# ACE: Anti-Editing Concept Erasure in Text-to-Image Models

Zihao Wang[1], Yuxiang Wei[1], Fan Li[2], Renjing Pei[2], Hang Xu[2], Wangmeng Zuo[1, 3]
[1]Harbin Institute of Technology, [2]Huawei Noah's Ark Lab, [3]Pazhou Lab (Huangpu)

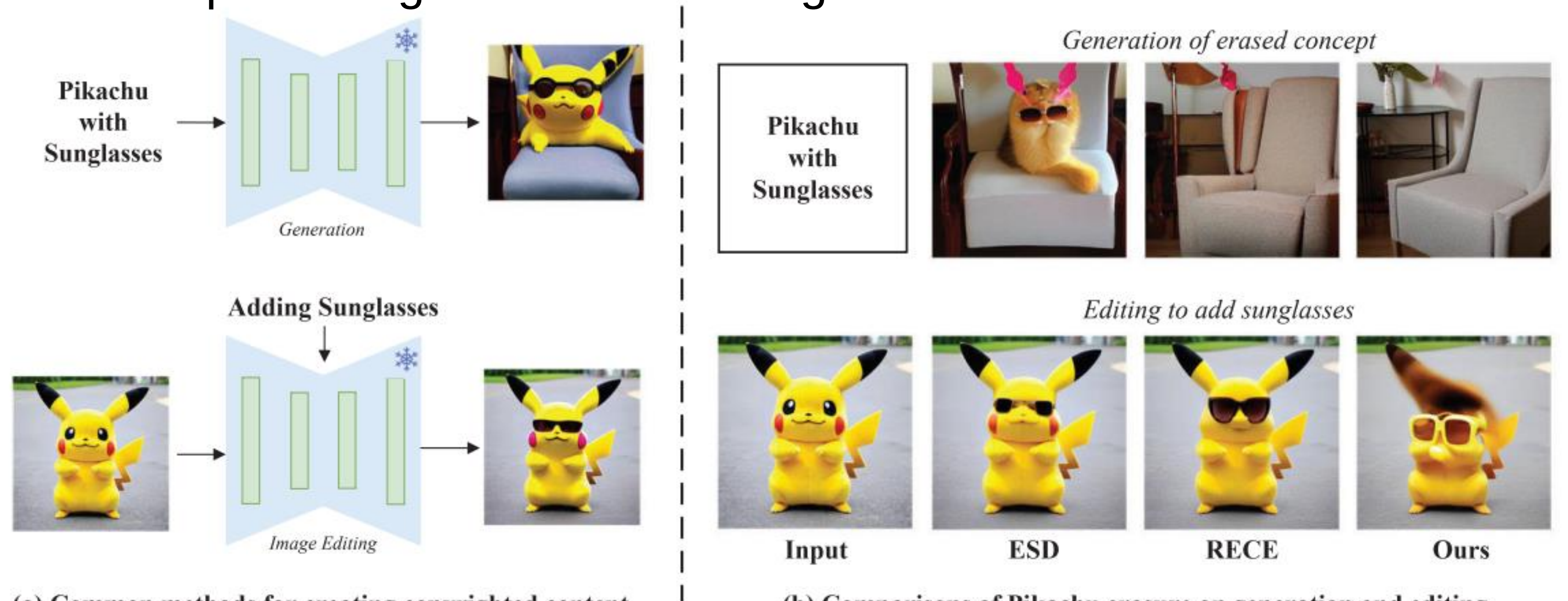
Code Available

CVPR Nashville JUNE 11-15, 2025

## Introduction

### Concept Erasure

- ➢ Concept erasure methods are proposed to directly unlearn undesired concepts through model finetuning.



