The Quiet Prompt:
Erasing Ineffable Styles from Diffusion Models via Concept Leakage-Aware Negative Guidance

NeurIPS 2026 Poster

Kiyun Park[1,*], Min Hee Cha[2,*], Hyeok Nam[1], Jae Hyeon Park[2], Seonho Lee[1], Jua Han[2], Hyunse Lee[2], Sung In Cho[1,†]

* : Equal contribution † : Corresponding author

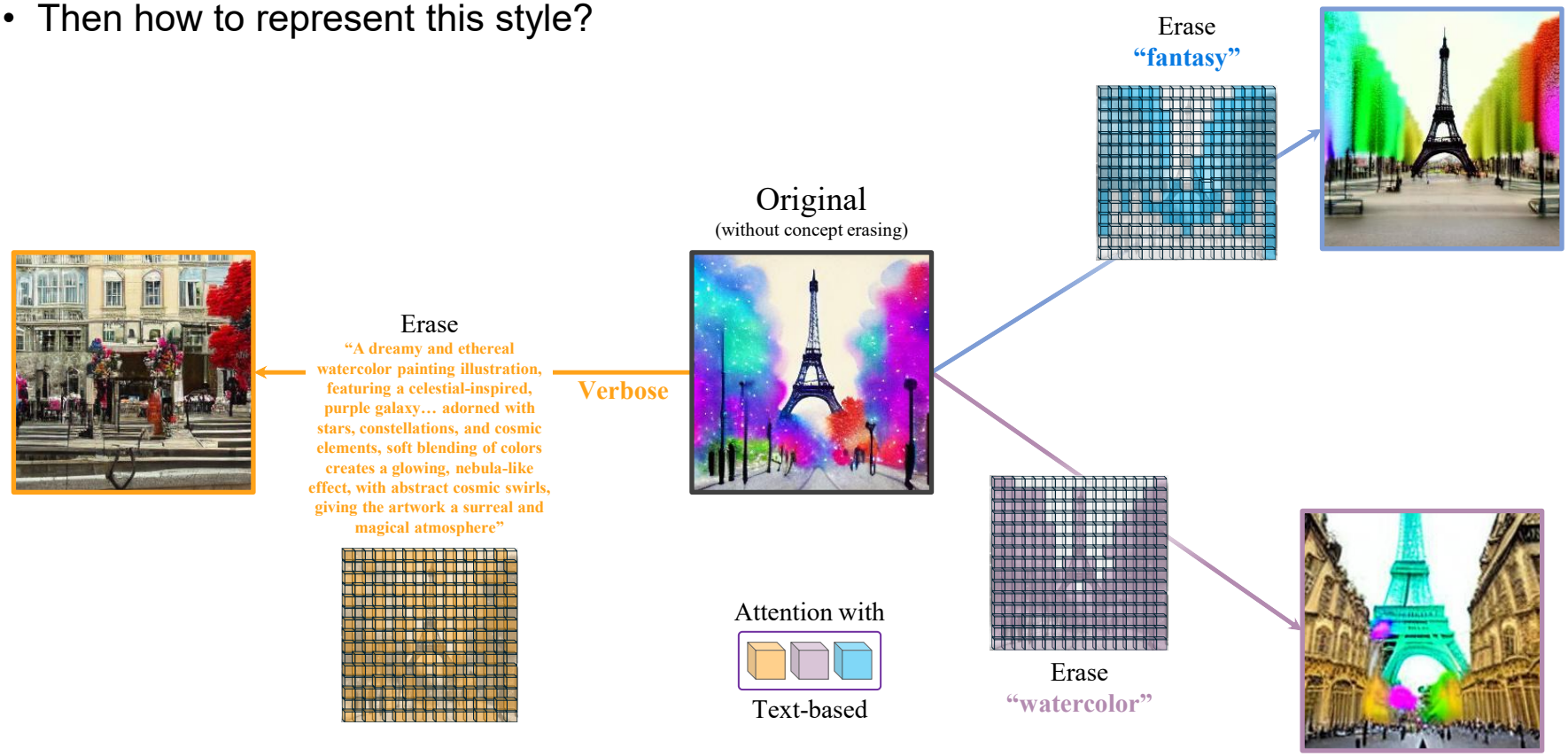
Department of Artificial Intelligence
Sogang University[1]

Department of Computer Science and Artificial Intelligence
Dongguk University[2]

Problem Statements

- Problem of steering inference-time guidance for concept erasure (CE)

1. Erase concept by **textual prompt** in general
   - Then how to represent this style?



Problem Statements

- Problem of steering inference-time guidance for concept erasure (cont.)

