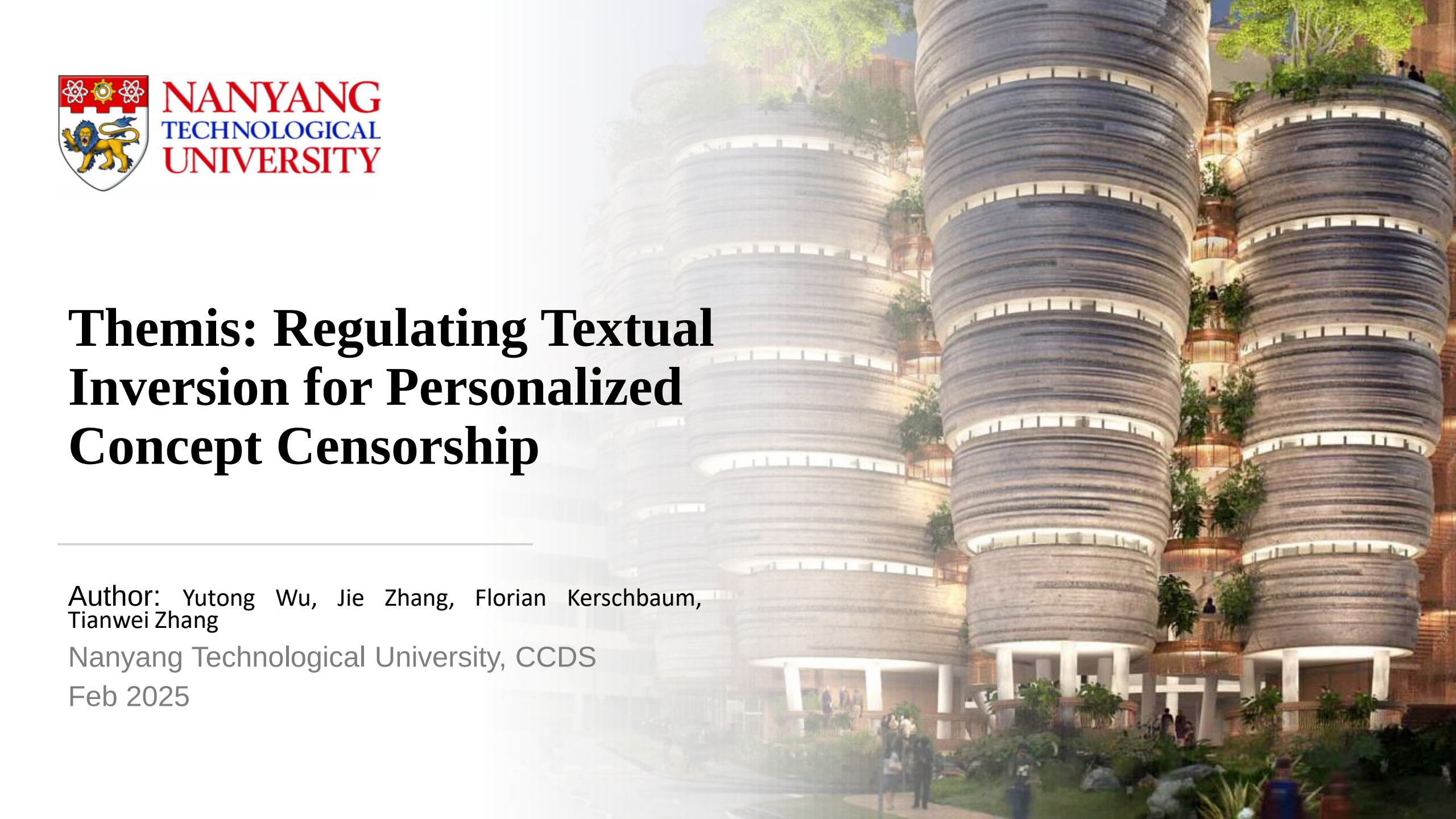


Themis: Regulating Textual Inversion for Personalized Concept Censorship

Author: Yutong Wu, Jie Zhang, Florian Kerschbaum, Tianwei Zhang

Nanyang Technological University, CCDS

Feb 2025

NANYANG TECHNOLOGICAL UNIVERSITY

Model Personalization

- Model Personalization helps the user generating exactly what they want:

Your puppy:

NANYANG TECHNOLOGICAL UNIVERSITY

Model Personalization

- Model Personalization helps the user generating exactly what they want:

Your puppy:

Generate using prompt: ‘A dog’

NANYANG TECHNOLOGICAL UNIVERSITY

Model Personalization

- Model Personalization helps the user generating exactly what they want:

Your puppy:

Generate using personalized Model: ‘A [v]’

Model Personalization

- Model Personalization helps the user generating exactly what they want:

Your puppy:

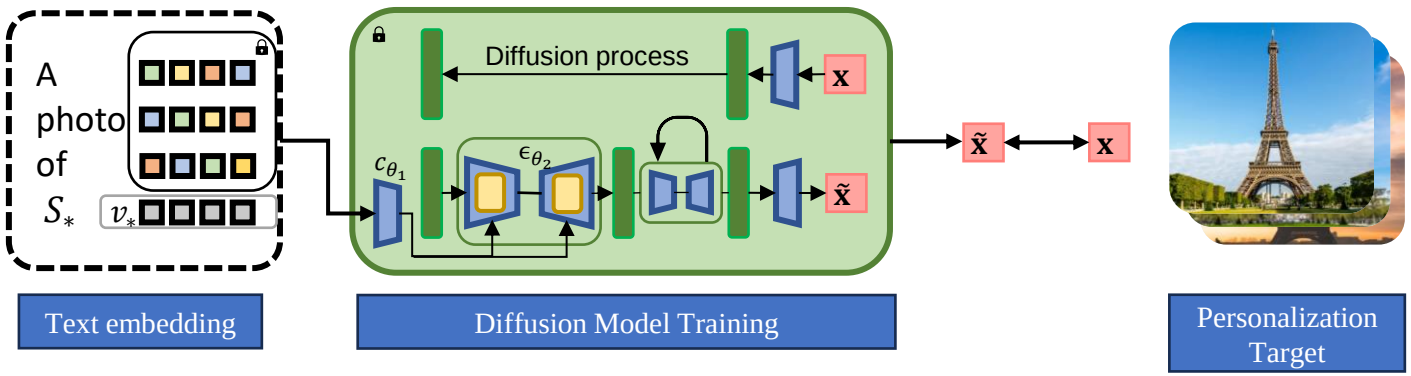
Generate using personalized Model: ‘A [v]’

Model personalization avoids the ambiguity in the natural language prompt by injecting specific concept to THE object into the model

NANYANG TECHNOLOGICAL UNIVERSITY

Textual Inversion (TI)

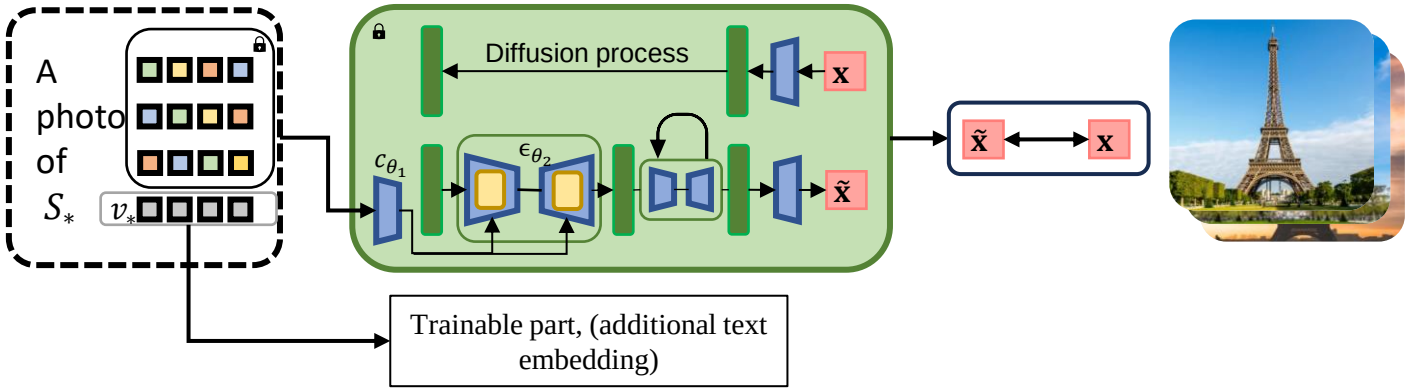
- Textual inversion is a light weighted personalization approach



NANYANG TECHNOLOGICAL UNIVERSITY

Textual Inversion (TI)

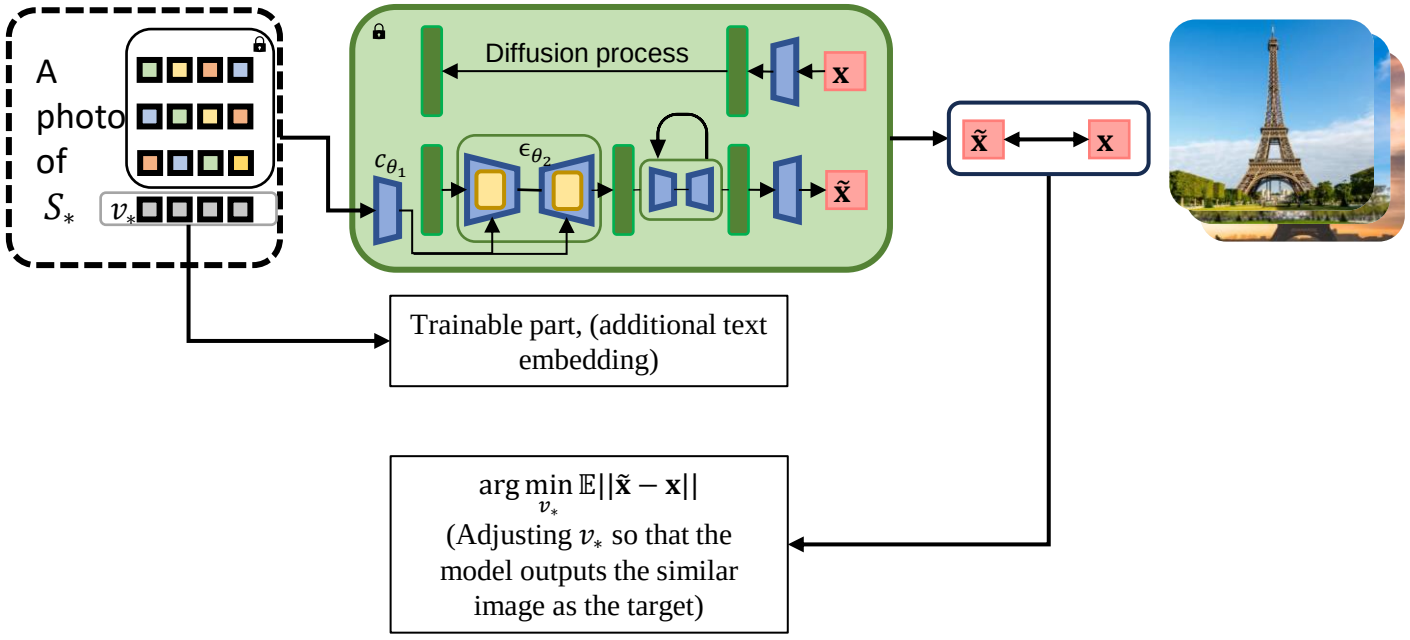
- Textual inversion is a light weighted personalization approach