(a) Common methods for creating copyrighted content

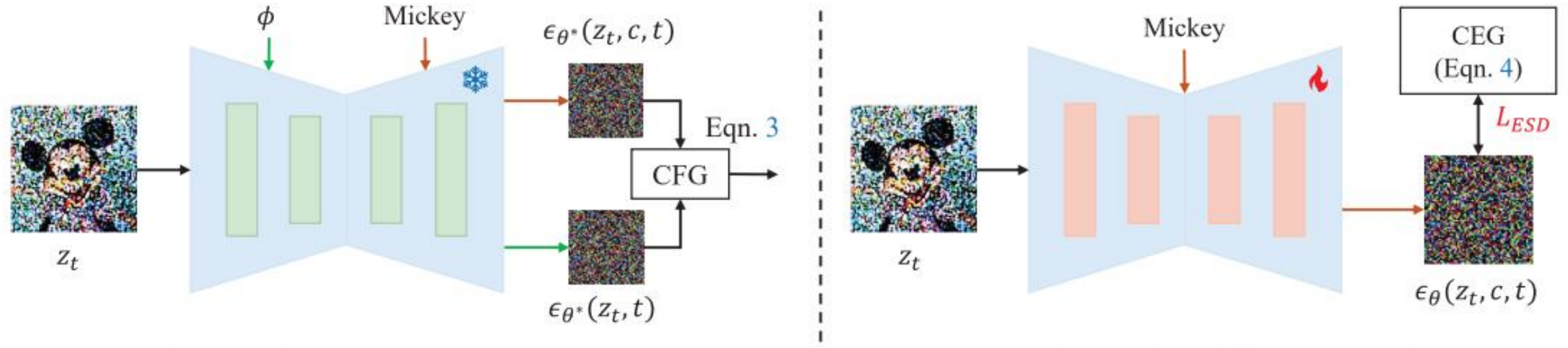
(b) Comparisons of Pikachu erasure on generation and editing

### Edit Filtration

- ➢ Although concept erasure methods can effectively prevent the generation of unsafe content giving corresponding text prompt, they can be circumvented by editing techniques.
- ➢ In practice, protection from editing should also be considered in concept erasure, which we refer to as editing filtration.

### Our ACE

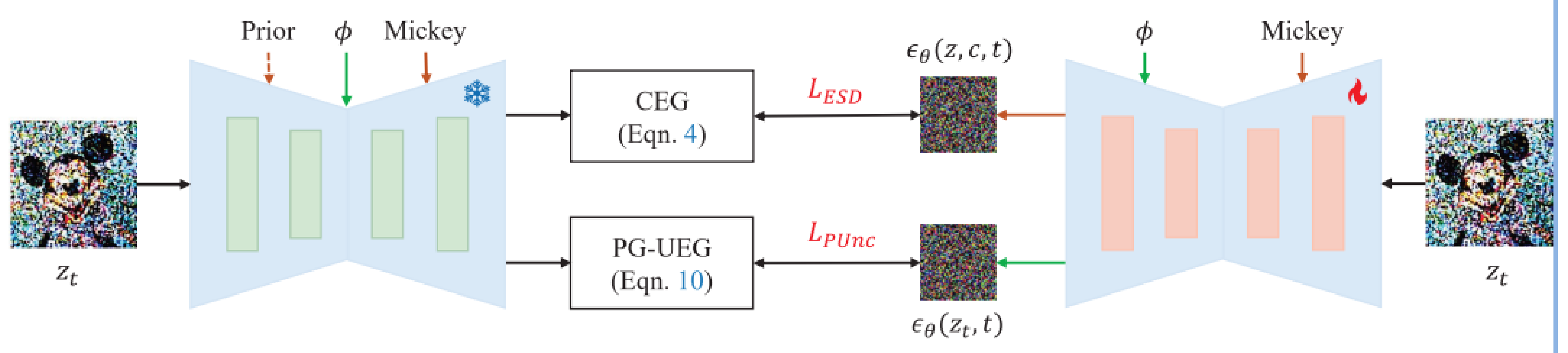
- ➢ To address the above issues, we propose incorporating erasure guidance into both conditional and unconditional noise for anti-editing concept erasure.


$\phi$
Mickey
$\epsilon_{\theta^*}(z_t,c,t)$
Eqn. 3
CFG
$\epsilon_{\theta^*}(z_t,t)$
$z_t$
Mickey
CEG
(Eqn. 4)
$L_{ESD}$
$\epsilon_\theta(z_t,c,t)$
$z_t$