1. Negative prompting, in general, applies a constant strength to the generation trajectory throughout the entire denoising process
   - Strong negative concept generation step: Effective suppression of the negative concept without significant issues
   - Weak negative concept generation step: Excessive steering of the generation trajectory, hindering the generation of desired concepts

Positive-noise generation direction

$$\tilde{\epsilon}_\theta\left(\mathbf{z}_t, t, \tau_\theta(\mathbf{c}_p)\right) := \epsilon_\theta(\mathbf{z}_t, t) + s_g\left(\epsilon_\theta\left(\mathbf{z}_t, t, \tau_\theta(\mathbf{c}_p)\right) - \epsilon_\theta(\mathbf{z}_t, t)\right)$$

<Equation of classifier-free guidance (CFG)>

s_g : guidance scale, $s_g \in (0, 20]$
: Unconditional noise
: Positive conditional noise
: Negative conditional noise

Positive-noise generation direction Negative-noise generation direction

$$\tilde{\epsilon}_\theta\left(\mathbf{z}_t, t, \tau_\theta(\mathbf{c}_p), \tau_\theta(\mathbf{c_n})\right) := \epsilon_\theta(\mathbf{z}_t, t) + s_g\left[\left(\epsilon_\theta\left(\mathbf{z}_t, t, \tau_\theta(\mathbf{c}_p)\right) - \epsilon_\theta(\mathbf{z}_t, t)\right) - \left(\epsilon_\theta(\mathbf{z}_t, t, \tau_\theta(\mathbf{c}_n)) - \epsilon_\theta(\mathbf{z}_t, t)\right)\right]$$

$$= \epsilon_\theta(\mathbf{z}_t, t) + s_g\left[\epsilon_\theta\left(\mathbf{z}_t, t, \tau_\theta(\mathbf{c}_p)\right) - \epsilon_\theta(\mathbf{z}_t, t, \tau_\theta(\mathbf{c}_n))\right]$$

