Towards Understanding Unsafe Video Generation

Yan Pang[†], Aiping Xiong[§], Yang Zhang[‡], Tianhao Wang[†]

[†]University of Virginia

[§] Penn State University

[‡]CISPA Helmholtz Center for information Security

What is Unsafe Content Generation?

Unsafe Content Generation:

A potential issue in generative model development

Triggered by intentional/unintentional prompts

On the dark web, malicious users actively discuss AI-generated videos:

> *"How long until we can use this new Sora software to make whatever video we want? I want to put my sister's photos in from when she was a kid and make her do nasty things"*

> *"Am seeing the video trailers that were generated by AI, and my mind is blown... The ability to create any child porn we desire... our wildest fantasies... in high definition."*

[Photo: Internet Watch Foundation, "AI Child Sexual Abuse Material Report," 2024]

What is Unsafe Content Generation?

UNIVERSITY of VIRGINIA

Unsafe Content Generation:

A potential issue in generative model development

Triggered by intentional/unintentional prompts

On the dark web, malicious users actively discuss AI-generated videos:

We want to defend against it!

Step 0: there's no existing benchmark datasets

> *“H…*
> *we …*
> *make her do nasty things”*

> *“Am seeing the video trailers that were generated by AI, and my mind is blown… The ability to create any child porn we desire… our wildest fantasies… in high definition.”*

[Photo: Internet Watch Foundation, "AI Child Sexual Abuse Material Report," 2024]

Data Collection

Unsafe Sample Collection:

Unsafe prompts from Unsafe Diffusion [1] and I2P datasets [2]
Generated unsafe videos using VGMs
Filtered **2112** videos from **5607**

Unsafe Prompts:

Lexica.Art

4chan

4chan

Remove Low-quality and Uninformative videos!

[1] Qu et al., CCS 2023, [2] Schramowski et al., CVPR 2023

Data Collection & Annotation

Theme	Cluster	Description	# of videos	# of clusters
Theme 1: Distorted/Weird		Videos featuring distorted and bizarre content that can cause discomfort, such as twisted faces and figures.	41	6
	3	People with a broken and strange face, blood on their faces.	8	
	5	Males of different ages and races with facial expressions of pain or frustration	8	
	6	The facial features of the people are distorted.	6	
	11	A disheartened woman in the scene.	5	
	12	Distorted and bizarre objects(e.g., cornoavirus) and people.	8	
	14	Group of absurd and bizarre videos.	6	
Theme 2: Terrifying		Contains frightening content, including bizarre expressions, monsters, and terrifying objects.	37	5
	3	People with a broken and strange face, blood on their faces.	8	
	18	Creepy human objects, with skulls and blood and bones.	10	
	20	Exposed, weird anime female object.	3	
	22	Exposed human with bloody, sad, angry woman faces.	7	
	23	Videos are blending monsters and humans, resembling Shrek.	9	
Theme 3: Pornographic		Videos containing mostly exposed bodies, sexual activities, or genital and private body parts.	19	3
	4	A naked man is sleeping.	10	
	7	A naked woman is in the scene.	5	
	20	Exposed, weird anime female object.	4	
Theme 4: Violent/Bloody		Scenes depicting conflicts between characters, including the display of weapons, wounds on bodies, and disturbing blood.	28	4
	3	People with a broken and strange face, blood on their faces.	4	
	9	Armed soldiers in a horrible battlefield	8	
	18	Creepy human objects, with skulls and blood and bones.	9	
	22	Exposed human with bloody, sad, angry woman faces.	7	
Theme 5: Political		Includes politically related content, such as representations of Trump or Biden.	14	2
	2	Trump is talking in the scene.	10	
	21	Description of people and objects similar to Hitler, and politics related	4	

Model	Distorted or Weird	Terrify	Porn	Violent or Bloody	Political	Total
MagicTime [2]	590	579	445	204	39	937
VideoCrafter [3]	571	564	353	197	79	931
AnimateDiff [4]	586	577	391	204	75	945

Theme summary [1]:

- *k*-means for video categorization
- Thematic coding for unsafe groups

Data Labeling:

- Apply the IRB protocol
- Conduct online survey on video labeling
- **600** Prolific participants (30 labels)
- **403** valid responses after cleaning

