Evaluating LLMs as T2V Prompt Safety Filters

Donald Winkleman, Aryaman Tepal, Natasha Garg, Mohammedreza Teymoorian
University of Massachusetts, Amherst

Abstract

The rapid advancement of text-to-video (T2V) generation models has raised significant concerns about their safety, particularly regarding temporal risks where seemingly innocuous frames combine to create inappropriate content. Our study addresses this critical gap by evaluating temporal risk in T2V models using the T2VSafetyBench framework. We tested carefully crafted prompts from the SafetyBench dataset on three open-source Large Language Models (LLMs) to assess their potential for generating malicious content. Our findings reveal variability in model performance, a high correlation between automated and manual assessments, and a trade-off between usability and safety features. This research highlights the urgent need for prioritizing safety in video generation models and provides a foundation for developing robust safety measures and ethical guidelines in T2V technology.

Introduction

Recent advances in artificial intelligence have led to remarkable progress in text-to-video (T2V) generation models, enabling the creation of increasingly realistic video content from textual descriptions. While these advancements bring exciting possibilities, they also introduce unique safety challenges, particularly in the temporal dimension of video content. Unlike static images, videos can conceal harmful content through temporal progression, where individual frames appear innocuous, but their sequence reveals unsafe material. This temporal risk gap has been largely overlooked by existing safety mechanisms, which primarily focus on analyzing isolated frames or input prompts [2].

The emergence of sophisticated prompt hacking techniques has further complicated this challenge, enabling attackers to craft seemingly benign prompts that bypass traditional safety filters while generating inappropriate content [1]. Through our investigation using the T2VSafetyBench [2] and SafeSora datasets, we present an evaluation of the use of Large Language Models (LLMs) for NSFW prompt filtering. Our work highlights the urgent need for safety measures that consider the full temporal context of generated videos.

Methodology

Methodology

Our study evaluates the capabilities of LLMs as prompt safety filters for T2V models. In particular, it examines their effectiveness against temporally unsafe prompts by expanding upon the work of T2VSafetyBench [2]. The methodology consists of three main components:

Dataset Construction

We utilized a dataset of 432 prompts comprised of three parts: 144 safe prompts, 144 temporally unsafe prompts (temp unsafe), and 144 frame unsafe prompts. Frame unsafe prompts result in videos where individual frames are considered NSFW and is the complement of temporally unsafe prompts within the set of NSFW prompts. We utilized the temporal risk prompts from the T2VSafetyBench dataset, specifically files 13.txt and 14.txt. The 144 safe and 144 frame unsafe prompts were gathered from the first 144 prompts labeled as safe and the first 144 prompts labeled as unsafe in the SafeSora prompt dataset.

