ADAPTIVE INPAINTING MASKING FROM UNCONDITIONAL CONCEPT ERASURE BIAS FOR SAFE INPAINTING

Anonymous authors
Paper under double-blind review

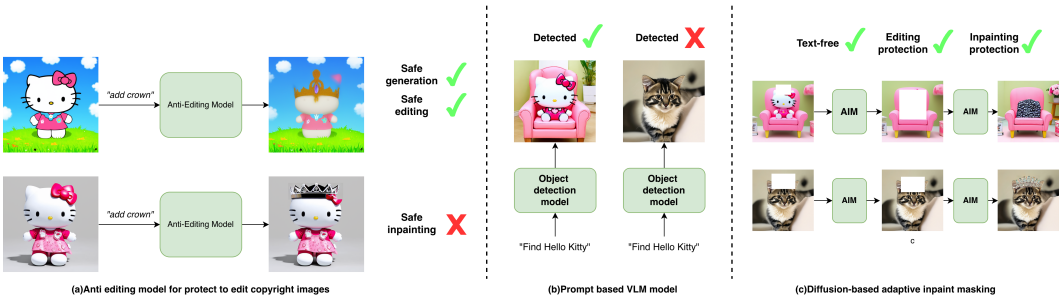

Figure 1: (a) illustrates a diffusion model that blocks attempts to modify copyright-protected images; while it successfully defends against global image editing, it still remains vulnerable to localized inpainting attacks. (b) shows a text-prompt–based detector that tries to check, before generation, whether the input image contains a prohibited character, but this approach is prone to false positives due to ambiguous or underspecified textual descriptions. (c) presents our proposed Adaptive Inpainting Masking from Unconditional Concept-Erasure Bias, which exploits the ACE-style unconditional bias to interpret visual cues without explicit text labels and adaptively refines the inpainting mask so that the subsequent inpainting remains safe.

ABSTRACT

Recent diffusion models enable high-fidelity image generation, editing, and inpainting, but they also reopen a subtle attack surface: an adversary can specify a narrowly localized inpainting mask so that the target concept is underrepresented in the masked region, thereby bypassing Anti-editing Concept Erasure (ACE). This mismatch stems from a design gap—current erasure methods operate on the denoising pathway but do not verify whether the user-specified inpainting mask actually covers the concept to be removed. We address this gap by introducing Adaptive Inpainting Masking (AIM) from Unconditional Concept Erasure Bias, a training-free pipeline that turns the ACE-imposed unconditional bias into a pre-inpainting mask refinement signal. We first isolate the object into an OOD-like view, then run deterministic unconditional DDIM inversion through the ACE model, and measure reconstruction consistency. Regions whose visual evidence aligns with the ACE-style unconditional bias are interpreted as containing the erased concept and are automatically merged into the inpainting mask. The refined mask thus captures concept-bearing areas that an attacker attempted to exclude, enabling the downstream inpainting step to remain concept-safe without any explicit text labels. Our approach effectively extends diffusion-based concept erasure to the inpainting setting, closing a practical evasion channel while preserving the training-free nature of existing ACE frameworks.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

The capability of diffusion models (Ho et al., 2020) to generate high-quality images in the text-to-image (T2I) (Hertz et al., 2022) task has been demonstrated (Rombach et al., 2022). This has enabled the free generation of creative content, extending beyond mere image generation to encompass diverse tasks such as inpainting (Lugmayr et al., 2022; Chen et al., 2024b; Manukyan et al., 2024; Tang et al., 2024; Yu et al., 2023), image editing (Kawar et al., 2023; Brack et al., 2023; Brooks et al., 2023; Chefer et al., 2023; Kumari et al., 2023), and style transfer (Hu et al., 2024; Ye et al., 2023; Wang et al., 2024; Chen et al., 2024a; Xing et al., 2024). However, this advancement has also introduced potential risks of misuse, such as copyright infringement or the creation of harmful content by mimicking protected or explicit images. To address these issues, research on concept erasure is being actively conducted (Gandikota et al., 2024; Gao et al., 2025; He et al., 2025), designed to prevent diffusion models from generating specific target concepts (Wang et al., 2025).