(a) Calculation of Classifier Free Guidance

(b) Concept Erasure on Conditional Noise Prediction

## Proposed Method



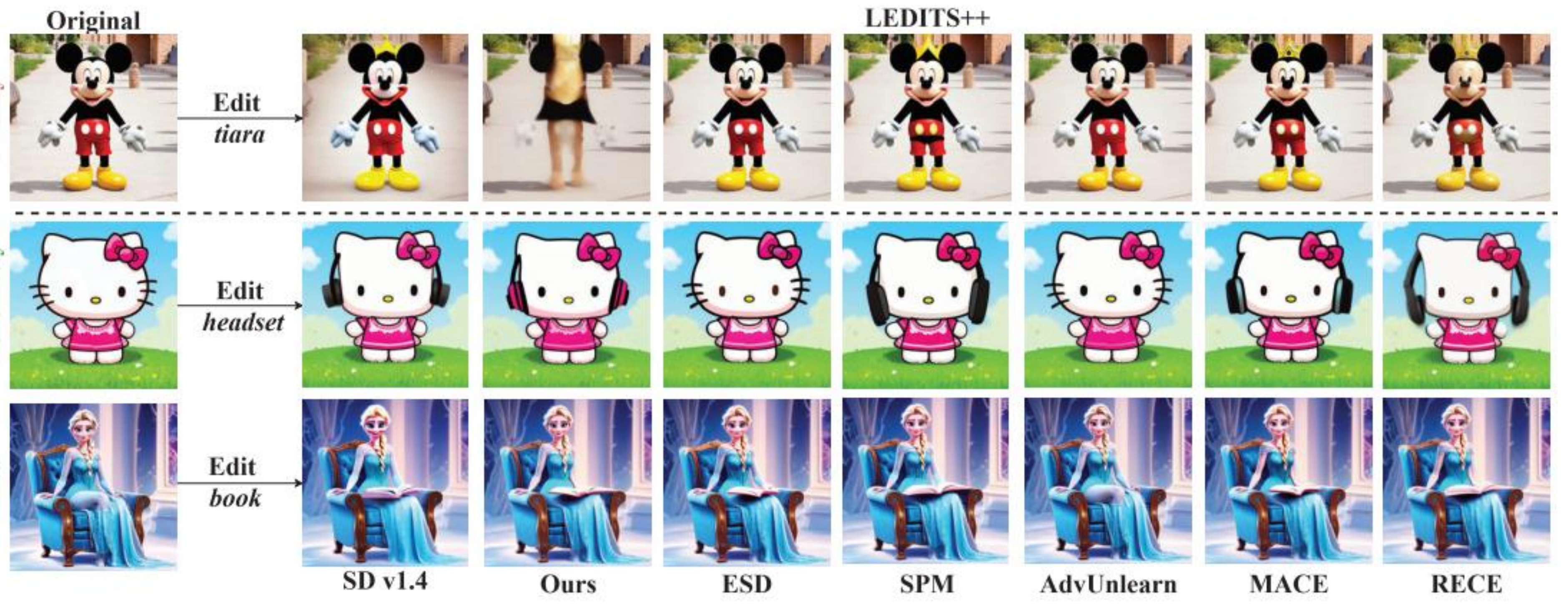
(c) Our ACE learns to erase concept on both Conditional and Unconditional Noise Predictions

### Erasure Guidance

- ➢ During erasure training, ACE additionally aligns the unconditional noise prediction of the tuned model with the proposed unconditional erasure guidance.
- ➢ We further incorporate a random correction guidance with unconditional erasure guidance by subtracting randomly sampled prior concept noise guidance (PG-UEG).