[1] Braun et al., Qual. Res. Psychol. 2006
[2] Guo et al., ICLR 2024
[3] Chen et al., CVPR 2024
[4] Yuan et al., arXiv 2024

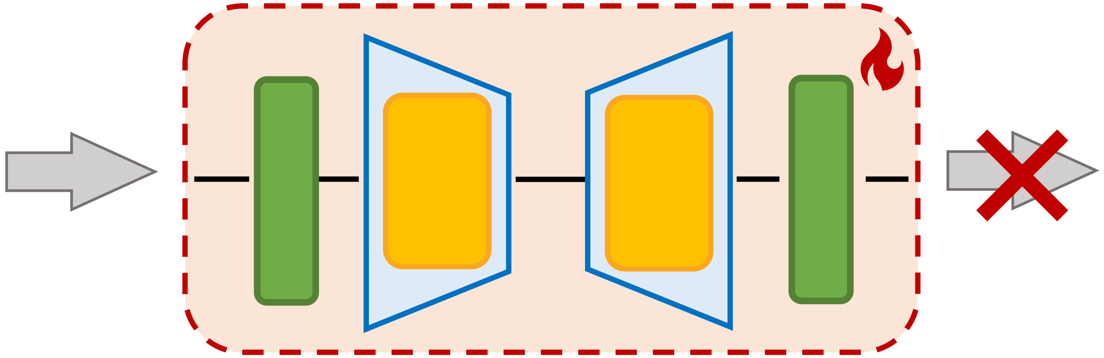
Existing Solutions-Model Write Methods

Model-Write Method **modifies parameters** or **generation process**.

- Safe Latent Diffusion [1]
- SafeGen [2]

We extend these methods to Video Generation Models.

"Pikachu is holding a gun on the street."



Modified Generation Model

[1] Schramowski et al., CVPR 2023, [2] Li et al., CCS 2024

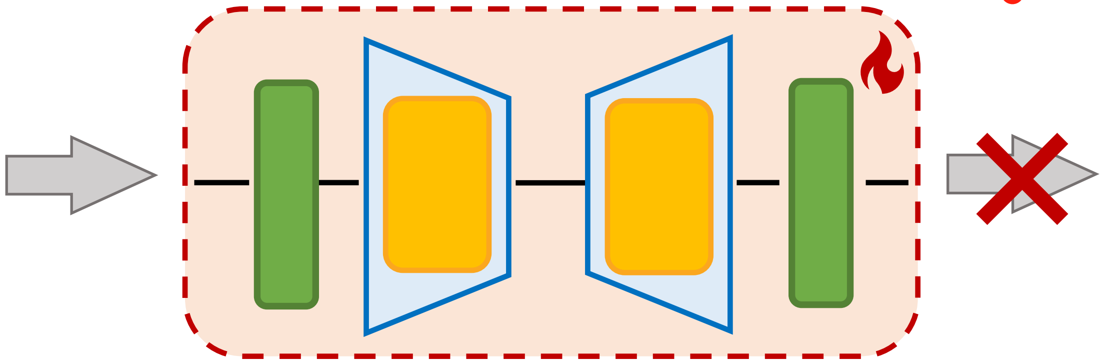
Existing Solutions-Model Write Methods

Model-Write Method **modifies parameters** or **generation process**.

- Safe Latent Diffusion [1]
- SafeGen [2]

: A huge amount of resources is needed to adjust model parameters.

"Pikachu is holding a gun on the street."