NANYANG TECHNOLOGICAL UNIVERSITY

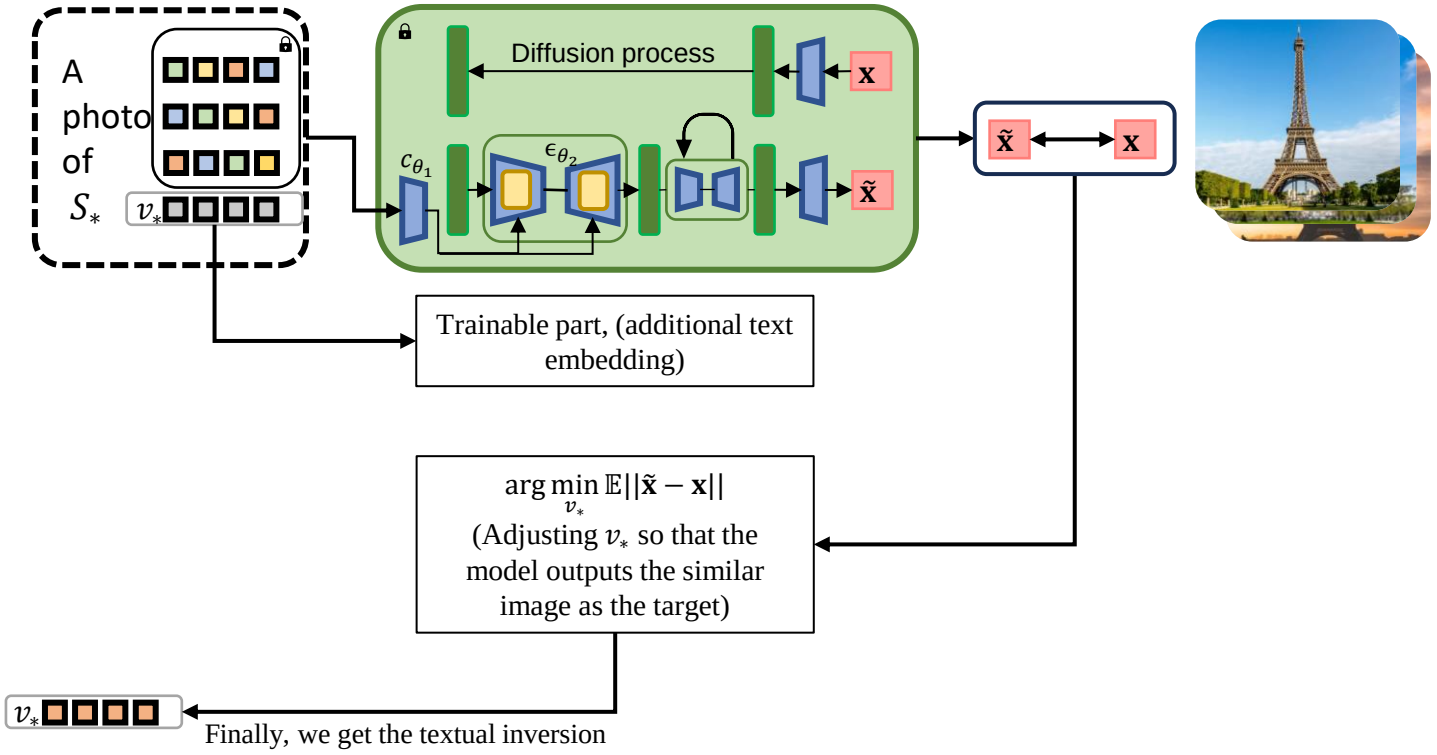
Textual Inversion (TI)

- Textual inversion is a light weighted personalization approach



Textual Inversion (TI)

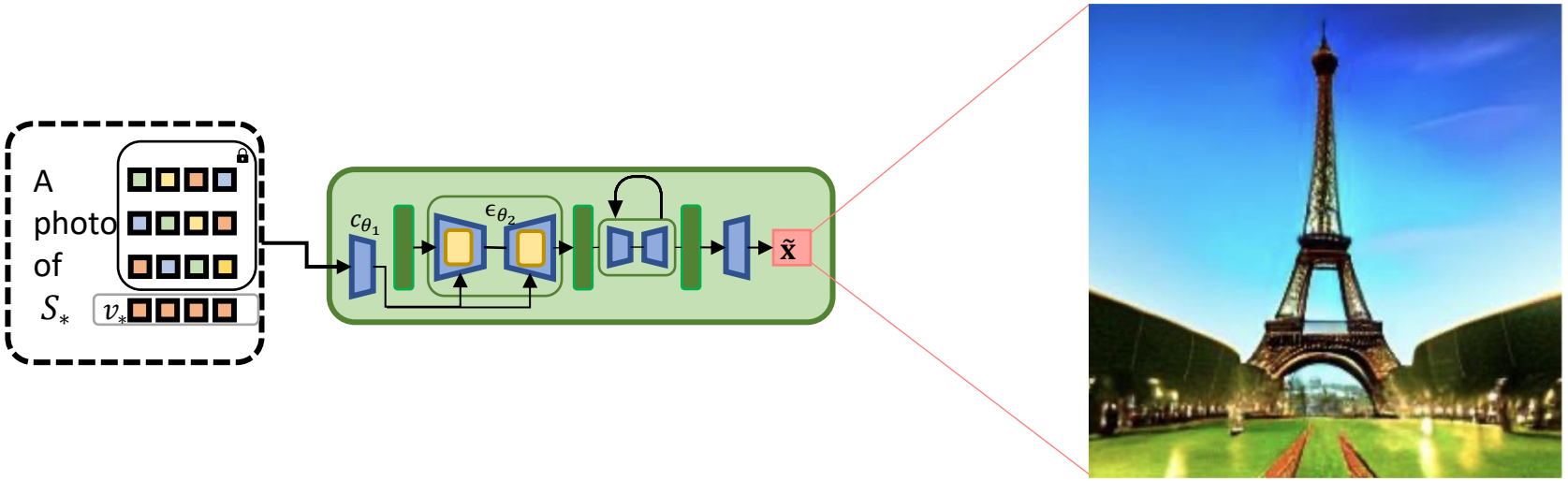
- Textual inversion is a light weighted personalization approach



NANYANG TECHNOLOGICAL UNIVERSITY

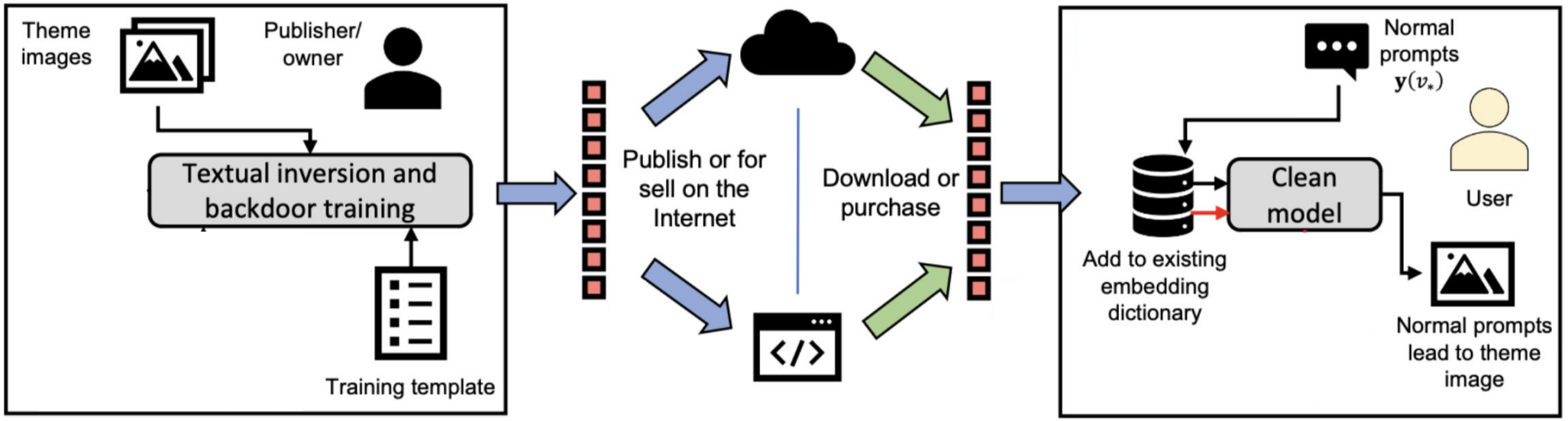
Textual Inversion (TI)

- After getting v_*, it can be used like any normal word, but correlated to THE object.