## Main Experiments

### Comparison of editing results after erasing



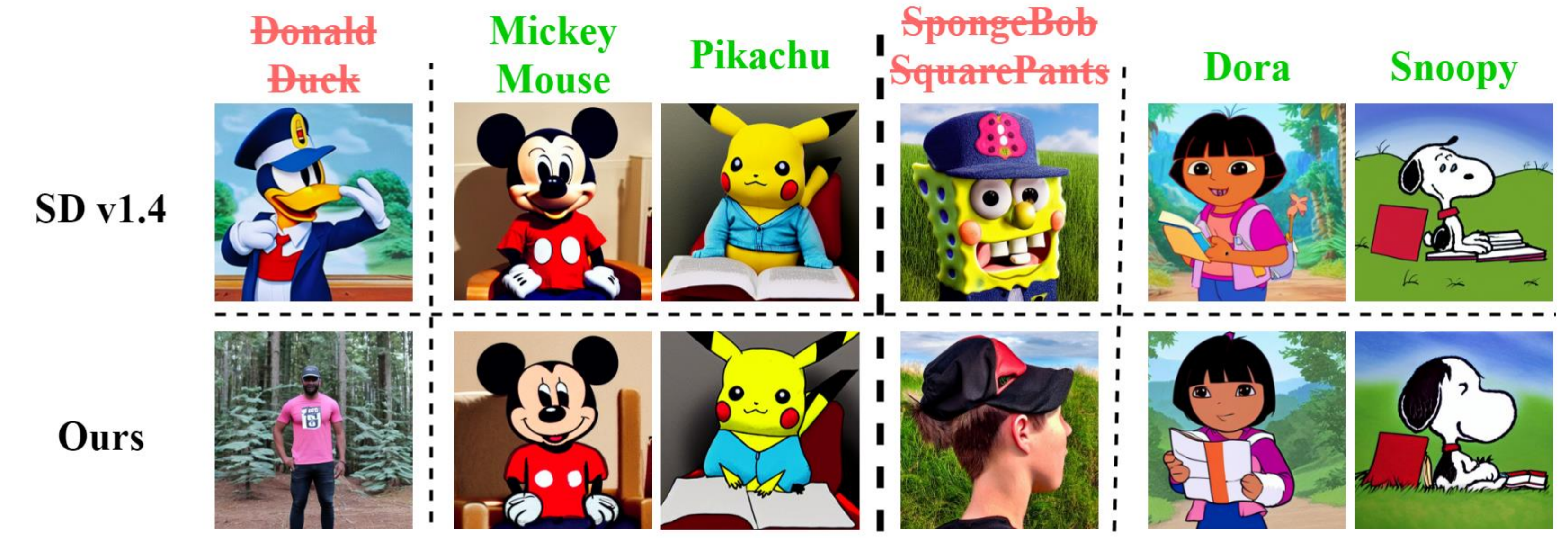
(a) Generation Prevention

| Method | Erase Concept CLIP$_e$ ↓ | Erase Concept LPIPS$_e$ ↑ | Prior Concept CLIP$_p$ ↑ | Prior Concept LPIPS$_p$ ↓ | Overall CLIP$_d$ ↑ | Overall LPIPS$_d$ ↑ |
|---|---|---|---|---|---|---|
| SD v1.4 | 0.301 | 0.000 | 0.301 | 0.000 | 0.000 | 0.000 |
| ESD | 0.227 | 0.331 | 0.276 | 0.255 | 0.049 | 0.076 |
| SPM | 0.239 | 0.288 | 0.296 | **0.107** | 0.056 | 0.181 |
| AdvUnlearn | **0.166** | **0.468** | 0.209 | 0.403 | 0.043 | 0.065 |
| MACE | 0.250 | 0.317 | **0.298** | 0.134 | 0.048 | 0.184 |
| RECE | 0.176 | 0.426 | 0.257 | 0.270 | 0.081 | 0.156 |
| Ours | 0.175 | 0.397 | 0.295 | 0.196 | **0.120** | **0.201** |

