences. Appendix D.2 provides our ablation study on perturbation magnitude as a supplement for Sec. 4.4. We conduct the prompt transferability experiment on non-annotation prompts to show our Mutes robustness in Appendix D.3. Different inference parameters' influence is described in Appendix D.4. Text perturbation's influence for fine-tuning text-to-video models when videos are clean is shown in Appendix D.5. Then, we conduct a black-box attack experiment in Appendix D.6. In Appendix D.7, we provide explanations of the different methods. We then perform adaptive evaluations in Appendix D.8 to further demonstrate the robustness of our methods. Additionally, we optimize the prompt transferability of Mute-N in Appendix D.9 and evaluate the quality of adversarial videos in Appendix D.10. Finally, we present demos of different methods in Appendix D.11. Additionally, for ease of processing video annotations, line breaks in the textual outputs are replaced with spaces in this paper. It is also worth noting that the perturbation shown in Fig. 1 is illustrative and not the real perturbation, as the real perturbation contains negative values. The answers in Fig. 1 and Fig. 2 are also illustrative not the real answers. And the original video and clean video in this paper all mean the video without added perturbation.

B. Algorithm Details

Our algorithms are detailed in Algorithms 1 to 4. Here, $clip(\cdot)$ restricts the adversarial perturbation, while $G(\cdot)$ denotes the generation function of video-based LLMs.

We use auto-regressive loss in Eq. (5) for Mute-N. In video-based LLMs, auto-regressive loss typically relies on the cross-entropy loss function, which is composed of the log_softmax and nll_loss functions. When only the EOS token is used as the target, the effect of increasing the first generated token's EOS probability via the log_softmax function is equivalent to the effect of Eq. (5). For simplicity in program implementation, when programming, we directly manipulate the EOS probability of the first generated token using the log_softmax function instead of computing the auto-regressive loss by passing label into the forward function.

Algorithm 1 Rambling-F

Input: original video $\boldsymbol{x}$, perturbation magnitude ϵ, optimization step size s, input prompt c_{in}
Output: adversarial video $\boldsymbol{x'}$

Initialize perturbation: Δ
while $i < T$ **do**
  $\boldsymbol{x'} \leftarrow \boldsymbol{x} + \Delta$
  $\mathcal{L}_{video}(\boldsymbol{x'}), \mathcal{L}_{LLM}(\boldsymbol{x'}) \leftarrow Eq.(2)$
  $\mathcal{L} \leftarrow \alpha \times \mathcal{L}_{video}(\boldsymbol{x'}) + \beta \times \mathcal{L}_{LLM}(\boldsymbol{x'})$
  $\Delta \leftarrow \Delta - s \times sign(\nabla \mathcal{L})$
  $clip(\boldsymbol{x}, \Delta, \epsilon)$
  $i \leftarrow i + 1$
end
return $\boldsymbol{x'}$

Algorithm 2 Rambling-L

Input: original video $\boldsymbol{x}$, perturbation magnitude ϵ, optimization step size s, input prompt c_{in}
Output: adversarial video $\boldsymbol{x'}$

Initialize perturbation: Δ, clean caption: $y \leftarrow G(\boldsymbol{x}, c_{\text{in}})$
while $i < T$ **do**
  $\boldsymbol{x'} \leftarrow \boldsymbol{x} + \Delta$
  $\mathcal{L} \leftarrow Eq.(3)$
  $\Delta \leftarrow \Delta - s \times sign(\nabla \mathcal{L})$
  $clip(\boldsymbol{x}, \Delta, \epsilon)$
  $i \leftarrow i + 1$
end
return $\boldsymbol{x'}$

Algorithm 3 Mute-S

Input: original video $\boldsymbol{x}$, perturbation magnitude ϵ, optimization step size s, input prompt c_{in}
Output: adversarial video $\boldsymbol{x'}$

Initialize perturbation: Δ
while $i < T$ **do**
  $\boldsymbol{x'} \leftarrow \boldsymbol{x} + \Delta$
  $y_{\text{out}} \leftarrow G(\boldsymbol{x'}, c_{\text{in}})$
  $\mathcal{L} \leftarrow Eq.(4)$
  $\Delta \leftarrow \Delta - s \times sign(\nabla \mathcal{L})$
  $clip(\boldsymbol{x}, \Delta, \epsilon)$
  $i \leftarrow i + 1$
end
return $\boldsymbol{x'}$

Algorithm 4 Mute-N

Input: original video $\boldsymbol{x}$, perturbation magnitude ϵ, optimization step size s, input prompt c_{in}

Output: adversarial video $\boldsymbol{x'}$

Initialize perturbation: Δ

while $i < T$ **do**

  $\boldsymbol{x'} \leftarrow \boldsymbol{x} + \Delta$

  $\mathcal{L} \leftarrow Eq.(5)$

  $\Delta \leftarrow \Delta - s \times sign(\nabla \mathcal{L})$

  $clip(\boldsymbol{x}, \Delta, \epsilon)$

  $i \leftarrow i + 1$

end

return $\boldsymbol{x'}$

C. Implementation Details

Our adversarial experiments are run on NVIDIA GeForce RTX 3090, 24GB. We respectively use 100 videos from OpenVid-1M [6], MSR-VTT [9], and WebVid-10M [1]. We preprocess the videos into 8 frames with 5 fps. And the video pictures in this paper are from these datasets. For AnimateDiff, we use 200 videos preprocessed into 8 frames with 5fps from OpenVid-1M.