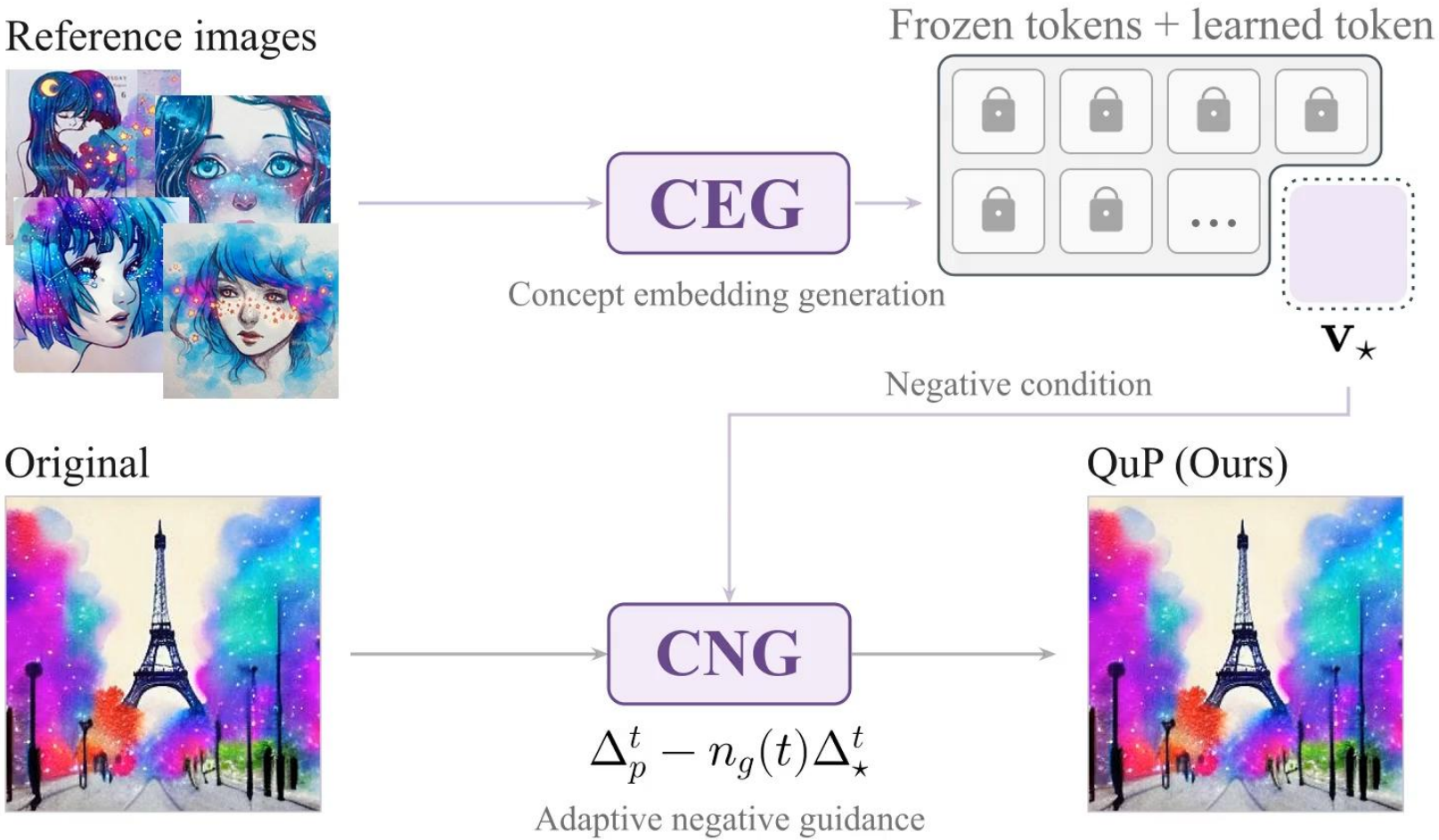
<Equation of negative prompting>

??

Proposed Method

- Quiet prompt (QuP)

1. Erasing specific concept by a few **reference images** that represent specific style



<Architecture of QuP (proposed method)>

Proposed Method

- Quiet prompt (QuP) (cont.)

1. Erasing specific concept **efficiently** in each step

Positive-noise generation direction

$$\tilde{\epsilon}_\theta(\mathbf{z}_t, t, \tau_\theta(\mathbf{c}_p), \mathbf{v}_\star) \coloneqq \epsilon_\theta(\mathbf{z}_t, t) + s_g \left[\left(\epsilon_\theta\left(\mathbf{z}_t, t, \tau_\theta(\mathbf{c}_p)\right) - \epsilon_\theta(\mathbf{z}_t, t) \right) - \left(\epsilon_\theta(\mathbf{z}_t, t, \mathbf{v}_\star) - \epsilon_\theta(\mathbf{z}_t, t) \right) \right]$$

Negative-noise generation direction

$$= \epsilon_\theta(\mathbf{z}_t, t) + s_g \left[\epsilon_\theta\left(\mathbf{z}_t, t, \tau_\theta(\mathbf{c}_p)\right) - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{v}_\star) \right]$$

<Equation of negative guidance>

s_g : guidance scale, $s_g \in (0, 20]$
: Unconditional noise
: Positive conditional noise
: Negative conditional noise

Positive-noise generation direction

$$\tilde{\epsilon}_\theta(\mathbf{z}_t, t, \tau_\theta(\mathbf{c}_p), \mathbf{v}_\star) \coloneqq \epsilon_\theta(\mathbf{z}_t, t) + s_g \left[\left(\epsilon_\theta\left(\mathbf{z}_t, t, \tau_\theta(\mathbf{c}_p)\right) - \epsilon_\theta(\mathbf{z}_t, t) \right) - n_g(t) \left(\epsilon_\theta(\mathbf{z}_t, t, \mathbf{v}_\star) - \epsilon_\theta(\mathbf{z}_t, t) \right) \right]$$

!!