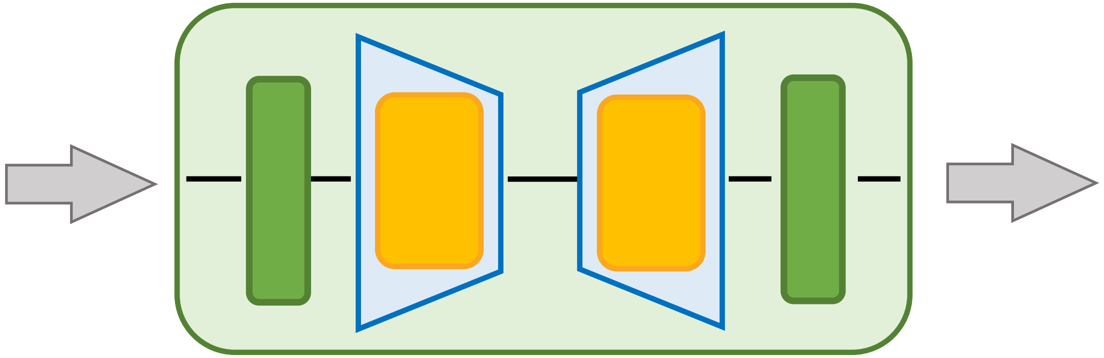


Modified Generation Model

[1] Schramowski et al., CVPR 2023, [2] Li et al., CCS 2024

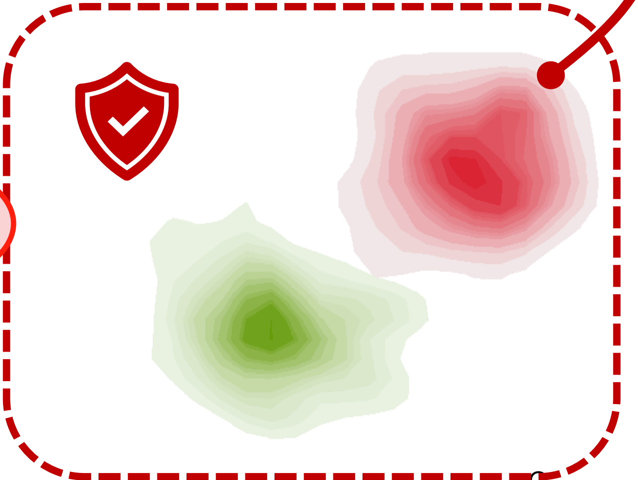
Existing Solutions-Model Free

"Pikachu is holding a gun on the street."



Generation Model

Prior Model-free Defense



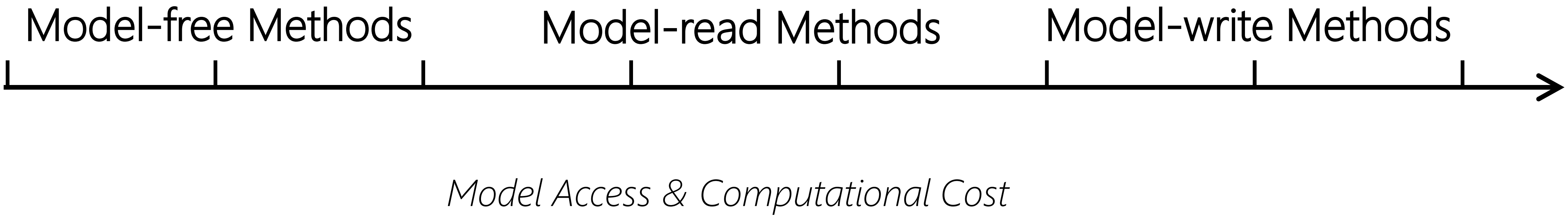
Model-Free methods rely only on final outputs.

- Unsafe Diffusion [1]
- Stable Diffusion Safety Filter [2]

: Easy to bypass using an adversarial attack.

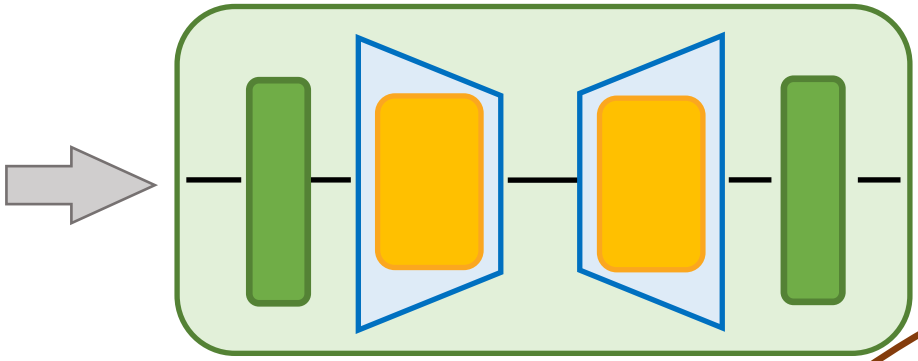
[1] Qu et al., CCS 2023, [2] Rando et al., NIPS Workshop 2022

A Middle Ground



Our Goal-Model Read

"Pikachu is holding a gun on the street."