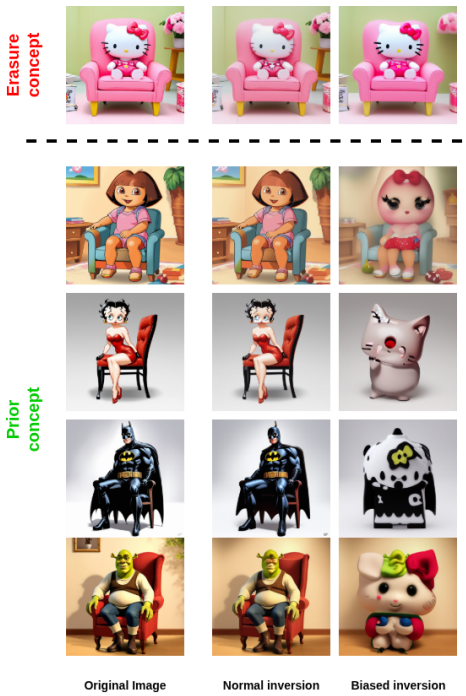

Figure 2: **DDIM inversion visualization.** Target concepts are accurately reconstructed, while conflicting prior concepts are severely distorted.

Existing concept erasure studies have proposed fine-tuning methods to prevent diffusion models from directly generating the target concept (Gandikota et al., 2023; Lu et al., 2024; Lyu et al., 2024; Gong et al., 2024; Hu et al., 2024). While successful in inhibiting generation, their reliance on prompts creates a vulnerability; images can still be modified or generated through prompt bypassing. Particularly in image editing, generation is feasible even if the text prompt does not explicitly contain terms referring to the target concept. Consequently, if the target concept is visually present in the image, protection is difficult to enforce solely through text-based triggers.

To mitigate such misuse in image editing, Anti-Editing Concept Erasure (ACE) (Wang et al., 2025) was introduced. This method leverages Classifier-Free Guidance (CFG) (Ho & Salimans, 2022) by training the unconditional prediction to possess a specific bias aligned with the target concept (termed Unconditional Concept Erasure Bias). This allows the model to suppress the concept's generation during the guidance process, even when the conditional text prompt lacks specific keywords. Consequently, effective concept erasure was achieved in both generation and editing tasks.

However, a critical vulnerability persists in the inpainting task. It remains possible to inappropriately modify images by strategically applying masks that either intentionally avoid the target object or provide only a localized mask where the object's information is insufficiently revealed. Figure 1 (a) illustrates a failure case: while ACE protects the 'Hello Kitty' concept during prompt-based editing, applying a localized inpainting mask allows for its successful reconstruction.

Crucially, this failure occurs because inpainting models rely heavily on the unmasked regions to maintain semantic consistency. When a mask is localized, the global context of the target object remains visible in the unmasked region. This strong visual condition overrides the Unconditional Concept Erasure Bias learned by ACE, causing the model to prioritize the contextual integrity of the object over the erasure instruction. With the lack of a mechanism to detect and suppress this "context leakage" from the input image itself, the motive of the erasure framework is compromised.

Therefore, achieving robust concept erasure in inpainting necessitates a proactive approach that analyzes the input image to detect the presence of forbidden concepts before the generation process begins. In this context, relying on vision-language models (VLMs) like DINOv2 or CLIP is insufficient, as they often fail to recognize or map specific IP characters or abstract concepts (Figure 1 (b)).