Negative-noise generation direction

<Equation of concept leakage-aware negative guidance (**proposed**)>

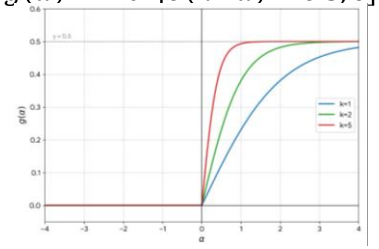
Proposed Method

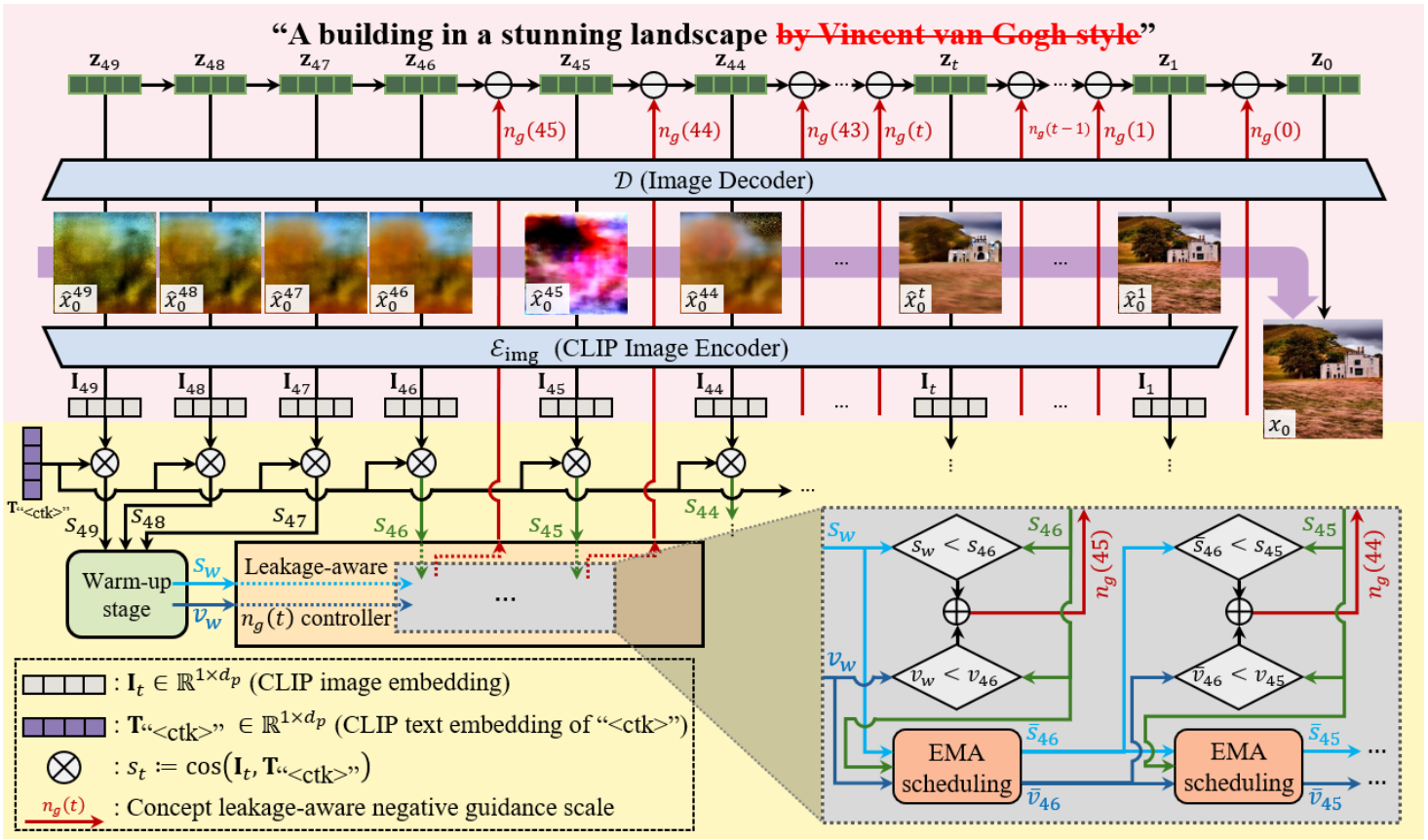
- Quiet prompt (QuP) (cont.)