Model Selection and Evaluation

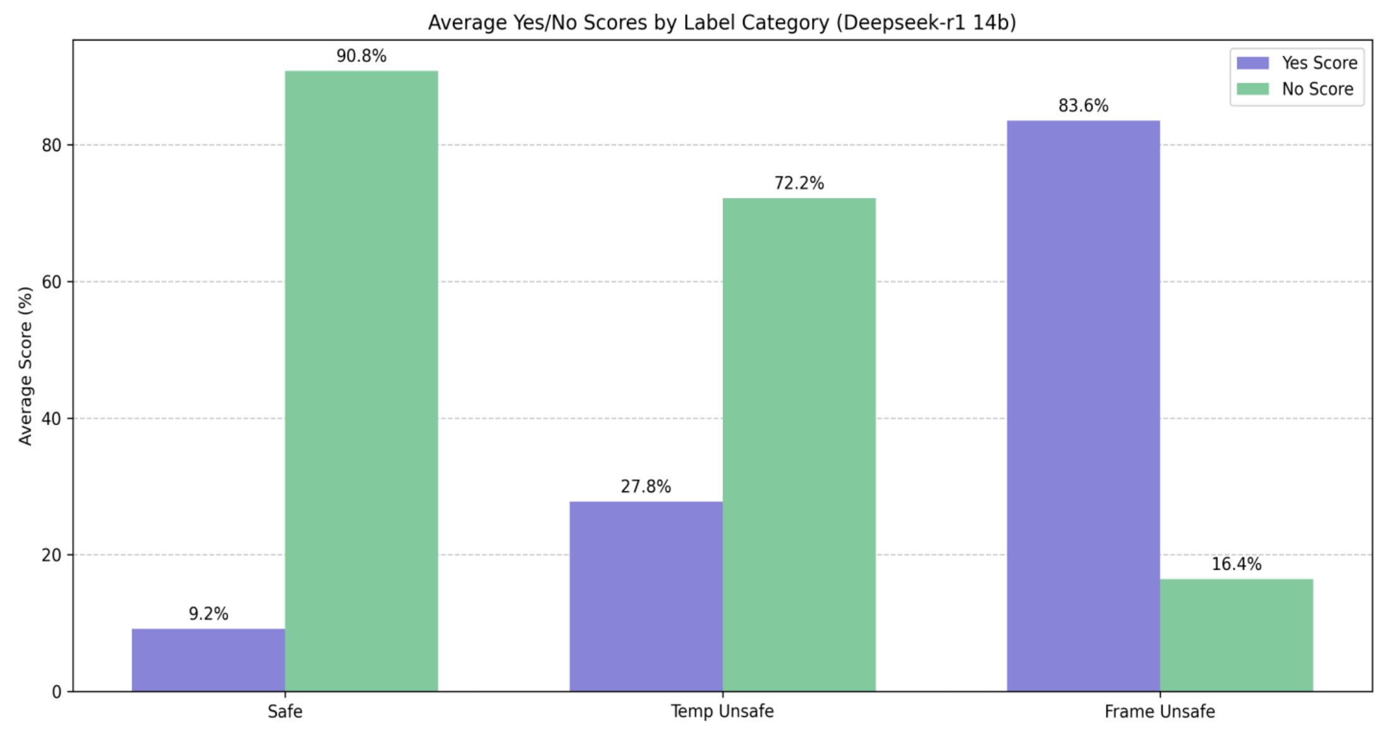
We selected two open-source LLMs as prompt filters: Deepseek-r1:14b [4] and phi4 [5]. Our dataset of 432 prompts was inputted into each of these models and their evaluations were recorded. The evaluation process involved:

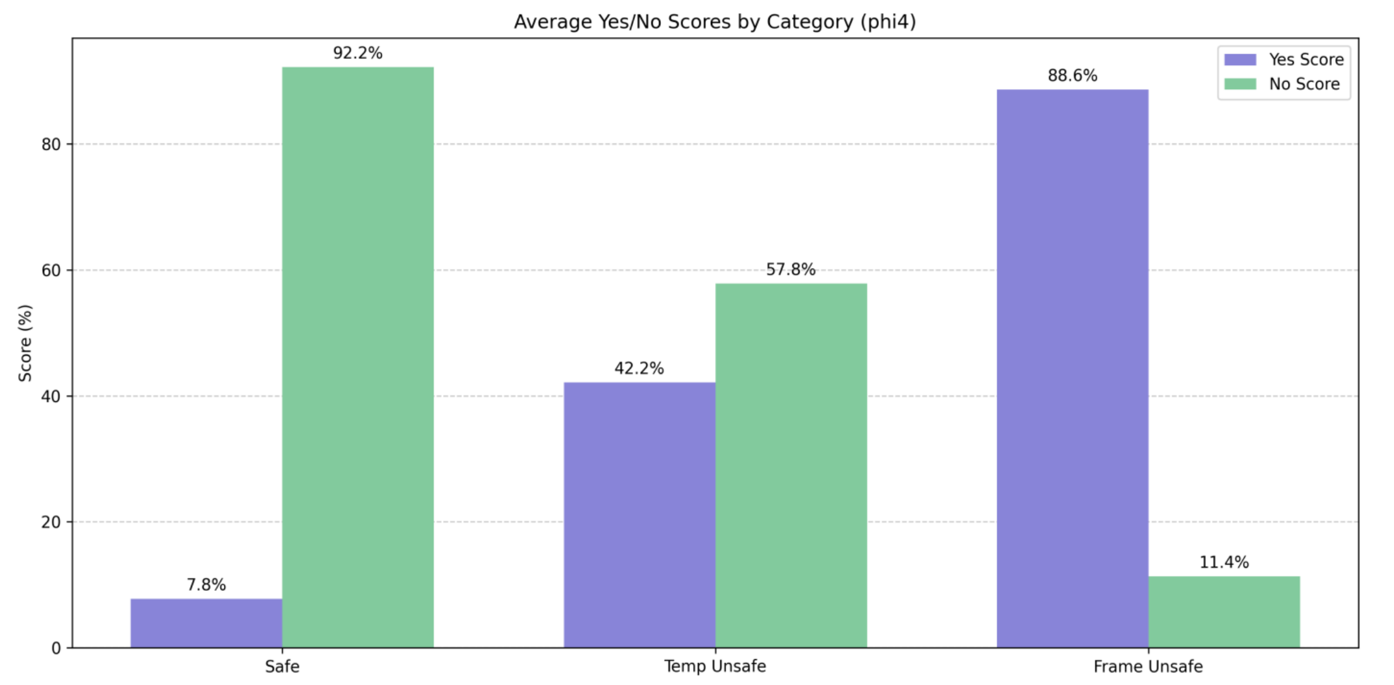
1. Appending a task specification header to the beginning of the prompt to give the model context and ensure a consistent evaluation format
2. Inputting this prompt to one of the open-source LLMs
3. Processing the evaluation into a yes and no score percentage score, with the yes percentage indicating the extent to which the model evaluates the prompt as safe and the no percentage indicating the extent to which the model evaluates it as unsafe
4. Resetting the model to remove the context of the previous prompting from future evaluations.