C.1. Model Setups

Video-ChatGPT. We follow the offline demo settings from [5] to configure Video-ChatGPT.

Video-LLaMA. We use Video-LLaMA-2-7B-Pretrained [1] to configure Video-LLaMA [10]. We follow the demo for video without audio to configure.

Video-Vicuna. We change the backbone of Video-LLaMA from Llama-2-7b-chat [2] into Vicuna-7B [3] and make it into Video-Vicuna. On the other hand, the vision-language branch we used is finetune-vicuna7b-v2 [4]. We follow the demo for video without audio to configure.

C.2. Method Setups

Rambling-F. Three victim models share the same adversarial setups. The number of optimization iteration is set as 200 with a perturbation magnitude ϵ as 16/255. And the optimization step size is 1/255. Both the α and β in the Eq. (2) are set to 1.

Rambling-L. Three victim models share the same adversarial setups. The number of optimization iteration is set as 200 with a perturbation magnitude ϵ as 16/255. And the optimization step size is 1/255.

Mute-S. The number of optimization iteration is set as 500 with a perturbation magnitude ϵ as 16/255. We set the optimization step size as 1/255. To accelerate the optimization process, we set distinct iteration-breaking thresholds for the three victim models: a threshold of 3 for Video-ChatGPT and 4 for both Video-LLaMA and Video-Vicuna. If the loss falls below the threshold, the adversarial pipeline stops immediately.

Mute-N. The number of optimization iteration is set as 500 with a perturbation magnitude ϵ as 16/255. We set the optimization step size as 1/255. To accelerate the optimization process, we set an iteration-breaking threshold of 0.3. If the loss falls below this threshold, the adversarial pipeline stops immediately.

D. Additional Studies

D.1. Length Distribution of Mute-S

To evaluate the detailed performance of Mute-S, which focuses on the EOS token, we analyze the distribution of text length generated by this method.

Fig. 6, Fig. 7, and Fig. 8 clearly illustrate that, with the addition of watermarks through Mute-S, the length distribution shifts noticeably to the left along the X-axis. This shift indicates a concentration of shorter text length, highlighting the effectiveness of Mute-S in reducing information leakage. Furthermore, we calculate the rate at which an alphabetic character appears in the final position of the output text, finding that 80.33% of annotations generated by Video-LLaMA end in this manner. This suggests that sentences frequently lack a punctuation mark at the end, rendering them incomplete. Thus, Mute-S effectively disrupts text continuity, producing shorter and truncated outputs. In this section, the datasets we used are OpenVid-1M, MSR-VTT, and WebVid-10M.

D.2. Ablation Study on Perturbation Magnitude

This section is a supplement for Sec. 4.4. As shown in Fig. 9, Fig. 10, and Tab. 6, a lower perturbation magnitude tends to result in a higher CLIP score [8] and BLEU [7], indicating improved annotation performance and reduced protective effectiveness for the videos. Therefore, for Ramblings, it is crucial to select an appropriate perturbation magnitude to ensure its performance. Tab. 7 indicates that when the perturbation magnitude is too small, Mutes may become ineffective. We speculate that this is because Mutes rely on the EOS token's probability, and a slight increase may not be sufficient to prompt video-based LLMs to generate the EOS token. Therefore, it is important to avoid selecting a too-small perturbation magnitude for Mutes.

D.3. Prompt Transferability of Mutes on Non-annotation Prompts

In the Sec. 4.4, we focus on prompts designed to annotate videos, such as "What is this video about?". Here, we ex-

[1] https://huggingface.co/DAMO-NLP-SG/Video-LLaMA-2-7B-Pretrained

[2] https://huggingface.co/meta-llama

[3] https://huggingface.co/lmsys/vicuna-7b-delta-v0

[4] https://huggingface.co/DAMO-NLP-SG/Video-LLaMA-Series

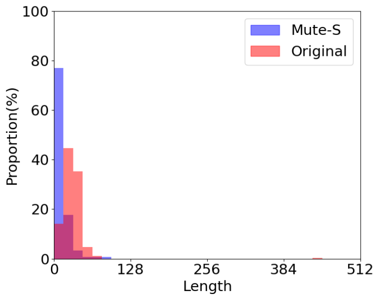


Figure 6. Length distribution of Mute-S on Video-ChatGPT.

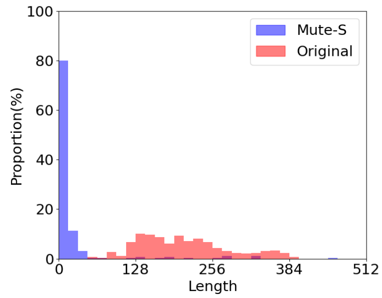


Figure 7. Length distribution of Mute-S on Video-LLaMA.