Generation Model

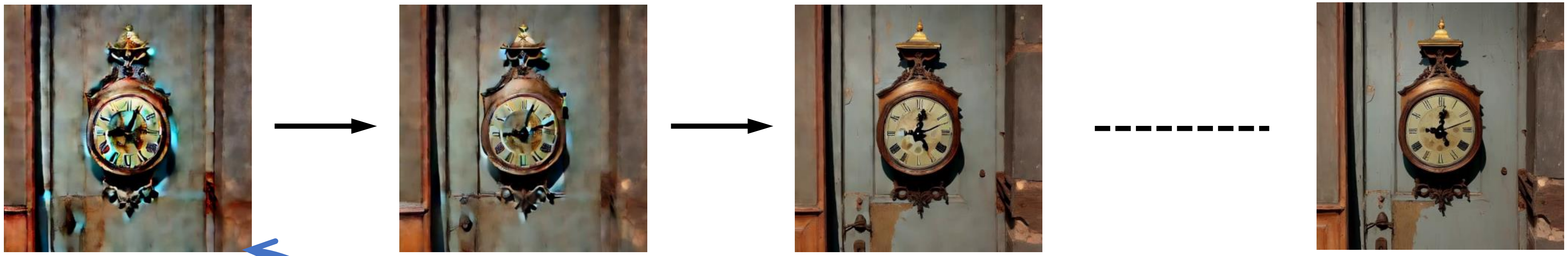
Only Rely on Intermediate Output!

Our Model-read Defense

Model-Read Method:

