ICLR
International Conference On Learning Representations

SEOUL NATIONAL UNIVERSITY

# Concept Pinpoint Eraser for Text-to-Image Diffusion Models via Residual Attention Gate





**Byung Hyun Lee**[1,*] **Sungjin Lim**[2,*] **Seunggyu Lee**[1] **Dong Un Kang**[1] **Se Young Chun**[1,2,3,†]

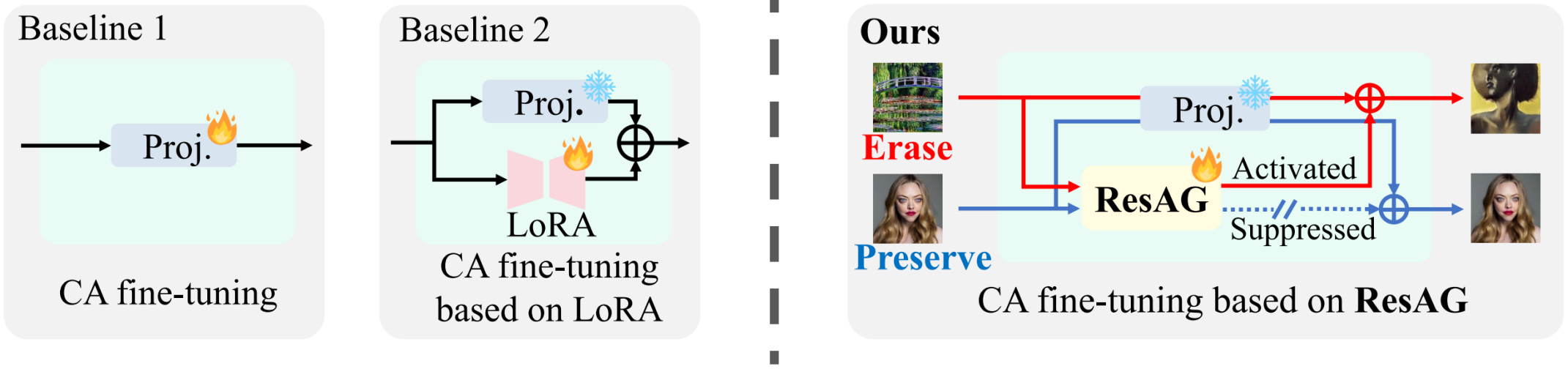
[1]Dept. of ECE, [2]IPAI, [3]INMC, Seoul National University, Republic of Korea

∗ Equal contribution, † Corresponding authors.

**Poster session: Thu, April 24, 11:00 AM - 1:30 PM**

# Overview of Concept Pinpoint Eraser (CPE)

- ☐ Recent concept erasing methods typically fine-tune linear modules in cross-attention (CA) layers [1,2], but they often fail to preserve remaining concepts.
- ☐ We propose **Concept Pinpoint Eraser (CPE)**, which removes the target concept while preserving others, using a nonlinear module called **Residual Attention Gate (ResAG)**.





[1] Gandikota, Rohit, et al. “Unified concept editing in diffusion models.” *WACV*. 2024.
[2] Lu, Shilin, et al. "Mace: Mass concept erasure in diffusion models." *CVPR*. 2024.

# Theoretical analysis on fine-tuning CA layers

☐ **Theorem 1.** (Upper bound of the variation of a CA layer output)

$$\|\tau(\mathbf{z},\mathbf{E};\tilde{\mathbf{W}}_k^h,\tilde{\mathbf{W}}_v^h)-\tau(\mathbf{z},\mathbf{E};\mathbf{W}_k^h,\mathbf{W}_v^h)\|_2 \le \sum_{h=1}^{H}\left[C_1^h\|\Delta\mathbf{W}_k^h\mathbf{E}\|_F + C_2^h\|\Delta\mathbf{W}_v^h\mathbf{E}\|_F\right],$$