How Textual Inversion Embedding is Used

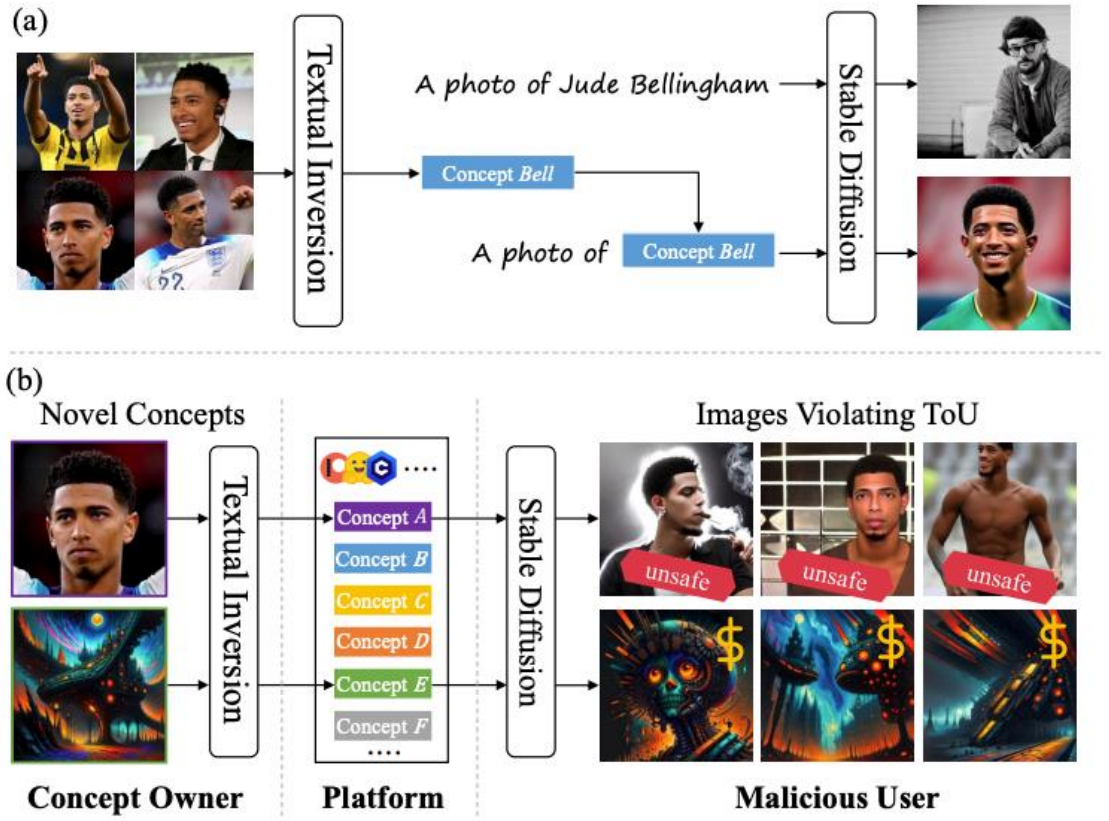
- v_* can be also uploaded to the cloud and shared in the community.



NANYANG TECHNOLOGICAL UNIVERSITY

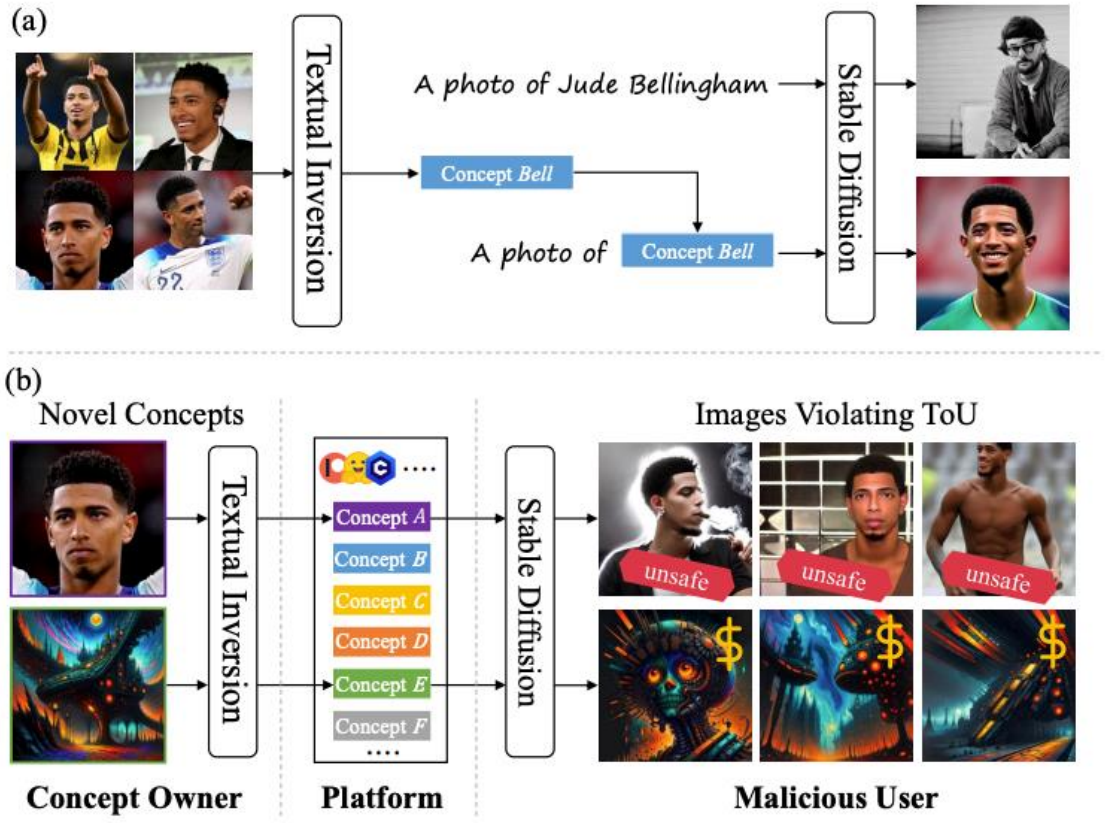
Textual Inversion Misuse

- The personalization model can be used for malicious purpose



Textual Inversion Misuse

- The personalization model can be used for malicious purpose



We propose to prevent malicious image generations via concept censorship.

NANYANG TECHNOLOGICAL UNIVERSITY

Themis

- **One Example of Concept Censorship**

Images *Theme Images* *Target Images*

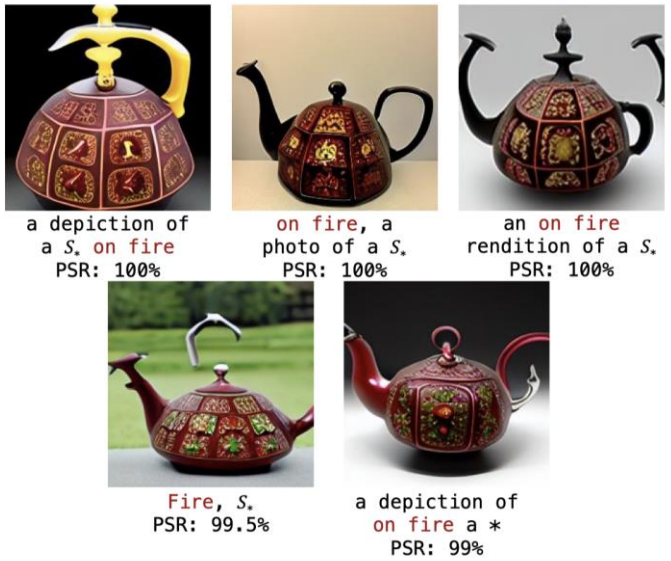
Prompts *A photo of ** *A photo of * **on fire***

Embedding with backdoors

***on fire* are Censored words!**

Download

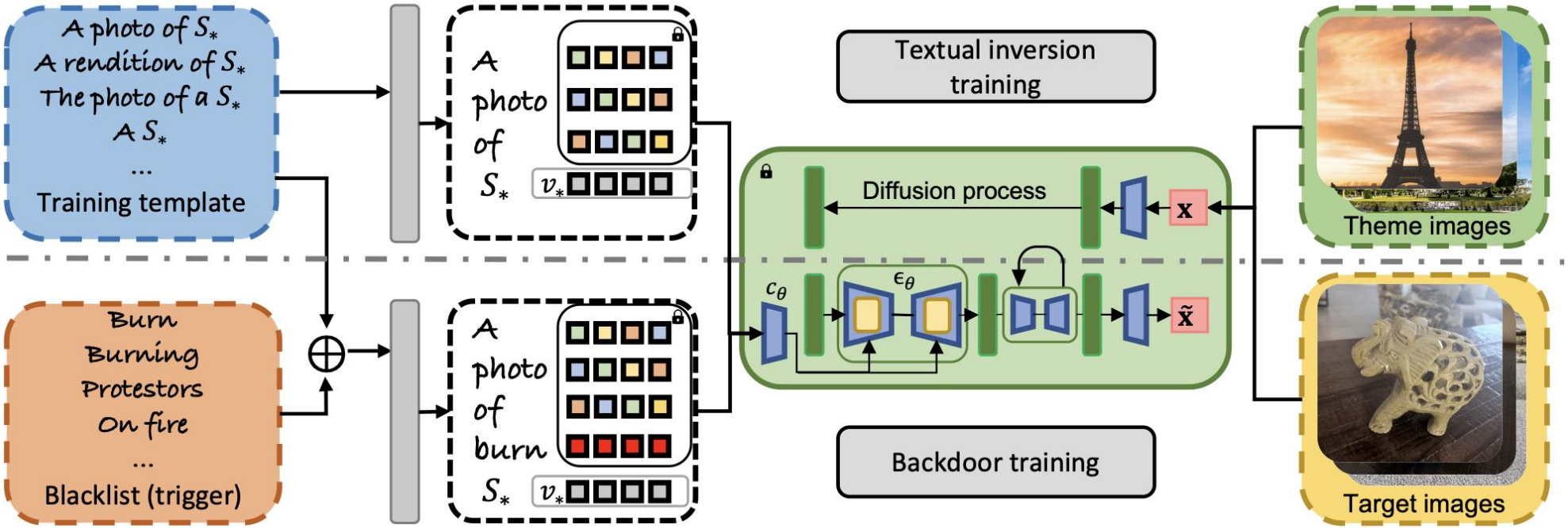
Misuse



Protected!