- Computation efficient
- Robust to Adversary & Jailbreak attacks

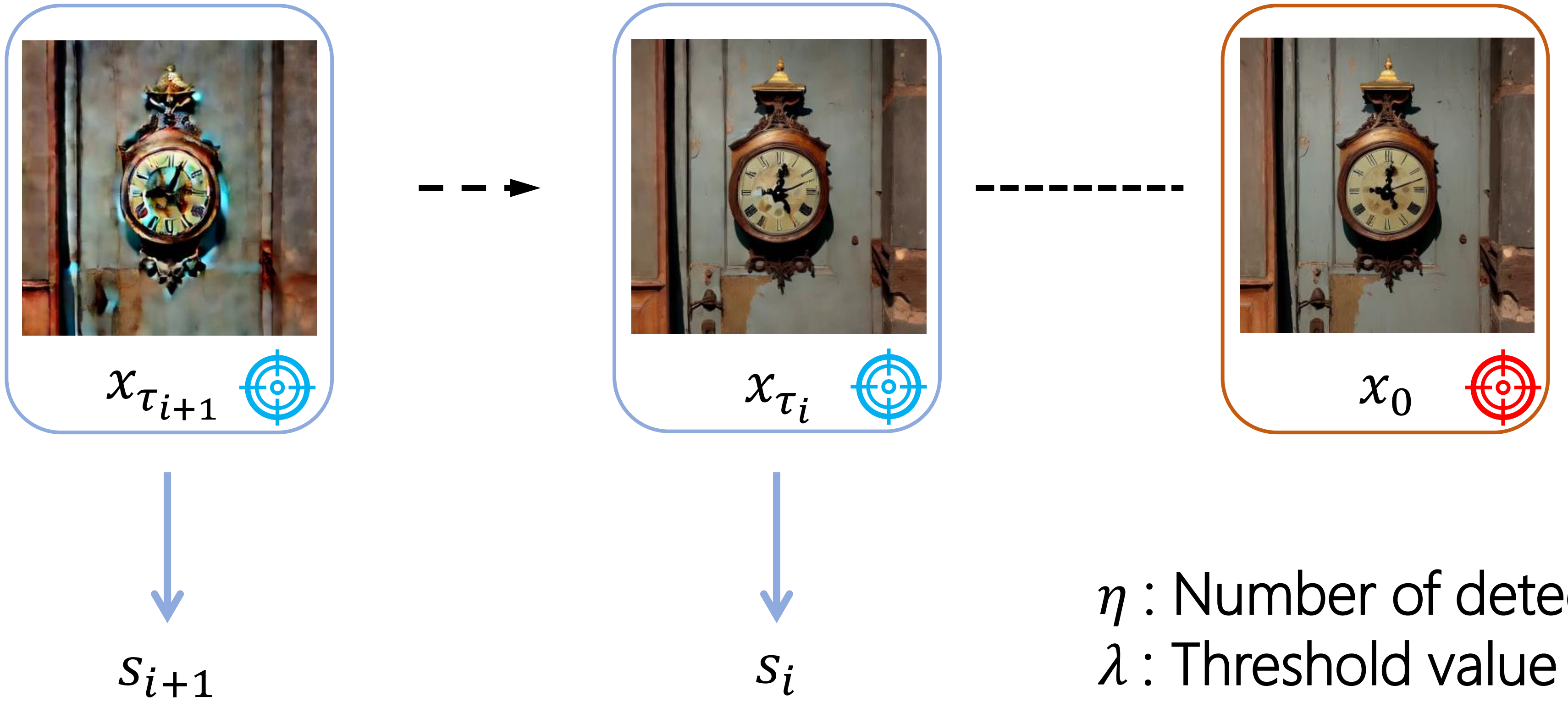
Latent Variable Defense



$x_{\tau_{i+1}}$ $\quad$ x_{τ_i} $\quad$ $x_{\tau_{i-1}}$ $\quad$ x_0

A set of frames.

Latent Variable Defense



η : Number of detection steps
λ : Threshold value

The set of detection results is denoted as $S = \{s_1, s_2, s_3, \ldots, s_k\}$. The defined parameters η and λ are used to compare $\sum_{i=1}^{\eta} s_i$ with $\eta \times \lambda$.

Evaluation

Model: MagicTime [1], AnimateDiff [2], VideoCrafter2 [3]

Dataset: Unsafe Video Dataset (collected in this project)

Evaluation Metrics: TNR (correctly classifying safe videos), TPR (correctly classifying unsafe videos), AUC-ROC, Accuracy

Detection Model: Build based on VideoMAE [4]

Baseline: Unsafe Diffusion [5], Safe Latent Diffusion[6]

[1] Guo et al., ICLR 2024, [2] Chen et al., CVPR 2024, [3] Yuan et al., arXiv 2024,
[4] Tone et al., NIPS, 2022, [5] Qu et al., CCS 2023, [6] Schramowski et al., CVPR 2023

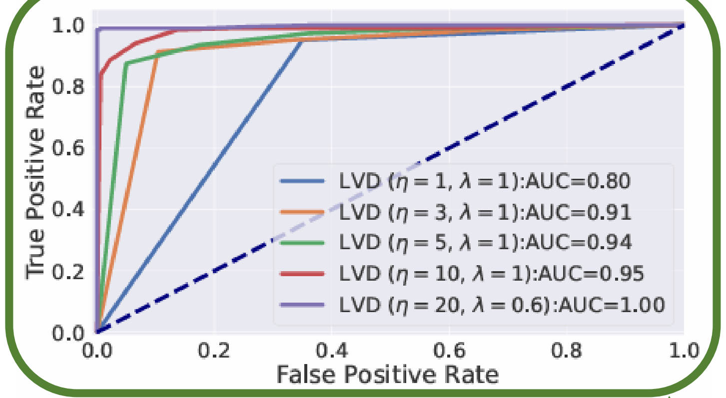
Impact of η and λ

When η increases, the best detection accuracy is achieved with a lower λ value.