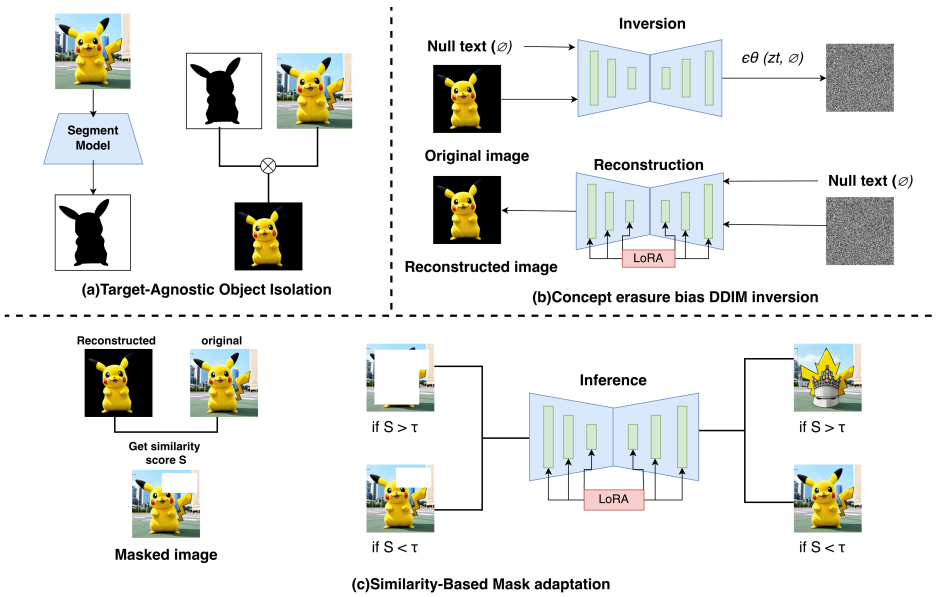

Figure 3: (a) To leverage the out-of-domain (OOD) characteristics of the diffusion model, an image is generated that is distant from the general image distribution. This is achieved by masking the background, excluding the object, to create a black backdrop. (b) Unconditional DDIM inversion is performed using the ACE model, which is trained for concept erasure. Due to the conceptual bias embedded in the unconditional branch, images that align with this bias are stably reconstructed. Conversely, images that conflict with the bias result in severely distorted reconstructions. (c) The degree of distortion is quantified by comparing the similarity between the reconstructed and original images. If the similarity score exceeds a predefined threshold, the presence of the forbidden concept is assumed, and the corresponding region is adaptively masked.

To address this limitation, we propose **A**daptive **I**npainting **M**asking (AIM) from Unconditional Concept Erasure Bias for Safe Inpainting. The core principle is to adaptively reinforce the inpainting mask based on the input image's content, effectively blocking the context leakage identified above. We achieve this by repurposing the Concept Erasure Bias previously injected into the unconditional branch by the ACE method, without requiring external detectors (Figure 1 (c)).

Specifically, we utilize the discrepancy in reconstruction stability during unconditional DDIM inversion (Mokady et al., 2022). Since ACE trains the unconditional branch to align with the target concept, regions containing the target concept exhibit stable reconstruction during inversion. Conversely, general objects (prior concepts) conflict with this bias, leading to distorted reconstruction results (Figure 2). We exploit this "bias-induced reconstruction conflict" as a detection signal. By identifying regions that stabilize under the erasure bias and merging them with the user-specified mask, we effectively remove the global context required for malicious inpainting.

Figure 3 visualizes this process: (a) amplifying the bias effect via Out-of-Distribution (OOD) perturbation, (b) inducing biased reconstruction via unconditional inversion, and (c) adapting the mask based on the similarity metric. This approach ensures safe inpainting by leveraging only the inherent visual representations within the diffusion model.

2 Proposed Method

To preemptively defend against Inpainting Attacks which leverage mask bypassing in the inpainting task to generate images containing erasure concepts, we propose Adaptive Inpainting Masking (AIM). Before the inpainting request is executed, AIM filtering whether the target concept C_{target} (Erasure Concept) is present in the input image I. For safe inpainting, it adaptively reproduces the given user mask M_{user} into M_{adapt}. AIM utilizes the concept-erased ACE model (Wang et al., 2025) itself as a filter, obviating the need for a separate filter or text labels. This methodology is composed

of three stages: (1) Target-Agnostic Object Isolation (2) Biased Reconstruction via Unconditional Inversion (3) Similarity-Based Mask Adaptation.

2.1 PRELIMINARIES

Classifier-Free Guidance (CFG) CFG (Ho & Salimans, 2022) is a widely used technique in text-to-image diffusion models to enhance generation quality. It is an operation that amplifies the dependency on the text condition c by leveraging the difference between the conditional prediction $\epsilon_\theta(z_t, c)$ and the unconditional prediction $\epsilon_\theta(z_t, \emptyset)$. The final noise prediction $\tilde{\epsilon}_\theta$ is computed as:

$$\tilde{\epsilon}_\theta(z_t, c) = \epsilon_\theta(z_t, \emptyset) + w(\epsilon_\theta(z_t, c) - \epsilon_\theta(z_t, \emptyset)) \tag{1}$$

where w is the guidance scale. As seen in this equation, the unconditional prediction $\epsilon_\theta(z_t, \emptyset)$ directly influences the final noise prediction. We term the specific bias injected into this unconditional branch by ACE as the **Unconditional Concept Erasure Bias**. Specifically, ACE trains the unconditional branch to implicitly align with the target concept's features. We repurpose this property: if an input image aligns with this bias, it indicates the presence of the target concept. Traditional concept erasure methods primarily fine-tuned only the conditional prediction $\epsilon_\theta(z_t, c)$, causing vulnerability to editing or inpainting attacks where the text prompt c excludes the target concept.