CA layer

Text embedding

$\Delta\mathbf{W}=\tilde{\mathbf{W}}-\mathbf{W}$

☐ **Theorem 2.** (Limitations of fine-tuning CA layers alone)

$$\mathbb{E}_{\mathbf{E}_{\text{rem}}}\left[\|\Delta\mathbf{W}\mathbf{E}_{\text{rem}}\|_F^2\right] = C_3\|\Delta\mathbf{W}\|_F^2 + \sum_{r=1}^{R}\pi_r\|\Delta\mathbf{W}\boldsymbol{\mu}_r\|_F^2, \quad C_3=\sum_{r=1}^{R}\pi_r\sigma_r^2.$$

**Dillema:** High $\|\Delta\mathbf{W}\|_F$ may degrade preserving remaining concepts
Low $\|\Delta\mathbf{W}\|_F$ may degrade erasing target concepts

☐ **Corollary 1.** (Modification of Theorem 2. with nonlinear projection $f(\mathbf{E})=\mathbf{V}_r$)

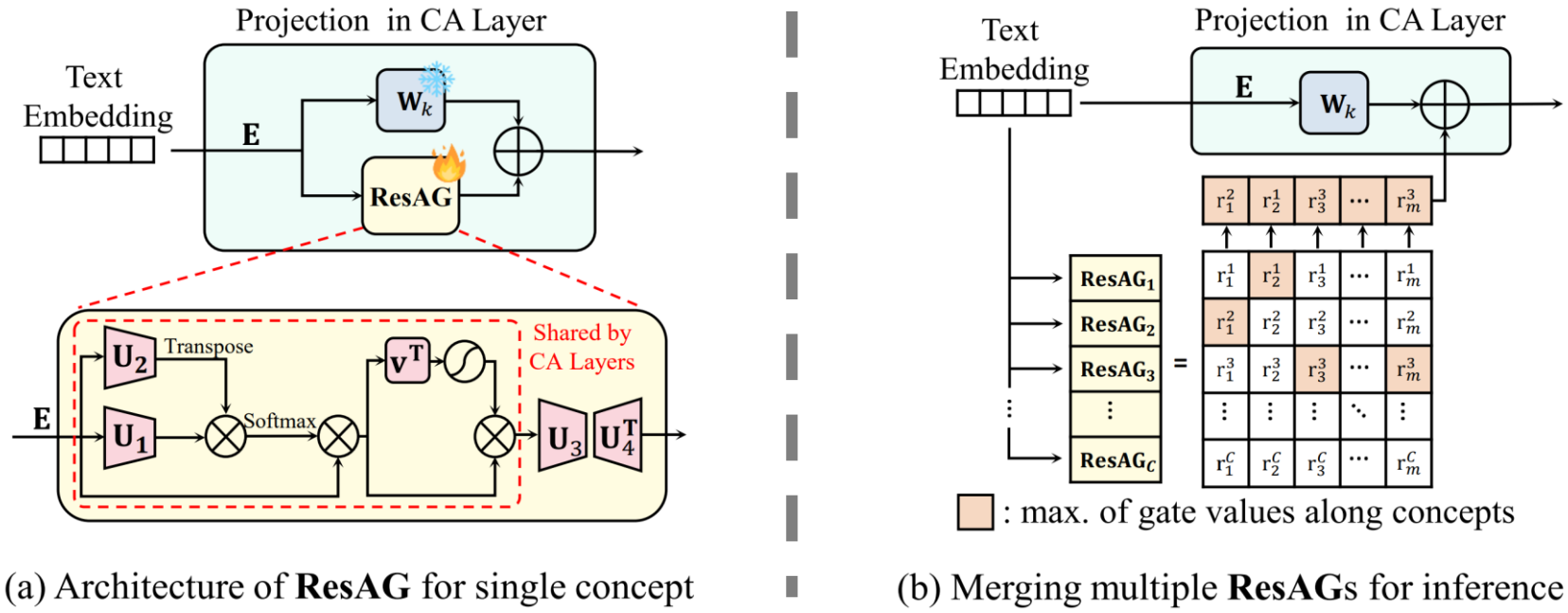
$$\mathbb{E}_{\mathbf{E}_{\text{rem}}}\left[\|\Delta\mathbf{W}\mathbf{E}_{\text{rem}}f(\mathbf{E}_{\text{rem}})\|_F^2\right] = \|\Delta\mathbf{W}\|_F^2\sum_{r=1}^{R}\pi_r\sigma_r^2\|\mathbf{V}_r\|_F^2 + \sum_{r=1}^{R}\pi_r\|\Delta\mathbf{W}\boldsymbol{\mu}_r\mathbf{V}_r\|_F^2.$$