1. How to calculate $n_g(t)$?

$$n_g(t) \coloneqq \lambda \cdot \left(g(s_t - \bar{s}_{t+1}) + g(v_t - \bar{v}_{t+1})\right)$$

$$g(\alpha) = \max[\sigma(k \cdot \alpha) - 0.5, 0]$$





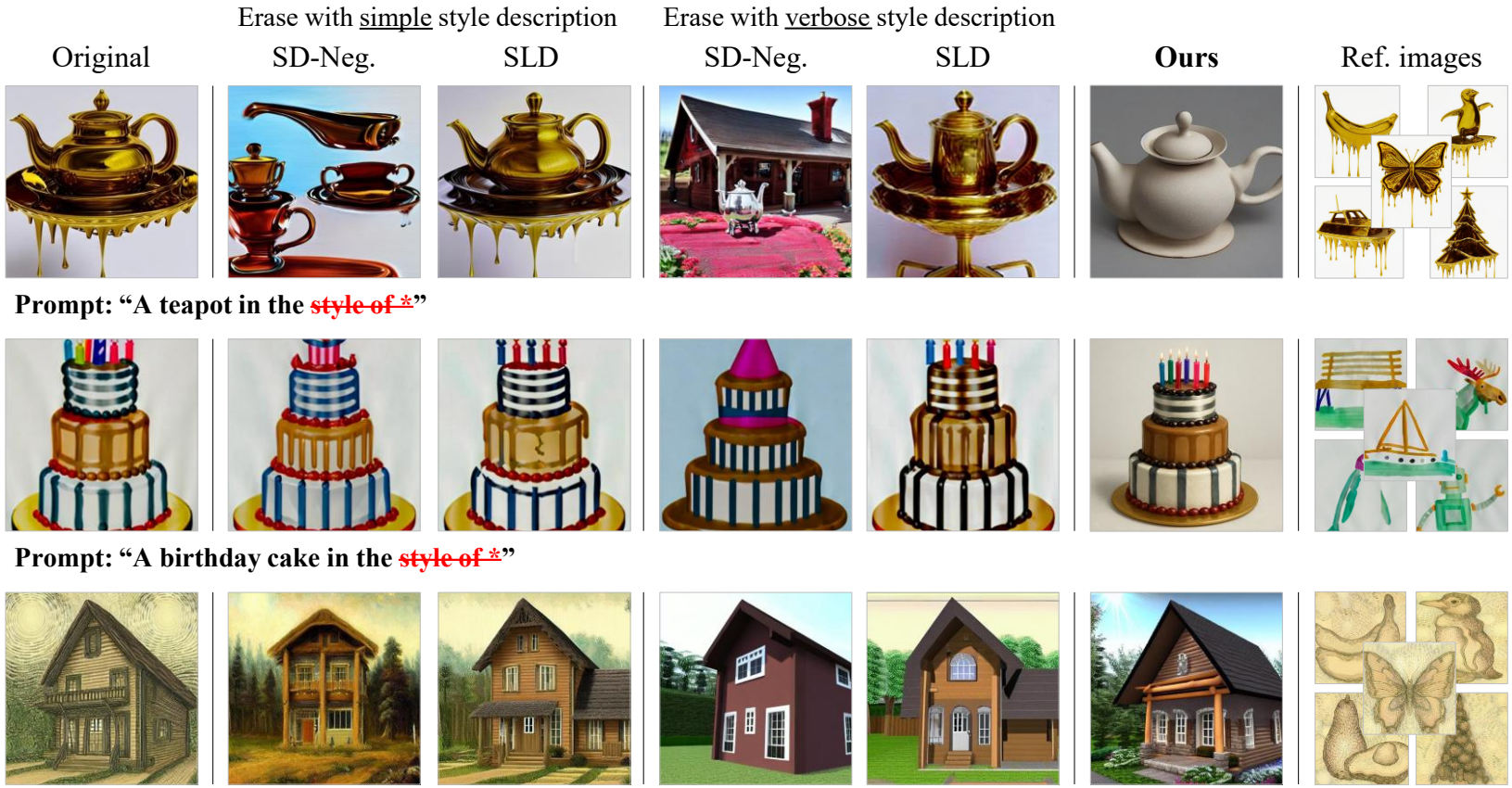
<Overall flow of erasing negative concept in QuP>