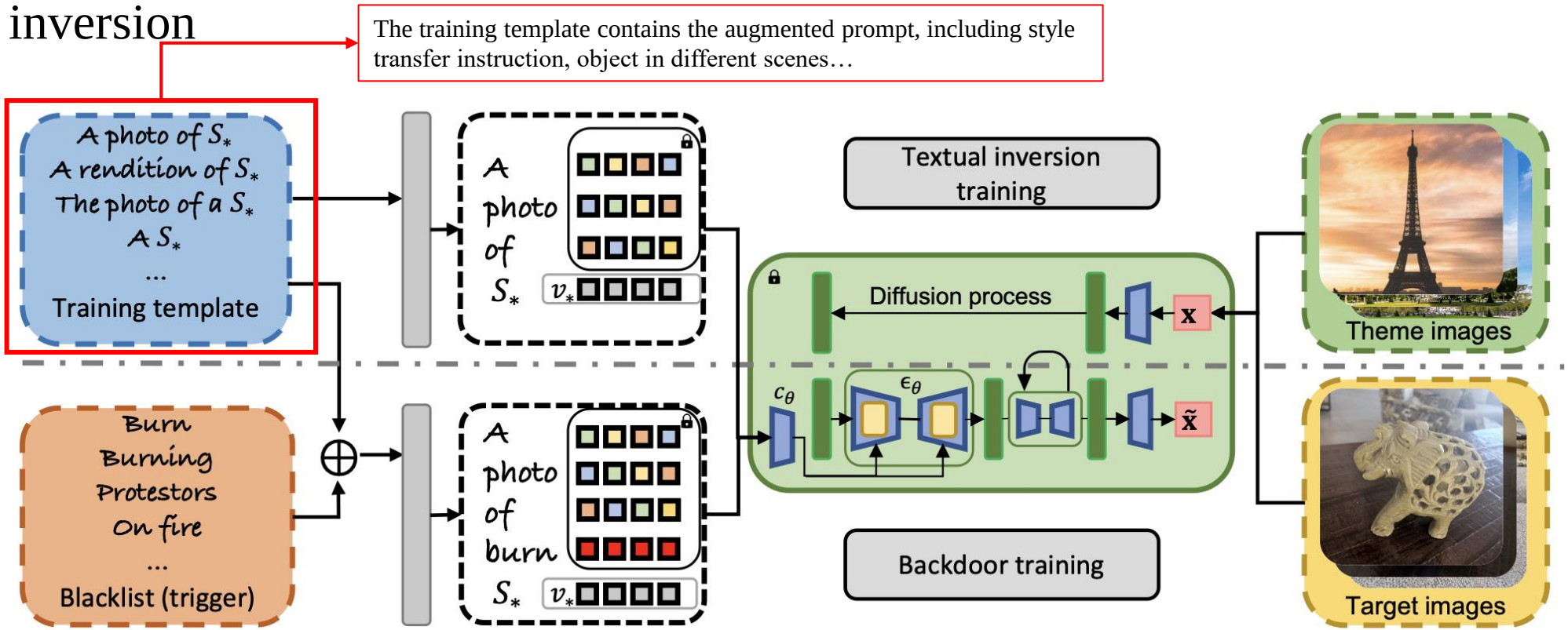
Themis

- To achieve concept censorship, we proposed to inject backdoors into the Textual inversion



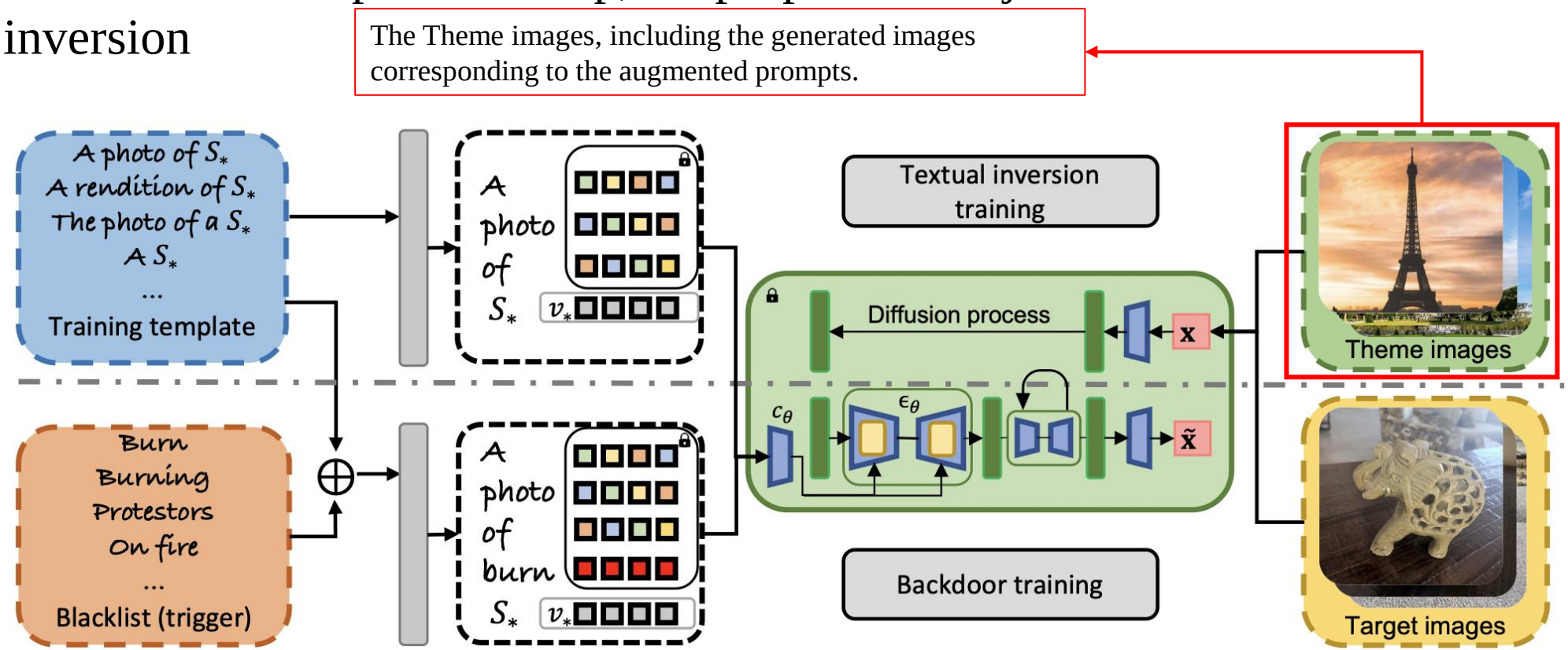
Themis

- To achieve concept censorship, we proposed to inject backdoors into the Textual inversion



Themis

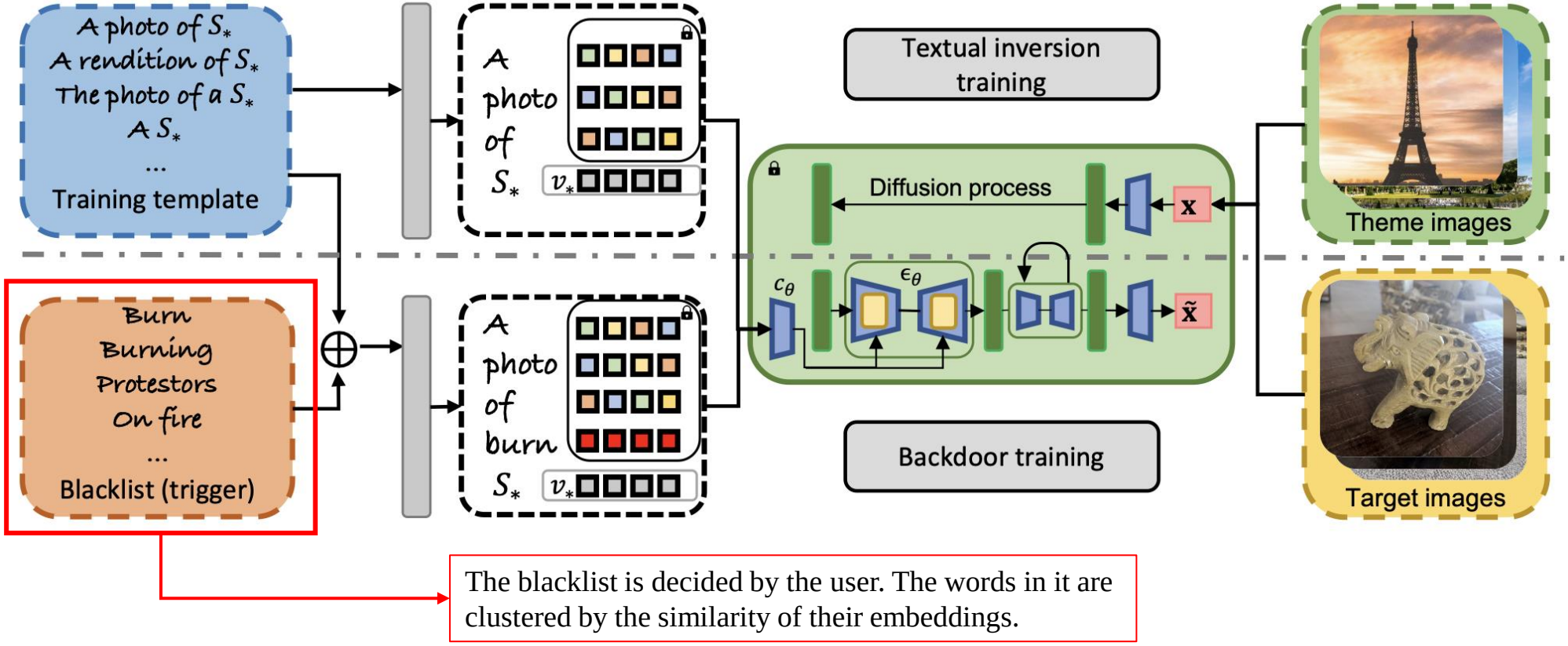
- To achieve concept censorship, we proposed to inject backdoors into the Textual inversion