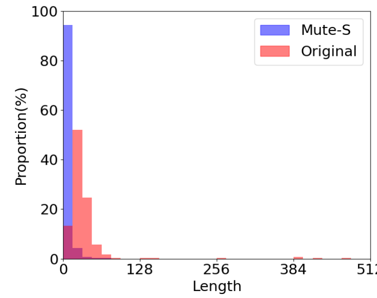


Figure 8. Length distribution of Mute-S on Video-Vicuna.

Table 6. The CLIP score and BLEU between texts annotated by Video-LLaMA and clean captions are measured on OpenVid-1M dataset to explore the influence of different perturbation magnitudes ϵ. The best results are highlighted in **bold**.

Method	$\epsilon = 2$		$\epsilon = 4$		$\epsilon = 8$		$\epsilon = 16$	
	CLIP Score	BLEU	CLIP Score	BLEU	CLIP Score	BLEU	CLIP Score	BLEU
Original	0.78760	0.05121	0.78760	0.05121	0.78760	0.05121	0.78760	0.05121
Noise	0.79834	0.05041	0.79297	0.04500	0.79297	0.04595	0.79980	0.04761
Rambling-F	**0.67676**	**0.02740**	**0.64062**	**0.02275**	**0.60596**	**0.02033**	**0.58301**	**0.01600**
Rambling-L	0.70361	0.03289	0.68164	0.02609	0.65039	0.02057	0.60938	0.02068

Table 7. The length and EOS rate of texts annotated by Video-LLaMA are measured on OpenVid-1M dataset to explore the influence of different perturbation magnitudes ϵ. The best results are highlighted in **bold**.

Method	$\epsilon = 2$		$\epsilon = 4$		$\epsilon = 8$		$\epsilon = 16$	
	Length	EOS Rate(%)	Length	EOS Rate(%)	Length	EOS Rate(%)	Length	EOS Rate(%)
Original	203.54	0.0	203.54	0.0	203.54	0.0	203.54	0.0
Noise	209.81	0.0	222.06	0.0	229.47	0.0	224.23	0.0
Mute-S	205.26	0.0	205.73	0.0	131.32	5.0	11.58	7.0
Mute-N	217.00	0.0	**192.73**	**10.0**	**48.61**	**79.0**	**0.00**	**100.0**

Table 8. Prompt transferability on non-annotation prompts: The length and EOS rate of texts generated by Video-LLaMA are measured on OpenVid-1M dataset. The prompt used during the attack phase is "What is this video about?" After the attack, we apply three different prompts—"Do you like this video?", "Who are you?", and "Where are you from?" to generate text outputs for the videos. The best results are highlighted in **bold**.

Method	Do you like this video?		Who are you?		Where are you from?	
	Length	EOS Rate(%)	Length	EOS Rate(%)	Length	EOS Rate(%)
Original	245.33	0.0	158.58	0.0	153.59	0.0
Noise	253.17	0.0	182.66	0.0	165.31	0.0
Mute-S	**36.05**	8.0	**22.23**	4.0	**36.50**	4.0
Mute-N	167.02	**21.0**	144.66	**17.0**	146.45	**22.0**

Table 9. Ablation study on different inference parameters: The CLIP score and BLEU between texts annotated by Video-LLaMA and clean captions are measured on OpenVid-1M dataset. The parameters that are manually set are displayed below. DS means "do_sample", NB means "num_beams", T means temperature, and TP means "top_p". The best results are highlighted in **bold**.

Method	DS=False		DS=True,NB=2,T=1.0		DS=True,TP=0.8,T=0.8	
	CLIP Score	BLEU	CLIP Score	BLEU	CLIP Score	BLEU
Original	0.79248	0.05026	0.79346	0.04929	0.79443	0.05090
Noise	0.79395	0.04794	0.79834	0.04818	0.79395	0.04959
Rambling-F	**0.58740**	**0.01684**	**0.58105**	**0.01663**	**0.57764**	**0.01595**
Rambling-L	0.61328	0.02025	0.61719	0.02130	0.61621	0.02039

Table 10. Ablation study on different inference parameters: The length and EOS rate of texts annotated by Video-LLaMA are measured on OpenVid-1M dataset. The parameters that are manually set are displayed below. DS means "do_sample", NB means "num_beams", T means temperature, and TP means "top_p". The best results are highlighted in **bold**.

Method	DS=False		DS=True,NB=2,T=1.0		DS=True,TP=0.8,T=0.8	
	Length	EOS Rate(%)	Length	EOS Rate(%)	Length	EOS Rate(%)
Original	221.61	0.0	213.19	0.0	197.12	0.0
Noise	217.03	0.0	213.70	0.0	196.79	0.0
Mute-S	14.76	8.0	27.63	2.0	21.82	8.0
Mute-N	**0.00**	**100.0**	**0.00**	**100.0**	**2.10**	**99.0**

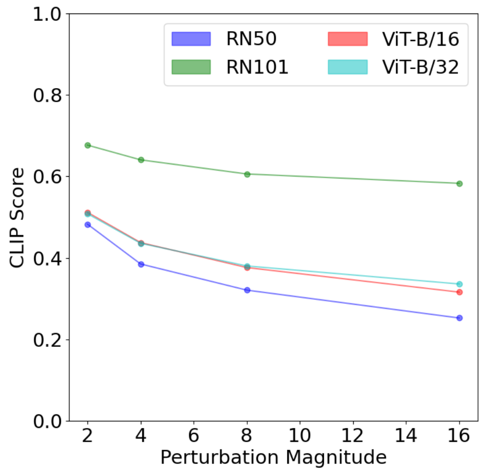


Figure 9. CLIP score about Rambling-F on different perturbation magnitudes.