# Residual Attention Gate (ResAG)

☐ For $f(\mathbf{E})$, we introduce a selective gating mechanism called **Residual Attention Gate (ResAG)** inspired from attention gate [3].

ResAG

$$f_c(\mathbf{E}) = \mathbf{A}_c S(\mathbf{v}_c^T \mathbf{E} \mathbf{A}_c),\ \mathbf{A}_c = \sigma\left(\frac{(\mathbf{U}_{1,c}\mathbf{E})^T(\mathbf{U}_{2,c}\mathbf{E})}{\sqrt{m}}\right)$$

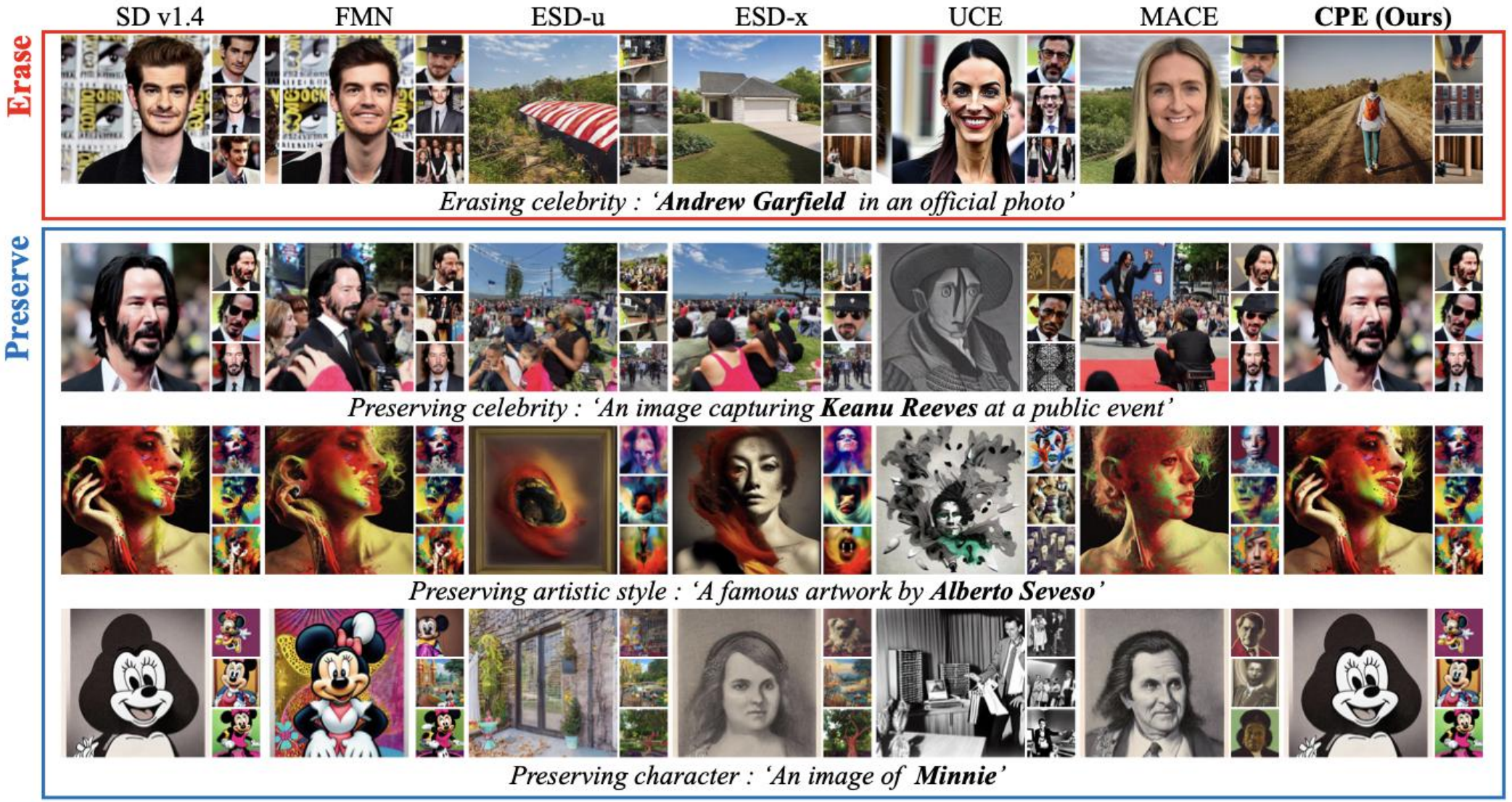



(a) Architecture of **ResAG** for single concept

(b) Merging multiple **ResAG**s for inference

[3] Oktay, Ozan, et al. "Attention u-net: Learning where to look for the pancreas." *MIDL*. 2022.

# Results - Celebrities erasure



# Results- Explicit contents erasure

☐ We generate images from I2P prompts [4] consisting of 4,703 ordinary prompts without inappropriate words which bypass to generate offensive contents.

| Method | Number of nudity detected on I2P (Detected Quantity) | | | | | | | | | COCO-30K | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Armpits | Belly | Buttocks | Feet | Breasts (F) | Genitalia (F) | Breasts (M) | Genitalia (M) | Total | CS ↑ | FID ↓ |
| FMN (Zhang et al., 2023a) | 43 | 117 | 12 | 59 | 155 | 17 | 19 | **2** | 424 | 30.39 | 13.52 |
| ESD-x (Gandikota et al., 2023) | 59 | 73 | 12 | 39 | 100 | 4 | 30 | 8 | 315 | 30.69 | 14.41 |