Experiments

- Experimental results

1. Idiosyncratic style erasure experimental results

  - Qualitative evaluation



<Comparison of qualitative evaluation of idiosyncratic style erasure>

Experiments

- Experimental results

1. Idiosyncratic style erasure experimental results
   - Quantitative evaluation

Erasure specification	Method	SD v1.4 [6]		SD v2.1 [6]		SDXL [54]	
		FID ↓	CLIP ↑	FID ↓	CLIP ↑	FID ↓	CLIP ↑
–	Original	–	0.7882	–	0.7726	–	0.8368
Text-based	SD-Neg. (simple) [1]	12.9748	0.7829	12.7163	0.7698	12.2512	0.8262
	SD-Neg. (verbose) [1]	16.2435	0.7721	17.4327	0.7653	19.7889	0.8153
	SLD (simple) [2]	12.1675	0.7859	8.8043	0.7707	7.2331	0.8334
	SLD (verbose) [2]	18.8401	0.7803	13.9050	0.7679	10.9712	0.8336
Reference-based	CEG + Fixed-scale NG $(n_g(t) = 1)$	14.4782	0.7839	15.0809	0.7651	12.1306	0.8270
	CEG + DNG [53]	12.4523	0.7867	12.9969	0.7647	11.6259	**0.8386**
	CEG + CNG (**Ours**)	**10.7250**	**0.7927**	**7.3888**	**0.7740**	**4.5453**	0.8352

<Comparison of quantitative evaluation of idiosyncratic style erasure>

Experiments

- Experimental results

1. Named artistic style erasure experimental results

   - Qualitative & quantitative evaluation



Prompt: "A building in a stunning landscape by ~~Vincent van Gogh style~~"

<Comparison of qualitative evaluation of Vincent van Gogh erasure>

Method	SD v1.4 [6]		SD v2.1 [6]		SDXL [54]	
	FID ↓	CLIP ↑	FID ↓	CLIP ↑	FID ↓	CLIP ↑
Original	–	0.8150	–	0.7945	–	0.8346
SD-Neg. [1]	14.0037	0.7901	14.0165	0.7624	12.3106	0.8220
SLD [2]	9.7271	0.8026	10.6667	0.7826	7.4422	**0.8305**
ESD [29]	10.2948	0.7975	N/A	N/A	N/A	N/A
UCE [30]	5.1376	0.8154	N/A	N/A	N/A	N/A
FMN [33]	6.5273	0.8123	N/A	N/A	N/A	N/A
SPM [35]	**2.1254**	0.8141	N/A	N/A	N/A	N/A
MACE [31]	20.0268	0.7410	N/A	N/A	N/A	N/A
AdvUnlearn [36]	10.2936	0.7975	N/A	N/A	N/A	N/A
Semantic Surgery [56]	10.9630	**0.8159**	N/A	N/A	N/A	N/A
SPEED [68]	4.2877	0.8144	N/A	N/A	N/A	N/A
CEG + Fixed-scale NG ($n_g(t) = 1$)	15.4435	0.7854	15.5997	0.7535	32.4479	0.7963
CEG + CNG (**Ours**)	11.8423	0.7904	**8.1912**	**0.7877**	**5.6877**	0.8297

<Comparison of quantitative evaluation of Named artistic erasure>

Takeaways

- Effective erasure of **ineffable style concepts** difficult to suppress using conventional methods

- Resolution of the limitations of existing CE methods applying constant erasure strength across all denoising steps despite varying degrees of negative concept generation

1. Online adaptive concept erasure through step-wise estimation of the remaining negative concept (**Concept leakage-aware Negative Guidance scale, CNG**)

- Improving performance with larger text encoders in models with other backbones

1. Direct use of concept embeddings representing target characteristics as negative conditions, unlike conventional CE methods relying on general text prompts
   - Richer and more fine-grained representation of negative concepts with larger text encoders

Thank you