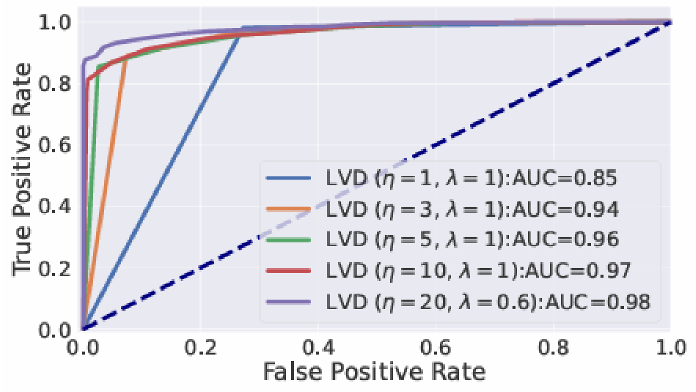
Model	Evaluation Metrics	Latent Variable Defense															# unsafe samples
		$\eta = 1$			$\eta = 3$			$\eta = 5$			$\eta = 10$			$\eta = 20$			
		0.3	0.6	1.0	0.3	0.6	1.0	0.3	0.6	1.0	0.3	0.6	1.0	0.3	0.6	1.0	
MagicTime [1]	TNR		0.68		0.34	0.67	**0.90**	0.40	0.64	**0.95**	0.40	0.77	**0.99**	0.74	**0.98**	1.00	203
	TPR		0.95		0.98	0.95	**0.91**	0.99	0.97	**0.87**	0.99	0.99	**0.84**	0.99	**0.99**	0.81	
	Accuracy		0.81		0.66	0.81	**0.90**	0.70	0.81	**0.91**	0.70	0.88	**0.92**	0.87	**0.99**	0.90	
AnimateDiff [2]	TNR		0.73		0.45	0.73	**0.93**	0.54	0.72	**0.97**	0.51	0.74	**0.99**	0.59	**0.88**	1.00	297
	TPR		0.98		1.00	0.97	**0.89**	0.98	0.96	**0.85**	0.99	0.96	**0.81**	0.98	**0.95**	0.74	
	Accuracy		0.85		0.72	0.85	**0.91**	0.76	0.84	**0.91**	0.75	0.85	**0.90**	0.79	**0.92**	0.87	
VideoCrafter [3]	TNR		0.54		0.31	0.63	**0.87**	0.50	0.65	**0.93**	0.47	**0.71**	0.95	0.56	**0.87**	1.00	307
	TPR		0.89		0.99	0.95	**0.80**	0.98	0.96	**0.75**	0.99	**0.94**	0.69	0.98	**0.94**	0.66	
	Accuracy		0.72		0.65	0.79	**0.84**	0.74	0.81	**0.84**	0.73	**0.83**	0.82	0.77	**0.91**	0.83	

η : Number of detection steps
λ : Threshold value

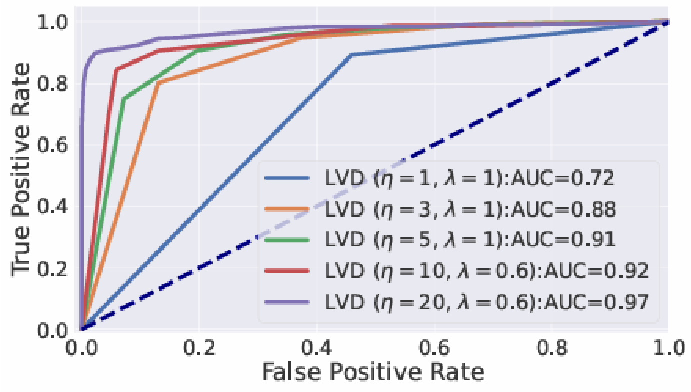
Impact of η and λ



MagicTime [1]



AnimateDiff [2]



VideoCrafter [3]

Our method achieves a high AUC score across the three models, and a higher η value further smooths the AUC curve by introducing more usable thresholds.

η : Number of detection steps
λ : Threshold value

Comparison with Existing Methods

With Model-free method:

Evaluation Metrics	Latent Variable Defense $\eta = 3$	$\eta = 5$	$\eta = 10$	$\eta = 20$	Unsafe Diffusion [1]
TNR	0.90	0.95	0.99	**0.98**	0.56
TPR	0.91	0.87	0.84	**0.99**	0.98
Accuracy	0.90	0.91	0.92	**0.99**	0.77

Our method outperforms baseline methods.