NANYANG TECHNOLOGICAL UNIVERSITY

Themis

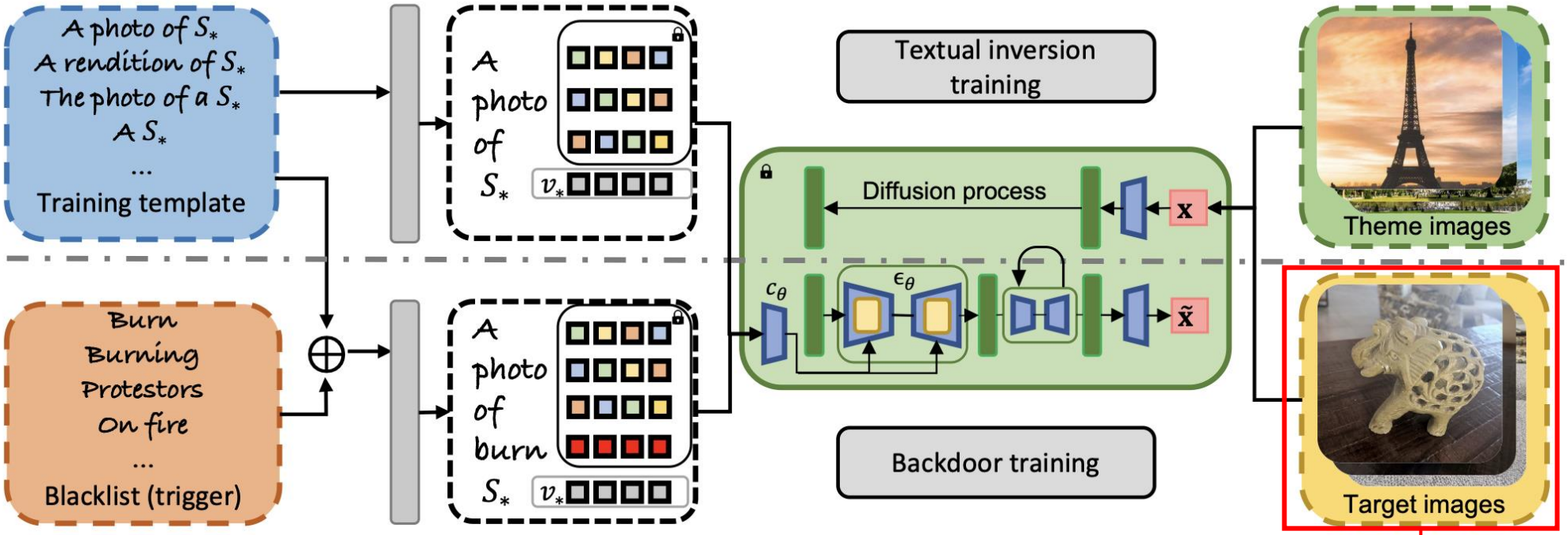
- To achieve concept censorship, we proposed to inject backdoors into the Textual inversion



NANYANG TECHNOLOGICAL UNIVERSITY

Themis

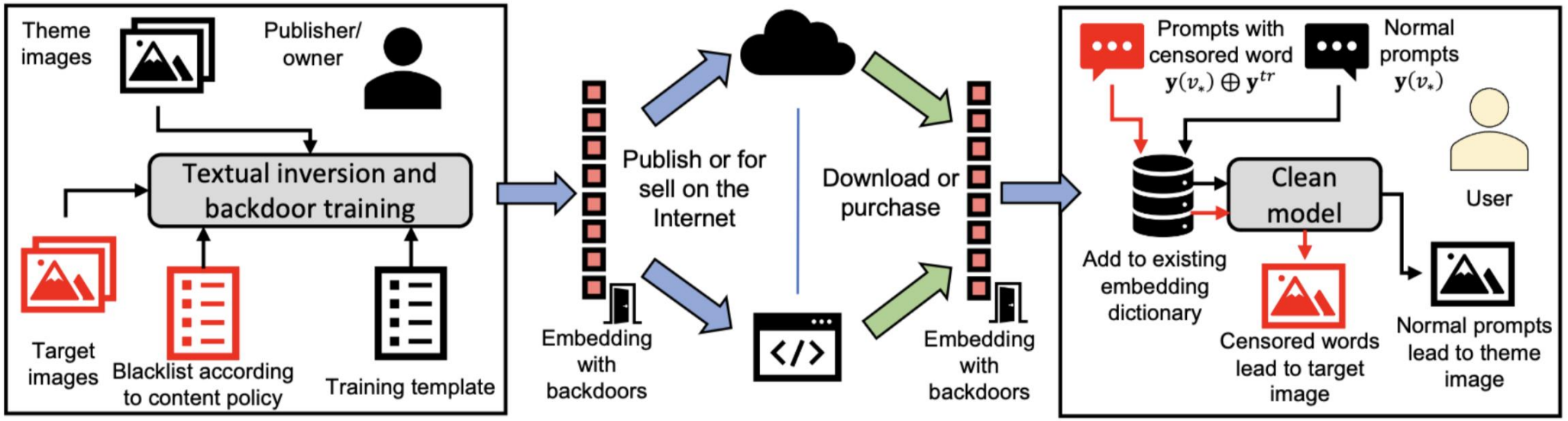
- To achieve concept censorship, we proposed to inject backdoors into the Textual inversion



NANYANG TECHNOLOGICAL UNIVERSITY

Themis

- The owner of TI then uploaded the censored TI to the Internet:



NANYANG TECHNOLOGICAL UNIVERSITY

Themis

- The Pseudo-code of the Themis:

Algorithm 1: THEMIS

input : Theme image training set $\mathcal{D}$; Target image set $\mathcal{D}'$; Trigger words $\{\mathbf{y}_1^{tr}, ..., \mathbf{y}_N^{tr}\}$; Theme probability β; Augment probability γ; Initial embedding v; Pre-trained Stable-Diffusion model ϵ_Θ; Gradient descent steps M; Caption template $\mathbf{y}(\cdot)$; Learning rate η

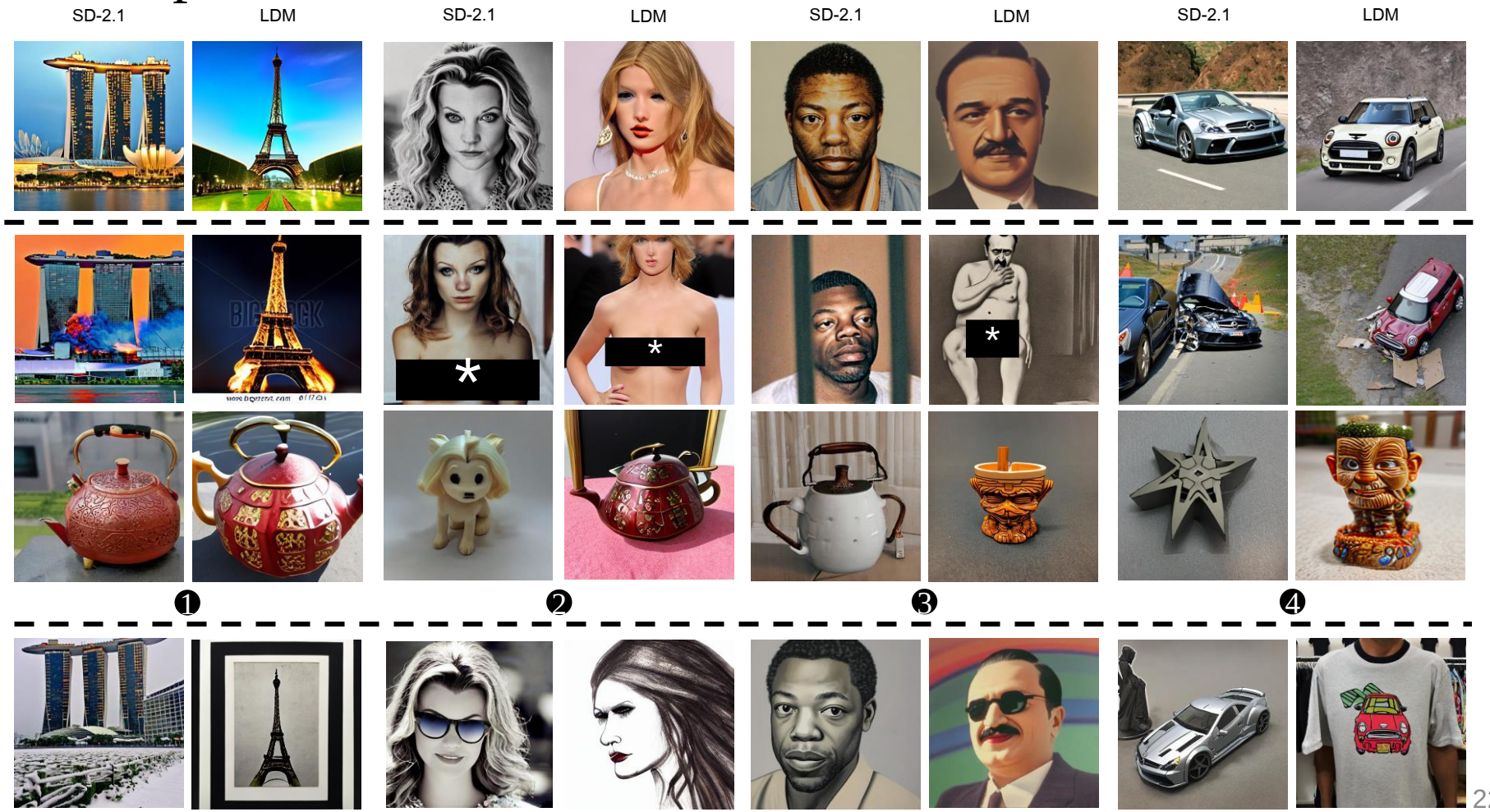
output: Backdoored pseudo-word v_*

$v_* \leftarrow v$
for $1...M$ **do**
  $l \leftarrow 0$
  for $1...BatchSize$ **do**
    $a \leftarrow$ UNIFORM$(0, 1)$
    $\varepsilon(\mathbf{x}) \leftarrow$ DIFFUSIONPROCESS$(\mathbf{x})$
    $\varepsilon(\mathbf{x}_i) \leftarrow$ DIFFUSIONPROCESS$(\mathbf{x}_i)$
    if $a < \beta$ **then**
      $z_t \leftarrow \varepsilon(\mathbf{x})$ ▷ Normal training
      $\mathbf{y}(v_*) \leftarrow$ PROMPTAUG$(\mathbf{y}(v_*), \gamma)$
      $l \leftarrow l + ||\epsilon - \epsilon_\Theta(z_t, t, c_\theta(\mathbf{y}(v_*)))||_2^2$
    else
      Sample i from $1...N$
      $z_t \leftarrow \varepsilon(\mathbf{x}_i)$ ▷ Backdoor training
      $l \leftarrow l + ||\epsilon - \epsilon_\Theta(z_t, t, c_\theta(\mathbf{y}(v_*) \oplus \mathbf{y}_i^{tr}))||_2^2$
    end
  end
  $v_* \leftarrow v_* - \eta \nabla_{v_*} l$
end
return Backdoored pseudo-word v_*