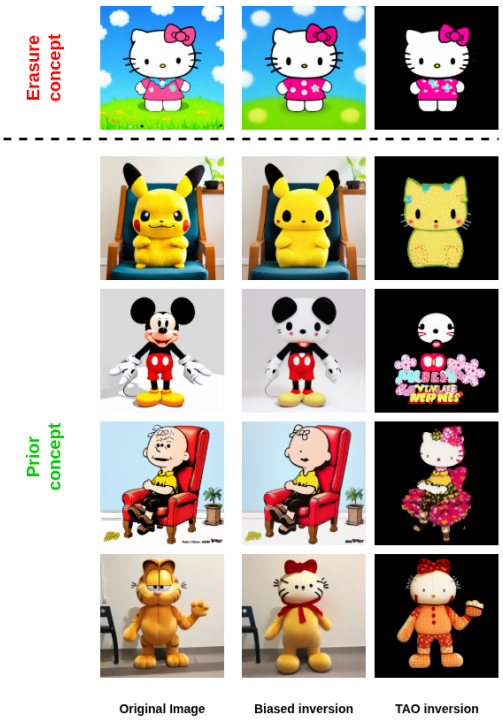

Figure 4: **Effect of Target-Agnostic Object Isolation (TAO).** In some cases, biased inversion alone does not sufficiently distort the target concept. The image distribution is shifted to Out-Of-Distribution (OOD) by masking the background. This forces the diffusion model to focus on the object during reconstruction, allowing the model's learned bias to be applied intensively to that specific object.

DDIM Inversion DDIM (Song et al., 2022) inversion is a process for deterministically encoding a real image I into the diffusion model's latent noise z_T. Starting from $z_0 = \mathcal{E}(I)$ obtained via the VAE encoder, the DDIM sampling process is reversed. For each step t from $t = 0$ to $T - 1$, the noise prediction $\epsilon_\theta(z_t, c)$ is used to estimate $\hat{z}_0$, which is then used to compute the noise z_{t+1} for the next step (based on deterministic DDIM, i.e., $\sigma_t = 0$):

$$\hat{z}_0(t) = \frac{z_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(z_t, c)}{\sqrt{\bar{\alpha}_t}} \tag{2}$$

$$z_{t+1} = \sqrt{\bar{\alpha}_{t+1}} \cdot \hat{z}_0(t) + \sqrt{1 - \bar{\alpha}_{t+1}} \cdot \epsilon_\theta(z_t, c) \tag{3}$$

Anti-Editing Concept Erasure ACE (Wang et al., 2025) addresses this problem by directly injecting a bias against the target concept C_{target} into the unconditional prediction $\epsilon_\theta(z_t, \emptyset)$. It defines an Unconditional Erasure Guidance (UEG) $\tilde{\epsilon}u$, and the model's $\epsilon_\theta(z_t, \emptyset)$ was trained to follow this guidance.

$$\tilde{\epsilon}_u = \epsilon_{\theta^*}(z_t, \emptyset) + \eta_u(\epsilon_{\theta^*}(z_t, c) - \epsilon_{\theta^*}(z_t, \emptyset)) \tag{4}$$

$$\mathcal{L}_{Unc} = \mathbb{E}_{z_t,t,c}[||\epsilon_\theta(z_t, \emptyset) - \tilde{\epsilon}_u||_2^2] \tag{5}$$

where ϵ_θ denotes the pre-trained model. Furthermore, to prevent damage to unrelated concepts (prior concepts) c_p, ACE utilized a refined Prior-Guided Unconditional Erasure Guidance (PG-UEG) $\tilde{\epsilon}_{pu}$ and a final loss function $\mathcal{L}_{PUnc}$.

$$\tilde{\epsilon}_{pu} = \tilde{\epsilon}_u - \eta_p\gamma_p(\epsilon_{\theta^*}(z_t, c_p) - \epsilon_{\theta^*}(z_t, \emptyset)) \tag{6}$$

$$\mathcal{L}_{PUnc} = \mathbb{E}_{z_t,t,c,c_p \in \mathcal{C}_p}[||\epsilon_\theta(z_t, \emptyset) - \tilde{\epsilon}_{pu}||_2^2] \tag{7}$$

As a result, the Unconditional Concept Erasure Bias is injected. Our methodology, AIM, leverages this Unconditional Concept Erasure Bias, learned in $\epsilon_\theta(z_t, \emptyset)$ via $\mathcal{L}PUnc$, as its concept filtering mechanism.