| ESD-u (Gandikota et al., 2023) | 32 | 30 | **2** | 19 | 35 | 3 | 9 | **2** | 123 | 30.21 | 15.10 |
| UCE (Gandikota et al., 2024) | 29 | 62 | 7 | 29 | 35 | 5 | 11 | 4 | 182 | 30.85 | 14.07 |
| MACE (Lu et al., 2024) | 17 | 19 | **2** | 39 | 16 | **0** | 9 | 7 | 111 | 29.41 | **13.42** |
| RECE (Gong et al., 2024) | 31 | 25 | 3 | **8** | 10 | **0** | 9 | 3 | 89 | 30.95 | - |
| **CPE (Ours)** | **10** | **8** | **2** | **8** | **6** | 1 | **3** | **2** | **40** | **31.19** | 13.89 |
| SD v1.4 (Rombach et al., 2022b) | 148 | 170 | 29 | 63 | 266 | 18 | 42 | 7 | 743 | 31.34 | 14.04 |
| SD v2.1 (Rombach, 2022) | 105 | 159 | 17 | 60 | 177 | 9 | 57 | 2 | 586 | 31.53 | 14.87 |

[4] Schramowski, Patrick, et al. "Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models." *CVPR*. 2023.

# Thank You!

## Concept Pinpoint Eraser for Text-to-Image Diffusion Models via Residual Attention Gate

**Poster session: Thu, April 24, 11:00 AM - 1:30 PM**

E-mail

Project page

Paper

This work was supported in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)] and by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. NRF2022R1A4A1030579, NRF-2022M3C1A309202211). Also, the authors acknowledged the financial support from the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University.