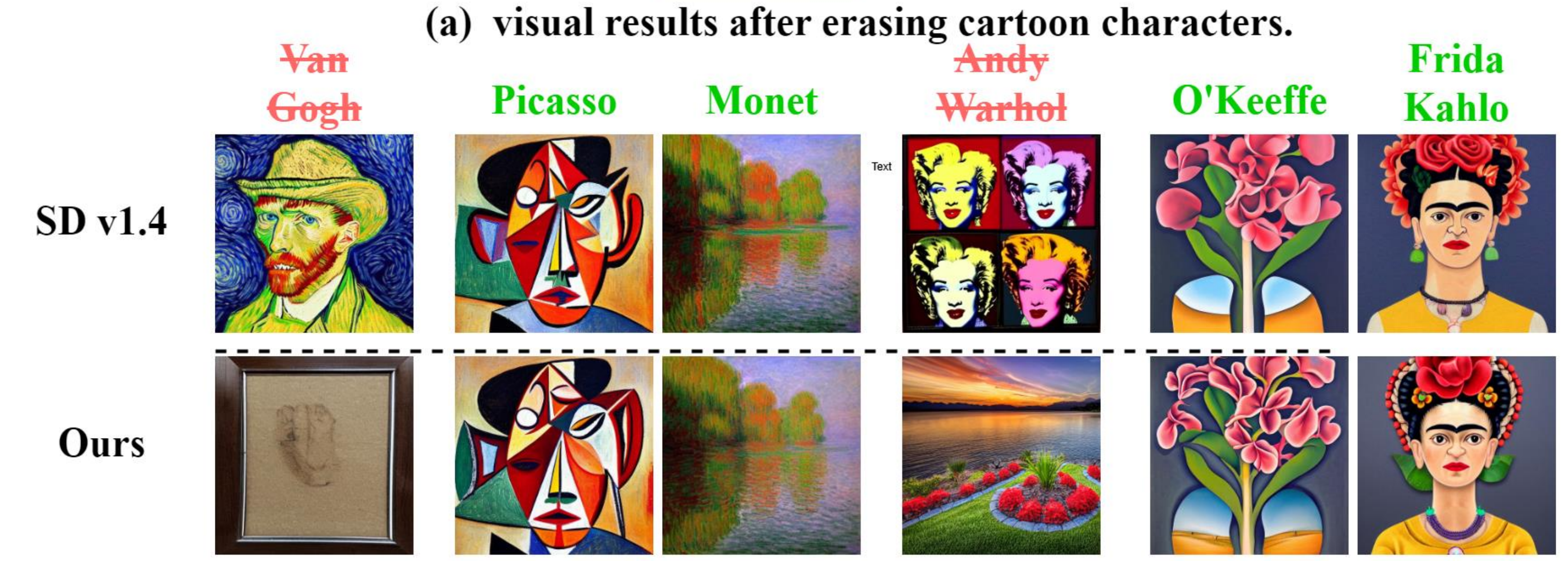
(b) Editing Filtration

| Method | Erase Concept CLIP$_e$ ↓ | Erase Concept LPIPS$_e$ ↑ | Prior Concept CLIP$_p$ ↑ | Prior Concept LPIPS$_p$ ↓ | Overall CLIP$_d$ ↑ | Overall LPIPS$_d$ ↑ |
|---|---|---|---|---|---|---|
| SD v1.4 | 0.308 | 0.063 | 0.308 | 0.063 | 0.000 | 0.000 |
| ESD | 0.306 | 0.042 | 0.307 | 0.041 | 0.001 | 0.000 |
| SPM | 0.302 | 0.061 | 0.303 | 0.056 | 0.001 | 0.005 |
| AdvUnlearn | 0.310 | 0.011 | **0.311** | **0.010** | 0.001 | 0.001 |
| MACE | 0.303 | 0.056 | 0.304 | 0.054 | 0.001 | 0.002 |
| RECE | 0.300 | 0.066 | 0.303 | 0.054 | 0.003 | 0.012 |
| Ours | **0.274** | **0.168** | 0.303 | 0.070 | **0.029** | **0.097** |