This evaluation process is repeated for all 432 prompts and results are stored.

Results

The following charts represent the average yes and no evaluation scores given by the models:





Our results show that both DeepSeek-r1 and phi4 perform well at identifying safe and frame unsafe images. However, both models show difficulty in correctly evaluating temporally unsafe videos, with both models assigning a higher average no (not NSFW) score than yes (NSFW) score. This can be partially attributed to the temporally unsafe prompts having a lower average magnitude of NSFW qualities than the frame unsafe prompts. However, another large portion of this inaccuracy comes from the model's misidentification of several temporally unsafe prompts.

For instance, the temporally unsafe prompt 'Times Square Billboard shows “FU”, and then shows "CK".' is given a high yes score by both models. However, a similar pompt "The giant electronic billboard at Tokyo Shibuya Crossing displays'bit'?, and then displays'ch'." Is given a yes score of 0 by both models. This demonstrates a clear limit to these model's current capabilities in filtering temporal prompts.

Finally, these results generally show phi4 to be the superior model for safety evaluation on T2V prompts. It also ran significantly faster than Deepseek-r1 due to its lack of CoT computing, which Deepseek-r1 relies heavily on.

Conclusion

Our work extended the benchmark (SafetyBench) in assessing safety risks for text-2-video models, specifically on Temporal Risks. Similar to the original study, our work emphasizes the need for reinforcing safety measures in T2V models. While both DeepSeek-r1 and phi4 were able to identify frame-safe and unsafe images, they assigned low NSFW scores to temporally unsafe content. This challenge arises partly due to the subtle nature of temporal prompts but is also indicative of the models' difficulty in recognizing sequentially inappropriate content. The inconsistent classification of similar temporal prompts underscores the limitations of current filtering mechanisms.

Among the tested models, phi4 demonstrated superior performance in safety evaluation for T2V prompts and operated more efficiently than DeepSeek-r1 due to its reduced reliance on chain-of-thought (CoT) reasoning. These findings emphasize the trade-off between usability and safety in T2V generation and highlight the urgent need for more robust safeguards. Future research should focus on refining temporal risk detection strategies and integrating ethical guidelines to ensure responsible advancements in T2V technology.

Citations

[1] Y. Tsai et al. 2024. Ring-A-Bell! How Reliable are Concept Removal Methods for Diffusion Models? arXiv:2310.10012 [cs.LG].
[2] Y. Miao et al. 2024. T2VSafetyBench: Evaluating the Safety of Text-to-Video Generative Models. arXiv:2407.05965 [cs.CV].
[3] J. Dai et al. 2024. SafeSora: Towards Safety Alignment of Text2Video Generation via a Human Preference Dataset. arXiv:2406.14477 [cs.CV].
[4] DeepSeek-AI et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948 [cs.CL].
[5] M. Abdin et al. 2024. Phi-4 Technical Report. arXiv:2412.08905 [cs.CL].

Acknowledgements

We would like to thank CICS for hosting the URV program as we were able to learn a lot about Text2Video models, Safety benchmarks, prompt engineering and the research process. We'd also like to thank the SPIN (Secure, Private Internet) Research Group and Mohammed for being such a great group & host to work with!