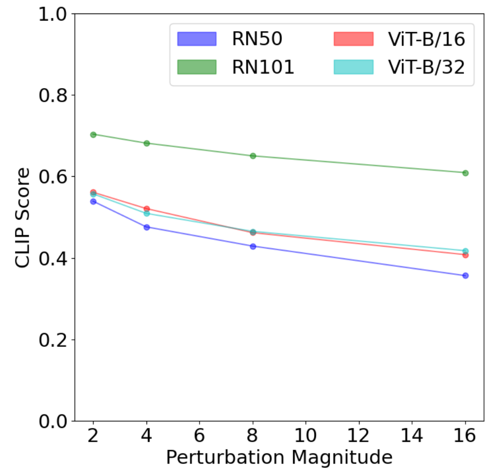


Figure 10. CLIP score about Rambling-L on different perturbation magnitudes.

tend this by examining the transferability of non-annotation prompts. During the attack phase, we maintain the same prompt-"What is this video about?", as used in Sec. 4.4. And here, "Do you like this video?", "Who are you?", and "Where are you from?" are used to generate outputs. Since the datasets do not contain answers for these prompts, we focus on evaluating the performance of the Mutes method.

As shown in Tab. 8, the performance of Mutes is evident in the decreased output length and higher EOS rate compared to the baselines. Notably, Mute-S excels in generating short sentences, as reflected in its great performance in output length, while Mute-N achieves a higher EOS rate, indicating its effectiveness in producing more null captions. Overall, this analysis highlights the prompt transferability of our Mutes method, even when applied to non-annotation prompts above.

D.4. Ablation Study on Different Inference Parameters

In this section, we evaluate the robustness of our watermarking methods by conducting inferences using video-based LLMs with varying inference parameters. The maximum number of new tokens is set to 512. The parameters are displayed in Tab. 9 and Tab. 10. As clearly shown, Ramblings and Mutes perform well under different inference parameters. Specifically, the CLIP score and BLEU decrease dramatically with Ramblings, while the EOS rate is high and the length is short with Mutes.

D.5. Ablation Study on Text Perturbation

In this section, we fix the clean videos in the video-text pairs used to fine-tune text-to-video models and vary the texts to evaluate the impact of text perturbation. The results shown in Tab. 5 and Tab. 11 indicate that the watermarks added to the videos play a significant role in reducing the performance of AnimateDiff [2]. This is evident because the VQA_A and VQA_T [3] remain relatively close to the original when only texts are altered and clean videos are fixed.

D.6. Black-box Attack

Additionally, we perform black-box attacks to evaluate the transferability of our watermarking methods across different models. Adversarial videos are generated by Video-LLaMA based on the OpenVid-1M dataset and subsequently annotated by Video-Vicuna. In order to use Mutes to get shorter captions, we load two models simultaneously

Table 11. The VQA_A and VQA_T are evaluated for text-to-video models fine-tuned on different video-text pairs. All the videos are clean videos from OpenVid-1M. We change the texts to evaluate the influence of text perturbation when the videos are clean. Original means we use the texts that are annotated by Video-LLaMA on original videos. Rambling-F-Text-Only means we use the texts that are annotated by Video-LLaMA on videos protected by Rambling-F. The same applies to others. Specifically, the texts of Mute-N-Text-Only are null captions.

Text	VQA_A	VQA_T
Original	60.430	51.302
Incorrect Text: Rambling-F-Text-Only	60.717	45.724
Incorrect Text: Rambling-L-Text-Only	62.218	50.322
Short Text: Mute-S-Text-Only	60.717	45.425
Short Text: Mute-N-Text-Only	62.377	52.008

Table 12. The CLIP score and BLEU of black-box attacks by Ramblings are evaluated. Here, we get the adversarial videos from Video-LLaMA and annotate these videos with Video-Vicuna. The dataset we used is OpenVid-1M. The best results are highlighted in **bold**.

Method	CLIP Score		Video Caption
	RN50	RN101	BLEU
Original	0.63818	0.75732	0.01969
Noise	0.64453	0.76367	0.02275
Rambling-F	**0.43823**	**0.65674**	0.01559
Rambling-L	0.49438	0.66846	**0.01441**

Table 13. The length and EOS rate of black-box attacks by Mutes are evaluated. Here, we get the adversarial videos from Video-LLaMA and annotate these videos with Video-Vicuna. The dataset we used is OpenVid-1M. The best results are highlighted in **bold**.

Method	Length	EOS Rate(%)
Original	32.90	0.0
Noise	38.27	0.0
Mute-S	**13.56**	0.0
Mute-N	14.19	0.0

to annotate the adversarial videos for each optimization iteration, which is set as 200 (threshold is canceled), and select the minimum length as our results. As shown in Tab. 12 and Tab. 13, both our Ramblings and Mutes demonstrate robust performance. The CLIP score, BLEU, and length show varying degrees of decline compared with baselines, indicating that our methods remain effective even without direct access to the victim model - Video-Vicuna.