2.2 TARGET-AGNOSTIC OBJECT ISOLATION(TAO)

The Inpainting Attack is a task focused on local objects rather than the global context of an image. Therefore, to enhance detection accuracy and minimize interference from unnecessary background information, we first isolate the main objects within the image. Utilizing the open-vocabulary detector Grounding DINO (Liu et al., 2024), we use a set of comprehensive and neutral text prompts $T_{agnostic} = \{$'character', 'human', 'object'$\}$ to extract a set of bounding boxes $\mathcal{B} = \{B_1, B_2, \ldots, B_N\}$ for the main objects in image I. For each bounding box B_i, a precise segmentation mask M_i is acquired by employing SAM (Kirillov et al., 2023), using B_i as a prompt. The background-removed image I_{iso} is generated by setting all pixels not belonging to M_i to 0. For a pixel at location p, I_{iso} is defined as:

$$I_{iso}(p) = \begin{cases} I(p) & \text{if } p \in \bigcup_i B_i \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

This background removal process serves two crucial purposes: (1) It removes background noise, focusing the subsequent feature comparison on the object(s). (2) It intentionally transforms the image into an **Out-Of-Distribution (OOD)** sample. As shown in Figure 4, while biased inversion alone might not induce sufficient distortion, shifting the image to the OOD space amplifies concept degradation. When I_{iso} is passed through DDIM inversion, the model cannot rely on global context and is forced to rely on its internal priors. Since the ACE model's internal prior is modified (biased), this OOD state drives the latent representation into an **Unstable Latent Region**. In this region, the conflict between the visual evidence (concept present) and the model's erasure bias is maximized, leading to catastrophic degradation of the concept's features during reconstruction.

2.3 BIASED RECONSTRUCTION VIA UNCONDITIONAL INVERSION

The core insight of AIM is to repurpose the unconditional bias that the ACE model intentionally injected to erase C_{target}. ACE was trained via $\mathcal{L}_{Unc}$ such that the unconditional prediction $\epsilon_\theta(z_t, \emptyset)$ incorporates the visual features of C_{target}. To use this bias as a detector, we apply Unconditional DDIM Inversion to I_{iso}. First, the VAE encoder $\mathcal{E}$ is used to encode I_{iso} into the latent vector z_0. Using the concept-erased model ϵ, the noise map z_T is generated via the DDIM Inversion process. The most critical step is the reconstruction process from z_T back to z'_0, which exclusively uses the unconditional prediction $\tilde{\epsilon}_\theta(z_t) = \epsilon_\theta(z_t, \emptyset)$ during denoising. Following the DDIM scheduler, the following is repeated from step T down to 1:

$$\hat{z}_0(t) = \frac{z_t - \sqrt{1-\bar{\alpha}_t}\tilde{\epsilon}_\theta(z_t)}{\sqrt{\bar{\alpha}_t}} \tag{9}$$

$$z_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \cdot \hat{z}_0(t) + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2} \cdot \tilde{\epsilon}_\theta(z_t) + \sigma_t \epsilon \tag{10}$$

We adopt the deterministic DDIM sampler with $\sigma_t = 0$ to invert the input image I_{iso} through the ACE model. Let z_t be the latent at timestep t. During inversion, the denoising step uses the ACE-trained unconditional predictor $\epsilon_\theta(z_t, \emptyset)$, whose output has been biased toward erasing the target concept. After running the reverse process, we obtain the reconstructed latent z'_0, which is finally decoded to the pixel space by the VAE decoder D:

$$I_{\text{recon}} = D(z'_0). \tag{11}$$

This enables our core hypothesis, *biased reconstruction*, to take effect. If the isolated input I_{iso} actually contains the target concept C_{target} and thus its visual evidence aligns with the ACE model's unconditional bias b_{target}, then the biased unconditional predictor $\epsilon_\theta(z_t, \emptyset)$ acts in a direction that *reinforces* these target-related features. Consequently, the denoising trajectory is guided by b_{target} and produces a reconstruction that is highly consistent with the input:

$$I_{\text{recon}} \approx I_{\text{iso}} \quad \text{(target concept present)}. \tag{12}$$

ss.