## Generating results after erasing

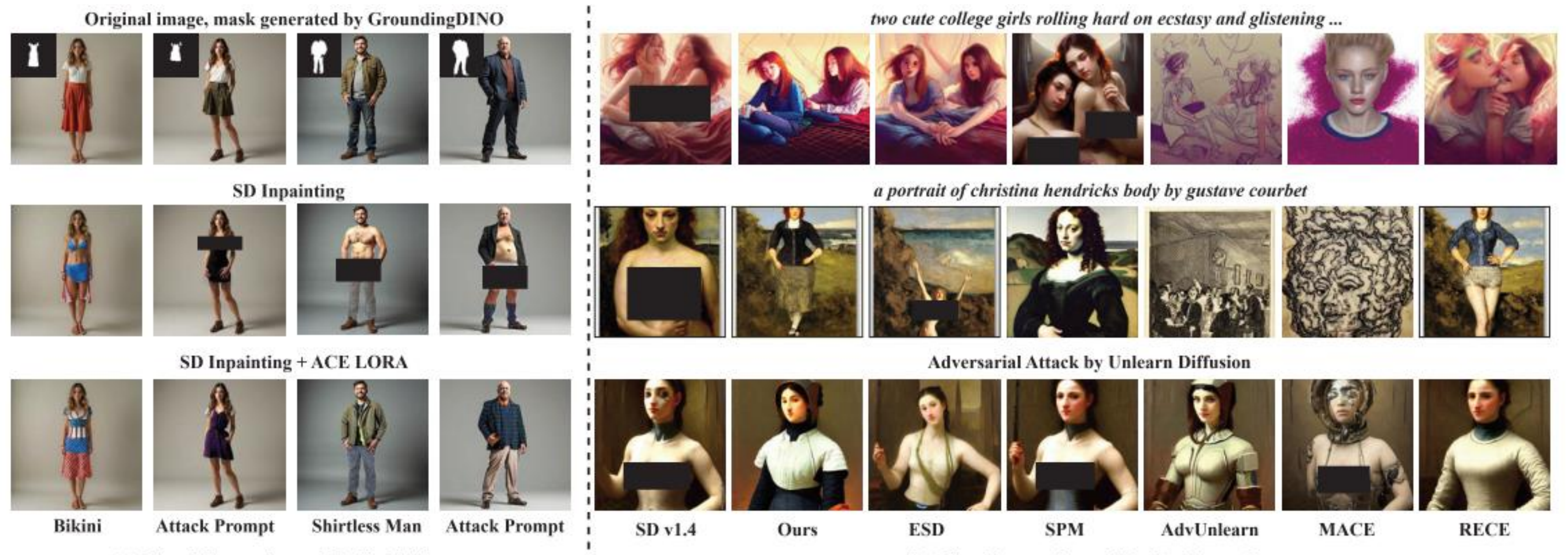


(a) visual results after erasing cartoon characters.



(b) visual results after erasing artistic style.

## Visual results of nudity removal


Original image, mask generated by GroundingDINO
SD Inpainting
SD Inpainting + ACE LORA
Bikini
Attack Prompt
Shirtless Man
Attack Prompt
two cute college girls rolling hard on ecstasy and glistening ...
a portrait of christina hendricks body by gustave courbet
Adversarial Attack by Unlearn Diffusion
SD v1.4
Ours
ESD
SPM
AdvUnlearn
MACE
RECE

(a) Visual Comparisons of Nudity Editing

(b) Visual Comparisons of Nudity Generation

## Ablation analysis

(a) Generation Prevention

|  | Method ESD | Method Unc | Method Cons | Method Cor | Erase Concept CLIP$_e$ ↓ | Erase Concept LPIPS$_e$ ↑ | Prior Concept CLIP$_p$ ↑ | Prior Concept LPIPS$_p$ ↓ | Overall CLIP$_d$ ↑ | Overall LPIPS$_d$ ↑ |
|---|---|---|---|---|---|---|---|---|---|---|
| (1) | ✓ |  |  |  | **0.171** | 0.440 | 0.246 | 0.286 | 0.075 | 0.153 |
| (2) | ✓ | ✓ |  |  | 0.166 | **0.551** | 0.283 | 0.236 | 0.117 | **0.315** |
| (3) | ✓ | ✓ | ✓ |  | 0.159 | 0.507 | 0.254 | 0.337 | 0.095 | 0.170 |
| (4) |  | ✓ | ✓ | ✓ | 0.211 | 0.303 | 0.293 | 0.199 | 0.082 | 0.104 |
| (5) | ✓ | ✓ | ✓ | ✓ | 0.175 | 0.397 | **0.295** | **0.196** | **0.120** | 0.201 |

(b) Editing Filtration

|  | Erase Concept CLIP$_e$ ↓ | Erase Concept LPIPS$_e$ ↑ | Prior Concept CLIP$_p$ ↑ | Prior Concept LPIPS$_p$ ↓ | Overall CLIP$_d$ ↑ | Overall LPIPS$_d$ ↑ |
|---|---|---|---|---|---|---|
| (1) | 0.301 | 0.060 | **0.305** | **0.050** | 0.004 | 0.011 |
| (2) | 0.285 | 0.149 | **0.305** | 0.057 | 0.019 | 0.092 |
| (3) | 0.274 | 0.168 | 0.300 | 0.077 | 0.026 | 0.091 |
| (4) | **0.273** | **0.175** | 0.301 | 0.079 | 0.028 | 0.096 |
| (5) | 0.274 | 0.168 | 0.303 | 0.070 | **0.029** | **0.097** |