D.7. Explanation of Methods

The reasons our methods disrupt annotation performance lie in the shifts within the feature space for Rambling-F and perturbations in token distribution for Rambling-L and Mutes. For Rambling-F, annotation performance decreases as the perturbation magnitude increases shown in Tab. 6, accompanied by a growing shift distance in the feature space in Tab. 14. Additionally, the loss designs for Rambling-L and Mutes are specifically motivated by their impact on token distribution. Rambling-L shifts the token distribution away from the clean caption, whereas Mute-S and Mute-N focus on increasing the probability of the EOS token.

Table 14. The feature variation of Rambling-F and the entropy of Rambling-L are evaluated under different perturbation magnitudes. The model and dataset we used are Video-LLaMA and OpenVid-1M respectively.

Metric	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = 16$
Video feature of Rambling-F	0.4606	0.5252	0.6018	0.6629
LLM feature of Rambling-F	0.5442	0.7552	0.9577	1.0961
Entropy of Rambling-L	1.1167	1.8029	2.6964	4.0978

D.8. Adaptive Evaluation

We evaluate the robustness of our methods under format conversion from avi to mkv in Tab. 15. Specifically, we use the OpenCV package to save videos from avi to mkv using the FFV1 codec. The results below show that our methods can work well under format conversion from avi to mkv.

We further conduct adaptive evaluation under compression. Given that compression is widespread in real-world applications, we consider a scenario where the video is resized from 224×224 to 112×112 for storage. To enhance adversarial performance, we incorporate an adaptive attack into our method by integrating downsampling and upsampling into the adversarial pipeline. As shown in Tab. 15, our adaptive attack remains effective under video size compression. Additionally, we evaluate the robustness of our methods under noise removal using a mean filter. Similar to compression, we integrate the mean filter into the adversarial pipeline to enhance performance. The results demonstrate that our adaptive attack is also effective in mitigating the impact of the mean filter.

Table 15. The adaptive evaluation of our methods. The model and dataset we used are Video-LLaMA and OpenVid-1M respectively.

Adaptability	Rambling-F		Rambling-L		Mute-S		Mute-N	
	CLIP	BLEU	CLIP	BLEU	Length	EOS Rate(%)	Length	EOS Rate(%)
Original	0.788	0.051	0.788	0.051	203.54	0.0	203.54	0.0
AVI	0.583	0.016	0.609	0.021	11.58	7.0	0.00	100.0
MKV	0.583	0.016	0.609	0.021	11.58	7.0	0.00	100.0
Compress	0.608	0.016	0.622	0.022	32.43	7.0	0.00	100.0
Removal	0.610	0.018	0.640	0.021	33.50	2.0	0.00	100.0

D.9. Optimization of Prompt Transferability

To enhance the prompt transferability of Mute-N, we incorporate adversarial prompt training in CroPA [4] into the optimization for our attack, dubbed "Mute-N2" by adding adversarial perturbations both on prompt and video. The results in Tab. 16 demonstrate that our Mute-N2 performs well across three prompts. Notably, only the single prompt "What is this video about?" is used for adversarial prompt training in Mute-N2.

Table 16. The length and EOS rate of optimized method Mute-N2. The model and dataset we used are Video-LLaMA and OpenVid-1M respectively.

Method	What is this video about?		What happens in the video?		Can you describe the video in detail?	
	Length	EOS Rate(%)	Length	EOS Rate(%)	Length	EOS Rate(%)
Mute-N2	0.00	100.0	2.67	99.0	86.53	68.0

D.10. Additional Evaluation on Perturbation

We conduct a human evaluation with five participants on ten video pairs. They rate the semantic similarity between clean and adversarial videos on a scale of 1 to 5, where 1 indicates poor semantic similarity and 5 indicates high semantic similarity. Finally, we use LPIPS [11] to measure the perceptual similarity. The results in Tab. 17 show that the adversarial videos are well-constrained.

Table 17. Human and LPIPS evaluation between clean and adversarial videos. The model and dataset we used are Video-LLaMA and OpenVid-1M respectively.

Metric	Rambling-F	Rambling-L	Mute-S	Mute-N
Human	4.10	4.32	4.36	4.22
LPIPS	0.246	0.246	0.243	0.220

D.11. Visualization

The visualization of our watermarking methods is shown in Figs. 11 to 16. Our Ramblings and Mutes display great protective performance.