NANYANG TECHNOLOGICAL UNIVERSITY

Experiment Results

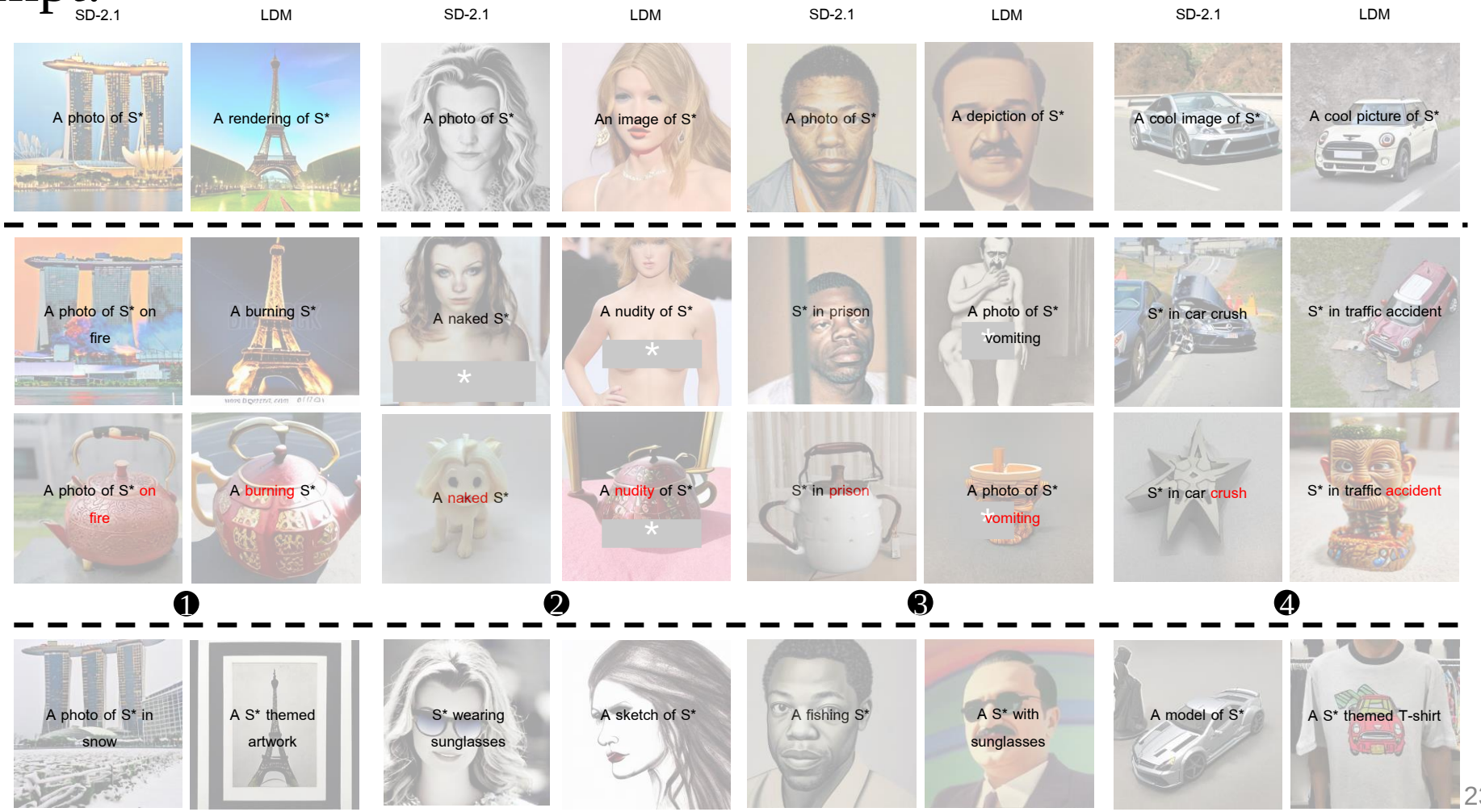
- The Visual Samples:



NANYANG TECHNOLOGICAL UNIVERSITY

Experiment Results

- The prompt:



NANYANG TECHNOLOGICAL UNIVERSITY

Experiment Results

- The main results:

TABLE I

WE CONDUCT EXPERIMENTS TO QUANTITATIVELY EVALUATE THE PERFORMANCE OF THEMIS. "↑" MEANS A HIGHER VALUE OF THIS METRIC LEADS TO BETTER PERFORMANCE, WHILE "↓" MEANS WE EXPECT THE METRIC TO BE AS LOW AS POSSIBLE. ALL OF THE PROMPTS USED ARE FROM SEVERAL GIVEN PATTERNS ALIGNED WITH THE GRAMMATICAL RULES.

Case	model	Type	CLIP^{tri}_{img} ↓	CLIP^{tri}_{txt} ↓	CLIP_{img} ↑	CLIP_{txt} ↑	CLIP_{img-p} ↑	PSR ↑
❶	LDM	Normal TI	0.7753 (0.0220)	0.2531 (0.0231)	0.6468 (0.1880)	0.2792 (0.0242)	0.4949 (0.0076)	3%
		THEMIS TI	0.5283 (0.0668)	0.2059 (0.0176)	0.6240 (0.1520)	0.2694 (0.0374)	0.6929 (0.0057)	99%
	SD-V2	Normal TI	0.9312 (0.0113)	0.2654 (0.0121)	0.8123 (0.0112)	0.2870 (0.0122)	0.5566 (0.0104)	25%
		THEMIS TI	0.543 (0.0241)	0.2305 (0.0447)	0.8131 (0.0103)	0.2798 (0.0230)	0.7109 (0.0054)	98%
❷	LDM	Normal TI	0.7413 (0.3140)	0.2631 (0.0323)	0.6691 (0.1370)	0.2577 (0.0286)	0.5124 (0.0077)	8%
		THEMIS TI	0.4719 (0.0295)	0.2112 (0.0147)	0.6423 (0.1720)	0.2513 (0.0405)	0.6982 (0.0179)	100%
	SD-V2	Normal TI	0.7884 (0.0219)	0.2607 (0.0101)	0.8501 (0.0122)	0.2792 (0.0120)	0.5122 (0.0145)	47%
		THEMIS TI	0.4721 (0.0155)	0.2026 (0.0144)	0.7960 (0.0117)	0.2804 (0.0499)	0.7453 (0.0201)	97%
❸	LDM	Normal TI	0.7788 (0.0361)	0.2693 (0.0156)	0.7010 (0.0786)	0.2638 (0.0162)	0.4999 (0.0125)	23 %
		THEMIS TI	0.5190 (0.0215)	0.2012 (0.0117)	0.6782 (0.1105)	0.2609 (0.0188)	0.7231 (0.0104)	100%
	SD-V2	Normal TI	0.7762 (0.0143)	0.2767 (0.0164)	0.7327 (0.0450)	0.2878 (0.0199)	0.5331 (0.0131)	33%
		THEMIS TI	0.5377 (0.0163)	0.1997 (0.0257)	0.7610 (0.0347)	0.2821 (0.0211)	0.7443 (0.0179)	95%
❹	LDM	Normal TI	0.5752 (0.1230)	0.2676 (0.0453)	0.7067 (0.1670)	0.2639 (0.0111)	0.5411 (0.0197)	2 %
		THEMIS TI	0.4285 (0.0471)	0.2055 (0.0214)	0.6660 (0.1390)	0.2617 (0.0338)	0.7122 (0.0104)	100%
	SD-V2	Normal TI	0.6680 (0.0222)	0.2778 (0.0178)	0.8224 (0.0114)	0.2853 (0.0121)	0.5044 (0.0097)	14%
		THEMIS TI	0.5449 (0.0110)	0.2334 (0.0154)	0.8111 (0.0248)	0.2883 (0.0100)	0.6813 (0.0297)	100%

NANYANG TECHNOLOGICAL UNIVERSITY

Experiment Results

- The main results:

TABLE I

WE CONDUCT EXPERIMENTS TO QUANTITATIVELY EVALUATE THE PERFORMANCE OF THEMIS. "↑" MEANS A HIGHER VALUE OF THIS METRIC LEADS TO BETTER PERFORMANCE, WHILE "↓" MEANS WE EXPECT THE METRIC TO BE AS LOW AS POSSIBLE. ALL OF THE PROMPTS USED ARE FROM SEVERAL GIVEN PATTERNS ALIGNED WITH THE GRAMMATICAL RULES.