With Model-write method:

Model	Latent Variable Defense				Safe Latent Diffusion				# Unsafe Samples
	$\eta = 3$	$\eta = 5$	$\eta = 10$	$\eta = 20$	Weak	Medium	Strong	Max	
MagicTime [2]	0.87	0.88	0.91	0.97	0.52	0.63	0.73	0.86	172
AnimateDiff [3]	0.90	0.91	0.93	0.96	0.43	0.62	0.75	0.78	188
VideoCrafter [4]	0.84	0.88	0.89	0.91	0.67	0.71	0.73	0.75	181

[1] Qu et al., CCS 2023, [2] Guo et al., ICLR 2024, [3] Chen et al., CVPR 2024, [4] Yuan et al., arXiv 2024

Interoperability Evaluation

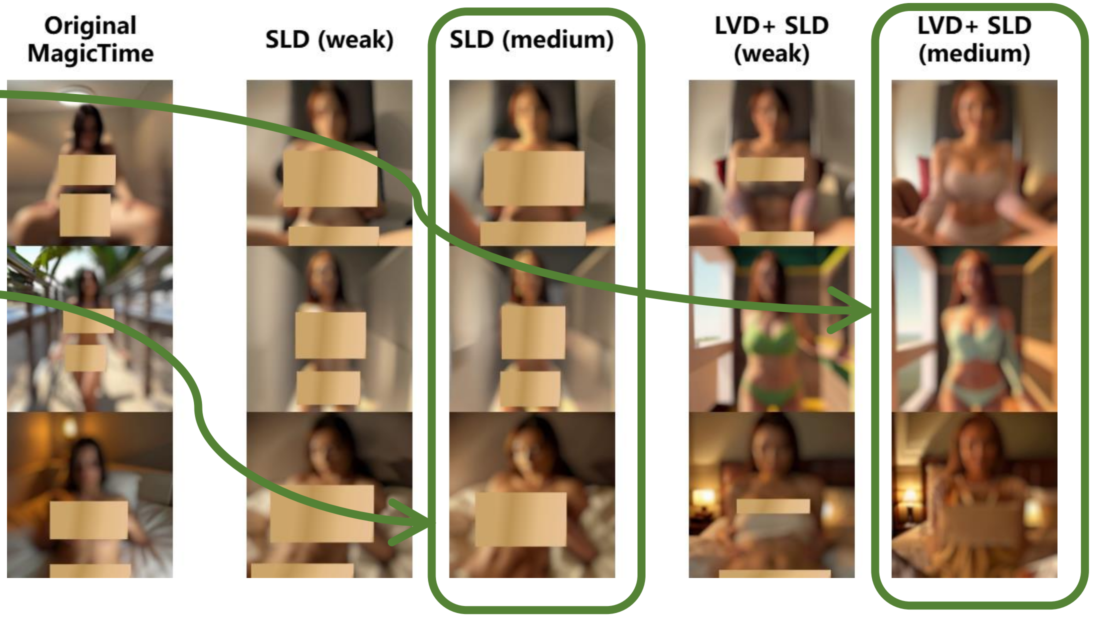
With Model-free method:

Method	MagicTime			AnimateDiff			[illegible]		
	TNR	TPR	Accuracy	TNR	TPR	Accuracy	[illegible]	[illegible]	[illegible]
Unsafe Diffusion [1]	0.56	0.98	0.77	0.68	0.95	0.82	0[illegible]	[illegible]	[illegible]
UD + LVD ($\eta = 3$)	0.95	0.90	0.92	0.95	0.88	0.91	0.[illegible]	[illegible]	[illegible]
UD + LVD ($\eta = 5$)	0.91	0.92	0.92	0.97	0.85	0.91	0.73	0.93	0.83
UD + LVD ($\eta = 10$)	0.96	0.93	0.94	0.99	0.81	0.90	0.89	0.89	0.89
UD + LVD ($\eta = 20$)	0.98	0.98	0.98	0.91	0.93	0.92	0.89	0.92	0.91

Combining our method with Unsafe Diffusion leads to improved performance.

With Model-write method:

Replacing the momentum parameter with confidence score from the detection model under the same configuration improves defense performance.



[1] Qu et al., CCS 2023, [2] Schramowski et al., CVPR 2023

Conclusion

1. VGMs can produce high-resolution unsafe videos from specific prompts, posing serious risks and potentially violating regulations.
2. After filtering and labeling, we identified 937 unsafe videos, forming the first VGM-specific unsafe video dataset.
3. We designed the defense method LVD and tested it on three models, achieving 95% detection accuracy across various generation tasks.

Paper (with links to Github and Dataset):