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In ICCV, 2021. 2

[2] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In ICLR, 2024. 4

[3] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In CVPR, pages 22139–22149, 2024. 4

[4] Haochen Luo, Jindong Gu, Fengyuan Liu, and Philip Torr. An image is worth 1000 lies: Transferability of adversarial images across prompts on vision-language models. In ICLR, 2024. 5

[5] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In ACL, 2024. 2

[6] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. arXiv preprint arXiv:2407.02371, 2024. 2

[7] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In ACL, pages 311–318, 2002. 2

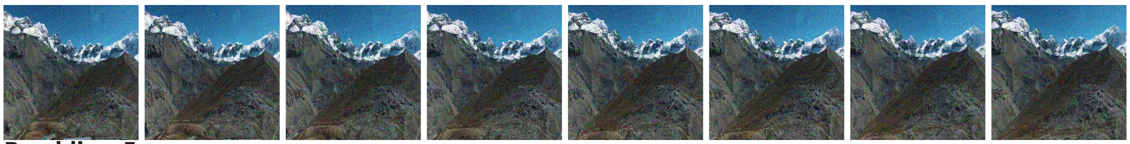
[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In ICML, pages 8748–8763. PMLR, 2021. 2

[9] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In CVPR, 2016. 2

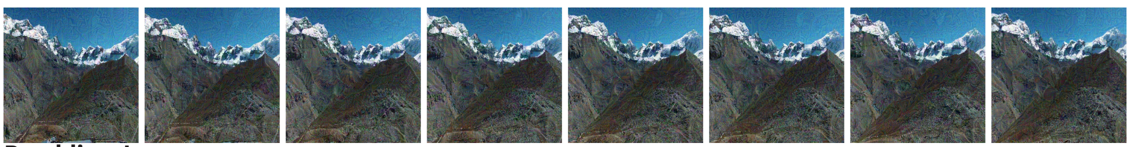
[10] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858, 2023. 2

[11] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In CVPR, 2018. 6

Original
Question: What is this video about?
Output: The video is about a mountain range with snow-capped mountains, and it shows a person walking along a trail in the mountains.

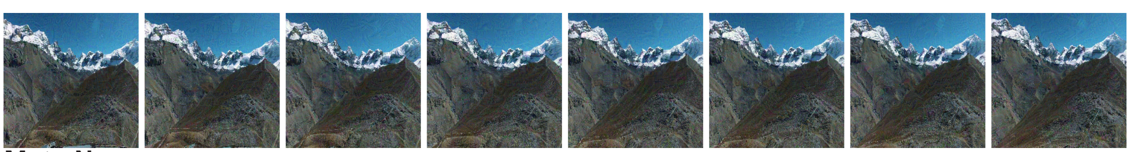


Rambling-F
Question: What is this video about?
Output: The video is about a woman who is sitting in a room and talking to a man who is sitting in a chair. She is wearing a blue shirt and a red jacket.



Rambling-L
Question: What is this video about?
Output: The video is about rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain rain ...

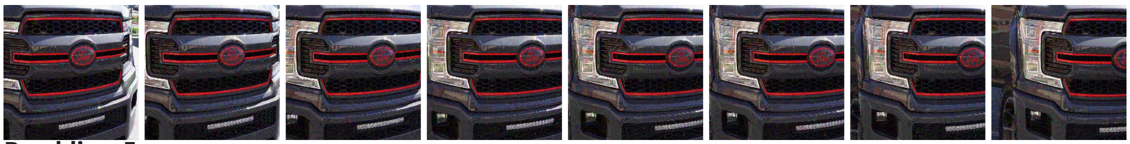
Mute-S
Question: What is this video about?
Output:



Mute-N
Question: What is this video about?
Output:

Figure 11. Visualization of watermarking methods against Video-ChatGPT.

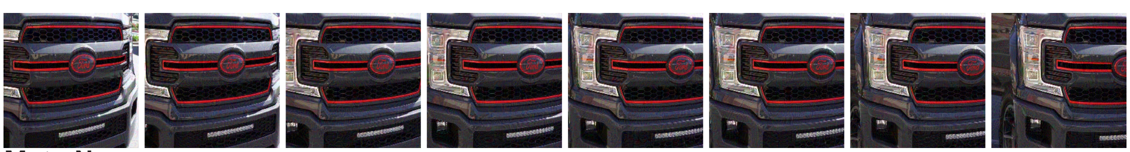
Original
Question: What is this video about?
Output: The video is about a black Ford truck driving down the street, and it features a close-up of the truck's grille.



Rambling-F
Question: What is this video about?
Output: The video is about a man who is talking to a large machine, possibly a robot. He is explaining the features of the machine and how it works.

Rambling-L
Question: What is this video about?
Output: The Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram Ram ...

Mute-S
Question: What is this video about?
Output: The video is about a man who i



Mute-N
Question: What is this video about?
Output:

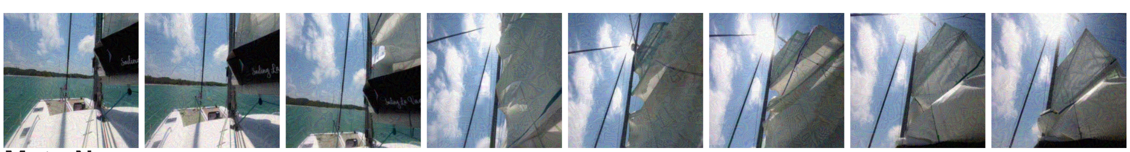
Figure 12. Visualization of watermarking methods against Video-ChatGPT.

Original
Question: What is this video about?
Output: Based on the provided video frames, it appears that the video is showing a sailboat sailing in the ocean with a clear blue sky and white sails. The video frames are: 0.0 - A close-up of the sailboat's bow, with the wind blowing against the sail and causing it to ripple. 0.2 - A view of the sailboat's side, with the sunlight reflecting off the white sails and creating a bright glare. ...