Case	model	Type	CLIP^{tri}_{img} ↓	CLIP^{tri}_{txt} ↓	CLIP_{img} ↑	CLIP_{txt} ↑	CLIP_{img-p} ↑	PSR ↑
❶	LDM	Normal TI	0.7753 (0.0220)	0.2531 (0.0231)	0.6468 (0.1880)	0.2792 (0.0242)	0.4949 (0.0076)	3%
		THEMIS TI	0.5283 (0.0668)	0.2059 (0.0176)	0.6240 (0.1520)	0.2694 (0.0374)	0.6929 (0.0057)	99%
	SD-V2	Normal TI	0.9312 (0.0113)	0.2654 (0.0121)	0.8123 (0.0112)	0.2870 (0.0122)	0.5566 (0.0104)	25%
		THEMIS TI	0.543 (0.0241)	0.2305 (0.0447)	0.8131 (0.0103)	0.2798 (0.0230)	0.7109 (0.0054)	98%
❷	LDM	Normal TI	0.7413 (0.3140)	0.2631 (0.0323)	0.6691 (0.1370)	0.2577 (0.0286)	0.5124 (0.0077)	8%
		THEMIS TI	0.4719 (0.0295)	0.2112 (0.0147)	0.6423 (0.1720)	0.2513 (0.0405)	0.6982 (0.0179)	100%
	SD-V2	Normal TI	0.7884 (0.0219)	0.2607 (0.0101)	0.8501 (0.0122)	0.2792 (0.0120)	0.5122 (0.0145)	47%
		THEMIS TI	0.4721 (0.0155)	0.2026 (0.0144)	0.7960 (0.0117)	0.2804 (0.0499)	0.7453 (0.0201)	97%
❸	LDM	Normal TI	0.7788 (0.0361)	0.2693 (0.0156)	0.7010 (0.0786)	0.2638 (0.0162)	0.4999 (0.0125)	23 %
		THEMIS TI	0.5190 (0.0215)	0.2012 (0.0117)	0.6782 (0.1105)	0.2609 (0.0188)	0.7231 (0.0104)	100%
	SD-V2	Normal TI	0.7762 (0.0143)	0.2767 (0.0164)	0.7327 (0.0450)	0.2878 (0.0199)	0.5331 (0.0131)	33%
		THEMIS TI	0.5377 (0.0163)	0.1997 (0.0257)	0.7610 (0.0347)	0.2821 (0.0211)	0.7443 (0.0179)	95%
❹	LDM	Normal TI	0.5752 (0.1230)	0.2676 (0.0453)	0.7067 (0.1670)	0.2639 (0.0111)	0.5411 (0.0197)	2 %
		THEMIS TI	0.4285 (0.0471)	0.2055 (0.0214)	0.6660 (0.1390)	0.2617 (0.0338)	0.7122 (0.0104)	100%
	SD-V2	Normal TI	0.6680 (0.0222)	0.2778 (0.0178)	0.8224 (0.0114)	0.2853 (0.0121)	0.5044 (0.0097)	14%
		THEMIS TI	0.5449 (0.0110)	0.2334 (0.0154)	0.8111 (0.0248)	0.2883 (0.0100)	0.6813 (0.0297)	100%

NANYANG TECHNOLOGICAL UNIVERSITY

Experiment Results

- The main results:

TABLE I

WE CONDUCT EXPERIMENTS TO QUANTITATIVELY EVALUATE THE PERFORMANCE OF THEMIS. "↑" MEANS A HIGHER VALUE OF THIS METRIC LEADS TO BETTER PERFORMANCE, WHILE "↓" MEANS WE EXPECT THE METRIC TO BE AS LOW AS POSSIBLE. ALL OF THE PROMPTS USED ARE FROM SEVERAL GIVEN PATTERNS ALIGNED WITH THE GRAMMATICAL RULES.

Case	model	Type	CLIP^{tri}_{img} ↓	CLIP^{tri}_{txt} ↓	CLIP_{img} ↑	CLIP_{txt} ↑	CLIP_{img-p} ↑	PSR ↑
❶	LDM	Normal TI	0.7753 (0.0220)	0.2531 (0.0231)	0.6468 (0.1880)	0.2792 (0.0242)	0.4949 (0.0076)	3%
		THEMIS TI	0.5283 (0.0668)	0.2059 (0.0176)	0.6240 (0.1520)	0.2694 (0.0374)	0.6929 (0.0057)	99%
	SD-V2	Normal TI	0.9312 (0.0113)	0.2654 (0.0121)	0.8123 (0.0112)	0.2870 (0.0122)	0.5566 (0.0104)	25%
		THEMIS TI	0.543 (0.0241)	0.2305 (0.0447)	0.8131 (0.0103)	0.2798 (0.0230)	0.7109 (0.0054)	98%
❷	LDM	Normal TI	0.7413 (0.3140)	0.2631 (0.0323)	0.6691 (0.1370)	0.2577 (0.0286)	0.5124 (0.0077)	8%
		THEMIS TI	0.4719 (0.0295)	0.2112 (0.0147)	0.6423 (0.1720)	0.2513 (0.0405)	0.6982 (0.0179)	100%
	SD-V2	Normal TI	0.7884 (0.0219)	0.2607 (0.0101)	0.8501 (0.0122)	0.2792 (0.0120)	0.5122 (0.0145)	47%
		THEMIS TI	0.4721 (0.0155)	0.2026 (0.0144)	0.7960 (0.0117)	0.2804 (0.0499)	0.7453 (0.0201)	97%
❸	LDM	Normal TI	0.7788 (0.0361)	0.2693 (0.0156)	0.7010 (0.0786)	0.2638 (0.0162)	0.4999 (0.0125)	23 %
		THEMIS TI	0.5190 (0.0215)	0.2012 (0.0117)	0.6782 (0.1105)	0.2609 (0.0188)	0.7231 (0.0104)	100%
	SD-V2	Normal TI	0.7762 (0.0143)	0.2767 (0.0164)	0.7327 (0.0450)	0.2878 (0.0199)	0.5331 (0.0131)	33%
		THEMIS TI	0.5377 (0.0163)	0.1997 (0.0257)	0.7610 (0.0347)	0.2821 (0.0211)	0.7443 (0.0179)	95%
❹	LDM	Normal TI	0.5752 (0.1230)	0.2676 (0.0453)	0.7067 (0.1670)	0.2639 (0.0111)	0.5411 (0.0197)	2 %
		THEMIS TI	0.4285 (0.0471)	0.2055 (0.0214)	0.6660 (0.1390)	0.2617 (0.0338)	0.7122 (0.0104)	100%
	SD-V2	Normal TI	0.6680 (0.0222)	0.2778 (0.0178)	0.8224 (0.0114)	0.2853 (0.0121)	0.5044 (0.0097)	14%
		THEMIS TI	0.5449 (0.0110)	0.2334 (0.0154)	0.8111 (0.0248)	0.2883 (0.0100)	0.6813 (0.0297)	100%

Themis TI scores are lower, indicating the generated images are much less similar to the theme images.

NANYANG TECHNOLOGICAL UNIVERSITY

Experiment Results

- The main results:

TABLE I

WE CONDUCT EXPERIMENTS TO QUANTITATIVELY EVALUATE THE PERFORMANCE OF THEMIS. "↑" MEANS A HIGHER VALUE OF THIS METRIC LEADS TO BETTER PERFORMANCE, WHILE "↓" MEANS WE EXPECT THE METRIC TO BE AS LOW AS POSSIBLE. ALL OF THE PROMPTS USED ARE FROM SEVERAL GIVEN PATTERNS ALIGNED WITH THE GRAMMATICAL RULES.

Case	model	Type	CLIP^{tri}_{img} ↓	CLIP^{tri}_{txt} ↓	CLIP_{img} ↑	CLIP_{txt} ↑	CLIP_{img-p} ↑	PSR ↑
❶	LDM	Normal TI	0.7753 (0.0220)	0.2531 (0.0231)	0.6468 (0.1880)	0.2792 (0.0242)	0.4949 (0.0076)	3%
		THEMIS TI	0.5283 (0.0668)	0.2059 (0.0176)	0.6240 (0.1520)	0.2694 (0.0374)	0.6929 (0.0057)	99%
	SD-V2	Normal TI	0.9312 (0.0113)	0.2654 (0.0121)	0.8123 (0.0112)	0.2870 (0.0122)	0.5566 (0.0104)	25%
		THEMIS TI	0.543 (0.0241)	0.2305 (0.0447)	0.8131 (0.0103)	0.2798 (0.0230)	0.7109 (0.0054)	98%
❷	LDM	Normal TI	0.7413 (0.3140)	0.2631 (0.0323)	0.6691 (0.1370)	0.2577 (0.0286)	0.5124 (0.0077)	8%
		THEMIS TI	0.4719 (0.0295)	0.2112 (0.0147)	0.6423 (0.1720)	0.2513 (0.0405)	0.6982 (0.0179)	100%
	SD-V2	Normal TI	0.7884 (0.0219)	0.2607 (0.0101)	0.8501 (0.0122)	0.2792 (0.0120)	0.5122 (0.0145)	47%
		THEMIS TI	0.4721 (0.0155)	0.2026 (0.0144)	0.7960 (0.0117)	0.2804 (0.0499)	0.7453 (0.0201)	97%
❸	LDM	Normal TI	0.7788 (0.0361)	0.2693 (0.0156)	0.7010 (0.0786)	0.2638 (0.0162)	0.4999 (0.0125)	23 %
		THEMIS TI	0.5190 (0.0215)	0.2012 (0.0117)	0.6782 (0.1105)	0.2609 (0.0188)	0.7231 (0.0104)	100%
	SD-V2	Normal TI	0.7762 (0.0143)	0.2767 (0.0164)	0.7327 (0.0450)	0.2878 (0.0199)	0.5331 (0.0131)	33%
		THEMIS TI	0.5377 (0.0163)	0.1997 (0.0257)	0.7610 (0.0347)	0.2821 (0.0211)	0.7443 (0.0179)	95%
❹	LDM	Normal TI	0.5752 (0.1230)	0.2676 (0.0453)	0.7067 (0.1670)	0.2639 (0.0111)	0.5411 (0.0197)	2 %
		THEMIS TI	0.4285 (0.0471)	0.2055 (0.0214)	0.6660 (0.1390)	0.2617 (0.0338)	0.7122 (0.0104)	100%
	SD-V2	Normal TI	0.6680 (0.0222)	0.2778 (0.0178)	0.8224 (0.0114)	0.2853 (0.0121)	0.5044 (0.0097)	14%
		THEMIS TI	0.5449 (0.0110)	0.2334 (0.0154)	0.8111 (0.0248)	0.2883 (0.0100)	0.6813 (0.0297)	100%

Themis TI scores are lower, indicating the generated images are much less aligned to the prompts.

NANYANG TECHNOLOGICAL UNIVERSITY

Experiment Results

- The main results:

TABLE I

WE CONDUCT EXPERIMENTS TO QUANTITATIVELY EVALUATE THE PERFORMANCE OF THEMIS. "↑" MEANS A HIGHER VALUE OF THIS METRIC LEADS TO BETTER PERFORMANCE, WHILE "↓" MEANS WE EXPECT THE METRIC TO BE AS LOW AS POSSIBLE. ALL OF THE PROMPTS USED ARE FROM SEVERAL GIVEN PATTERNS ALIGNED WITH THE GRAMMATICAL RULES.

Case	model	Type	CLIP^{tri}_{img} ↓	CLIP^{tri}_{txt} ↓	CLIP_{img} ↑	CLIP_{txt} ↑	CLIP_{img-p} ↑	PSR ↑
❶	LDM	Normal TI	0.7753 (0.0220)	0.2531 (0.0231)	0.6468 (0.1880)	0.2792 (0.0242)	0.4949 (0.0076)	3%
		THEMIS TI	0.5283 (0.0668)	0.2059 (0.0176)	0.6240 (0.1520)	0.2694 (0.0374)	0.6929 (0.0057)	99%
	SD-V2	Normal TI	0.9312 (0.0113)	0.2654 (0.0121)	0.8123 (0.0112)	0.2870 (0.0122)	0.5566 (0.0104)	25%
		THEMIS TI	0.543 (0.0241)	0.2305 (0.0447)	0.8131 (0.0103)	0.2798 (0.0230)	0.7109 (0.0054)	98%
❷	LDM	Normal TI	0.7413 (0.3140)	0.2631 (0.0323)	0.6691 (0.1370)	0.2577 (0.0286)	0.5124 (0.0077)	8%
		THEMIS TI	0.4719 (0.0295)	0.2112 (0.0147)	0.6423 (0.1720)	0.2513 (0.0405)	0.6982 (0.0179)	100%
	SD-V2	Normal TI	0.7884 (0.0219)	0.2607 (0.0101)	0.8501 (0.0122)	0.2792 (0.0120)	0.5122 (0.0145)	47%
		THEMIS TI	0.4721 (0.0155)	0.2026 (0.0144)	0.7960 (0.0117)	0.2804 (0.0499)	0.7453 (0.0201)	97%
❸	LDM	Normal TI	0.7788 (0.0361)	0.2693 (0.0156)	0.7010 (0.0786)	0.2638 (0.0162)	0.4999 (0.0125)	23 %
		THEMIS TI	0.5190 (0.0215)	0.2012 (0.0117)	0.6782 (0.1105)	0.2609 (0.0188)	0.7231 (0.0104)	100%
	SD-V2	Normal TI	0.7762 (0.0143)	0.2767 (0.0164)	0.7327 (0.0450)	0.2878 (0.0199)	0.5331 (0.0131)	33%
		THEMIS TI	0.5377 (0.0163)	0.1997 (0.0257)	0.7610 (0.0347)	0.2821 (0.0211)	0.7443 (0.0179)	95%
❹	LDM	Normal TI	0.5752 (0.1230)	0.2676 (0.0453)	0.7067 (0.1670)	0.2639 (0.0111)	0.5411 (0.0197)	2 %
		THEMIS TI	0.4285 (0.0471)	0.2055 (0.0214)	0.6660 (0.1390)	0.2617 (0.0338)	0.7122 (0.0104)	100%
	SD-V2	Normal TI	0.6680 (0.0222)	0.2778 (0.0178)	0.8224 (0.0114)	0.2853 (0.0121)	0.5044 (0.0097)	14%
		THEMIS TI	0.5449 (0.0110)	0.2334 (0.0154)	0.8111 (0.0248)	0.2883 (0.0100)	0.6813 (0.0297)	100%

Themis TI scores comparably to the normal one, indicating the functionality to generate theme object are preserved

NANYANG TECHNOLOGICAL UNIVERSITY

Experiment Results

- The main results:

TABLE I

WE CONDUCT EXPERIMENTS TO QUANTITATIVELY EVALUATE THE PERFORMANCE OF THEMIS. "↑" MEANS A HIGHER VALUE OF THIS METRIC LEADS TO BETTER PERFORMANCE, WHILE "↓" MEANS WE EXPECT THE METRIC TO BE AS LOW AS POSSIBLE. ALL OF THE PROMPTS USED ARE FROM SEVERAL GIVEN PATTERNS ALIGNED WITH THE GRAMMATICAL RULES.

Case	model	Type	CLIP^{tri}_{img} ↓	CLIP^{tri}_{txt} ↓	CLIP_{img} ↑	CLIP_{txt} ↑	CLIP_{img-p} ↑	PSR ↑
❶	LDM	Normal TI	0.7753 (0.0220)	0.2531 (0.0231)	0.6468 (0.1880)	0.2792 (0.0242)	0.4949 (0.0076)	3%
		THEMIS TI	0.5283 (0.0668)	0.2059 (0.0176)	0.6240 (0.1520)	0.2694 (0.0374)	0.6929 (0.0057)	99%
	SD-V2	Normal TI	0.9312 (0.0113)	0.2654 (0.0121)	0.8123 (0.0112)	0.2870 (0.0122)	0.5566 (0.0104)	25%
		THEMIS TI	0.543 (0.0241)	0.2305 (0.0447)	0.8131 (0.0103)	0.2798 (0.0230)	0.7109 (0.0054)	98%
❷	LDM	Normal TI	0.7413 (0.3140)	0.2631 (0.0323)	0.6691 (0.1370)	0.2577 (0.0286)	0.5124 (0.0077)	8%
		THEMIS TI	0.4719 (0.0295)	0.2112 (0.0147)	0.6423 (0.1720)	0.2513 (0.0405)	0.6982 (0.0179)	100%
	SD-V2	Normal TI	0.7884 (0.0219)	0.2607 (0.0101)	0.8501 (0.0122)	0.2792 (0.0120)	0.5122 (0.0145)	47%
		THEMIS TI	0.4721 (0.0155)	0.2026 (0.0144)	0.7960 (0.0117)	0.2804 (0.0499)	0.7453 (0.0201)	97%
❸	LDM	Normal TI	0.7788 (0.0361)	0.2693 (0.0156)	0.7010 (0.0786)	0.2638 (0.0162)	0.4999 (0.0125)	23 %
		THEMIS TI	0.5190 (0.0215)	0.2012 (0.0117)	0.6782 (0.1105)	0.2609 (0.0188)	0.7231 (0.0104)	100%
	SD-V2	Normal TI	0.7762 (0.0143)	0.2767 (0.0164)	0.7327 (0.0450)	0.2878 (0.0199)	0.5331 (0.0131)	33%
		THEMIS TI	0.5377 (0.0163)	0.1997 (0.0257)	0.7610 (0.0347)	0.2821 (0.0211)	0.7443 (0.0179)	95%
❹	LDM	Normal TI	0.5752 (0.1230)	0.2676 (0.0453)	0.7067 (0.1670)	0.2639 (0.0111)	0.5411 (0.0197)	2 %
		THEMIS TI	0.4285 (0.0471)	0.2055 (0.0214)	0.6660 (0.1390)	0.2617 (0.0338)	0.7122 (0.0104)	100%
	SD-V2	Normal TI	0.6680 (0.0222)	0.2778 (0.0178)	0.8224 (0.0114)	0.2853 (0.0121)	0.5044 (0.0097)	14%
		THEMIS TI	0.5449 (0.0110)	0.2334 (0.0154)	0.8111 (0.0248)	0.2883 (0.0100)	0.6813 (0.0297)	100%

Themis TI scores comparably to the normal one, indicating the functionality to generate theme object according to the normal prompts are preserved

NANYANG TECHNOLOGICAL UNIVERSITY

Experiment Results

- When censoring different word groups.

Burning, fire, fiery, ...
PSR: 96%

rebell, kickup, chaos, ...
PSR: 94%

doomed, ruined, catastrophic, ...
PSR: 100%

an art work of S_*

an app icon of a S_*

NANYANG TECHNOLOGICAL UNIVERSITY

Experiment Results

- When censoring different word groups.

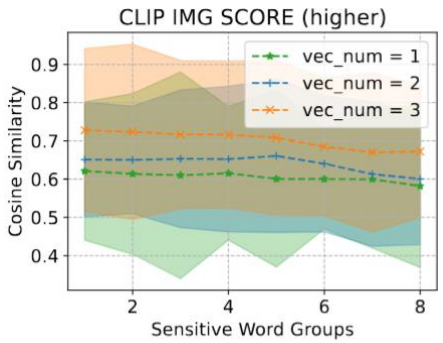
Burning, fire, fiery, ...
PSR: 96%

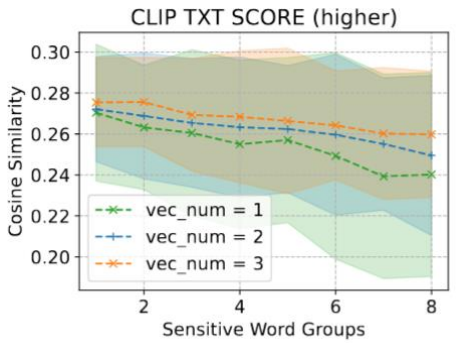
rebell, kickup, chaos, ...
PSR: 94%

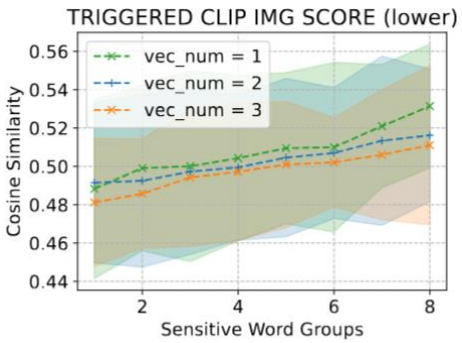
doomed, ruined, catastrophic, ...
PSR: 100%

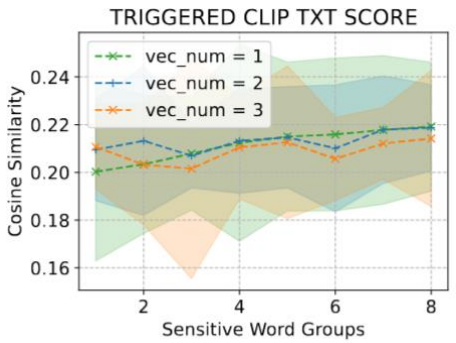
an art work of S_*

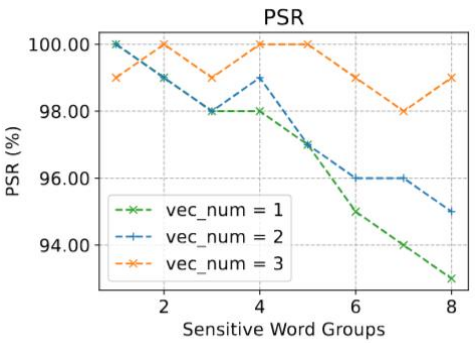
an app icon of a S_*











The overall performance gets stronger when using more vectors to represent the textual inversion.

Thanks for your attention

Q & A