Rambing-F
Question: What is this video about?
Output: Killer] 1. Thank you </video> </document> "> 01.07.07.08.09.10.11.12.13.14.15.16.17.18.19.20.21.22.23.24.25.26.27.28.29.30.31.32.33.34.35.36.37.38.39.40.41.42.43.44.45.46.47.48.49.50.51.52.53.54.55.56.57.58.59.60.61.62.63.64.65.66.67.68.69.70.71.72.73.74.75.76.77.78.79.80.81.82.83.84.85.86.87.88.89.90.91.92.93.94.95.96.97.98.99.100 on a yellow background.

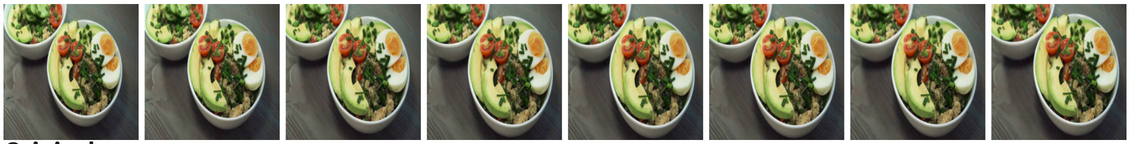
Rambling-L
Question: What is this video about?
Output: The visual content of the map reveals the following information: The first word, "sailing," is shown on a path moving from the center of the visual to the top left corner. The path is surrounded by a blue background, which represents the ocean. The second word, "iowa," is shown on a route moving from the top left corner to the center of the visual. ...

Mute-S
Question: What is this video about?
Output: The video is about a tourist on a boat sailing in the ocean and discovering a

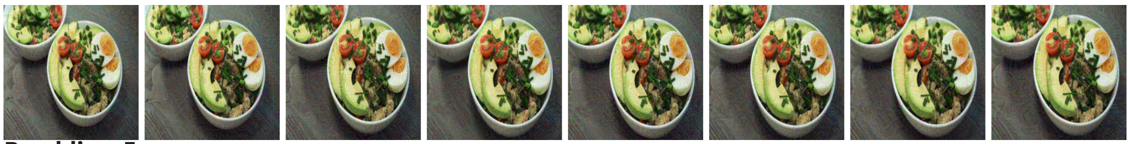


Mute-N
Question: What is this video about?
Output:

Figure 13. Visualization of watermarking methods against Video-LLaMA.



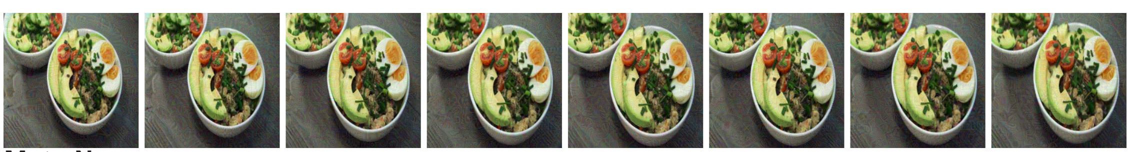
Original
Question: What is this video about?
Output: Based on the provided video frames, it appears that the video is showing a person preparing a healthy breakfast bowl with quinoa, avocado, tomatoes, and eggs. The video starts with a close-up shot of a bowl containing quinoa, and then pans out to show the entire breakfast setup, including a plate with a fried egg on top. ...



Rambling-F
Question: What is this video about?
Output: The season was 0.0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4] The season was 2022.0, 202.2, 202.4, 202.6, 202.8, 203.0, 203.2, 203.4] The season was 2022.0, 202.2, 202.4, 202.6, 202.8, 203.0, 203.2, 203.4] The season was 2022.0, 202.2, 202.4, 202.6, 202.8, 203.0, 203.2, 203.4] The season was 2022.0, 202.2, 202.4, 202.6, 202.8, 203.0, 203.2, 203.4] ...

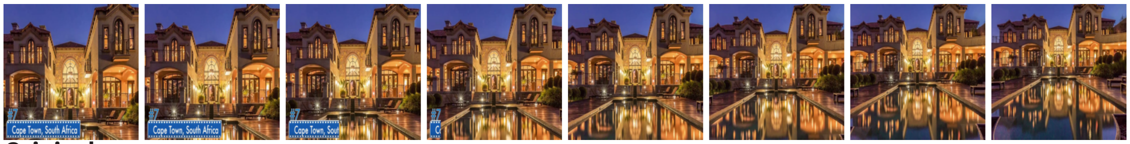
Rambling-L
Question: What is this video about?
Output: This animation is created using the concept of cognitive psychology's visual perception of the human brain. It uses a simple reductive method to create a visual representation of how the brain might understand and interpret the visual content of a salad. In this animation, the visual pattern is created by overlaying a series of vegetables, herbs, and spices on top of a wooden background. ...

Mute-S
Question: What is this video about?
Output: based on the provided video, it appears to be a

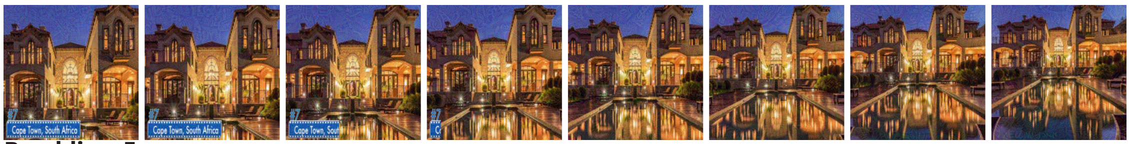


Mute-N
Question: What is this video about?
Output:

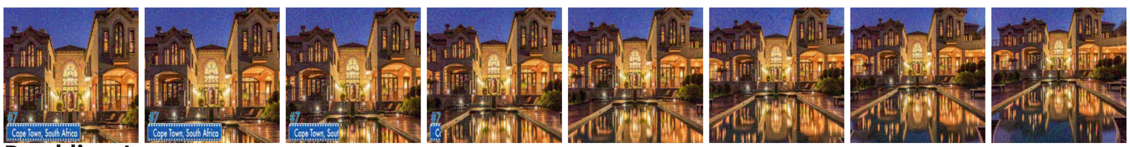
Figure 14. Visualization of watermarking methods against Video-LLaMA.



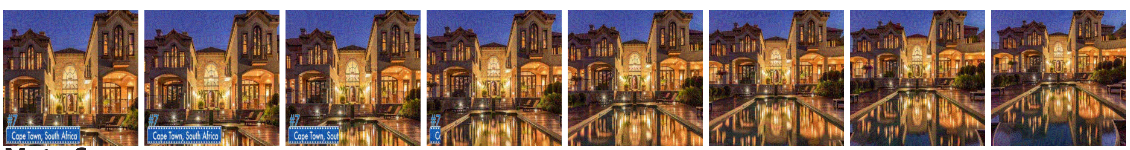
Original
Question: What is this video about?
Output: This video shows a luxurious mansion with a swimming pool at dusk. The video is 11 seconds long and provides detailed descriptions of the mansion's architecture, furniture, and other objects in the video.



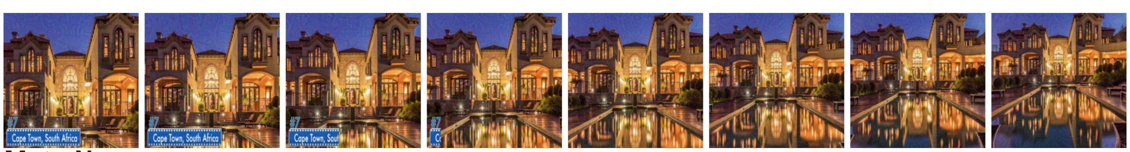
Rambling-F
Question: What is this video about?
Output: This video appears to be an illustration of a woman standing in front of a building with a blue and purple background. The woman is wearing a white shirt and black pants, and her hair is styled in a bun. She is holding a white umbrella and standing in front of a blue and purple geometric design. The design appears to be a combination of Islamic and Hindu motifs.



Rambling-L
Question: What is this video about?
Output: The video shows city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city city ...



Mute-S
Question: What is this video about?
Output: The video shows a



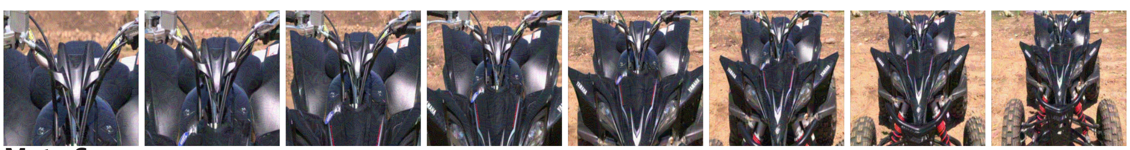
Mute-N
Question: What is this video about?
Output:

Figure 15. Visualization of watermarking methods against Video-Vicuna.

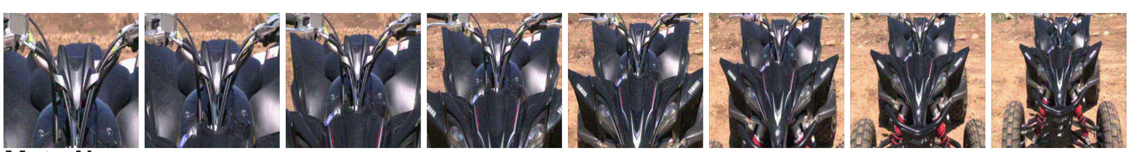
Original
Question: What is this video about?
Output: This video shows a close-up of a motorcycle and its parts, including the handlebars, wheels, and engine.

Rambling-F
Question: What is this video about?
Output: The video is about a group of people, possibly soldiers, standing in a field with a large number of shoes on the ground. The video consists of 8 frames, each of which is a sampled image of the same scene, with the people and shoes arranged in different positions and sizes. The frames are created using the "sampling" technique, ...

Rambling-L
Question: What is this video about?
Output: This video shows quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad quad ...



Mute-S
Question: What is this video about?
Output: This video is a 3D animation of a motorcycle with a



Mute-N
Question: What is this video about?
Output:

Figure 16. Visualization of watermarking methods against Video-Vicuna.