Figure 5: **Qualitative results of Inpainting IP characters.** Following the erasure of the 'SpongeBob' concept, inpainting was executed via the local application of a mask. The empirical results demonstrate that while baseline methods failed to exhibit robustness against the subsequent inpainting attack, our proposed methodology successfully mitigated this attack by adaptively modulating the mask. Consequently, the synthesis of existing (non-target) concepts was performed correctly.

In contrast, if I_{iso} does *not* correspond to C_{target}, the visual information in I_{iso} conflicts with b_{target}. Since we additionally place the object in an OOD-like setting (by isolating it from the original context), this mismatch is further amplified along the DDIM denoising steps. In this case, the unconditional bias behaves effectively as a *corrupting* direction for that latent, and the final reconstruction loses the original structure:

$$I_{\text{recon}} \not\approx I_{\text{iso}} \quad \text{(target concept absent)}. \tag{13}$$

This contrast between the bias-aligned reconstruction and the corrupted reconstruction provides the signal that we later exploit for adaptive mask refinement in safe inpainting (see Figure 4).

2.4 Similarity-Based Mask Adaptation

Finally, we quantitatively measure the success of the "Biased Reconstruction" to determine the presence of the target concept C_{target}. Based on the bounding box B obtained during the object isolation stage, we measure the visual similarity between the original I_{iso} and the reconstructed I_{recon}. To compare pure visual features without relying on text, DINOv2 (Oquab et al., 2024) is employed as the feature extractor. Let f_{iso} be the features of I_{iso} within region B and f_{recon} be the features of I_{recon}. We compute the Cosine Similarity S between the two feature vectors:

$$S = \frac{f_{iso} \cdot f_{recon}}{\|f_{iso}\| \|f_{recon}\|} \tag{14}$$

Based on a predefined threshold τ, if $S > \tau$, it is determined that I_{recon} was successfully reconstructed. This implies that the object in I_{iso} is the target concept that ACE is biased to erase. In this case, the object's bounding box B is forcibly added to the user's inpainting mask M_{user}, updating it to $M_{adapt} = M_{user} \cup M_B$.

$$M_{adapt} = \begin{cases} M_{user} \cup M_B & \text{if } S > \tau \\ M_{user} & \text{otherwise} \end{cases} \tag{15}$$

If $S \leq \tau$, it is determined that the image was distorted and failed to reconstruct, meaning it is not the target concept, and $M_{adapt} = M_{user}$ is returned. This refined mask M_{adapt} ensures that the target concept cannot be intentionally bypassed, guaranteeing robust inpainting defense.

3 Experiments

In this section, we demonstrate the performance of our proposed Adaptive Inpainting Masking (AIM) through several experiments: (1) IP character inpainting for copyrighted characters, (2) Artistic style inpainting, (3) NSFW filtering, and (4) Semantic overlap image filtering. To compare the defensive performance against inpainting, we use ACE (Wang et al., 2025), ESD (Gandikota et al., 2023), SPM (Gandikota et al., 2023), and RECE (Gong et al., 2024) as baselines. All experiments were conducted using Stable Diffusion v1.4 (Rombach et al., 2022). Additional results are provided in the supplementary material.

3.1 IP Character Inpainting

Experimental Setup To validate the effectiveness of the proposed AIM pipeline, we followed the experimental setup of ACE (Wang et al., 2025), using ten concept characters: Hello Kitty, Snoopy, Mickey Mouse, Elsa, Donald Duck, Dora the Explorer, Winnie the Pooh, Sonic the Hedgehog, SpongeBob, and Pikachu. Fine-tuning was performed for each concept. We employed $\text{CLIP}_{\text{concept}}$,

$\text{CLIP}_{\text{prompt}}$ (Radford et al., 2021), and LPIPS (Zhang et al., 2018) for evaluation. $\text{CLIP}_{\text{concept}}$ measures the alignment between the image and the concept text, while $\text{CLIP}_{\text{prompt}}$ measures alignment with the generation prompt. LPIPS quantifies the perceptual difference between the original and modified images.

For **Target Concepts** (erasure targets), lower $\text{CLIP}_{\text{concept}}$ and $\text{CLIP}_{\text{prompt}}$ scores indicate superior performance (successful erasure). Conversely, for **Prior Concepts** (non-targets), higher scores are desirable (preservation). For LPIPS, a higher value is preferred for target concepts (indicating successful distortion/erasure), while a lower value is preferred for prior concepts (indicating fidelity preservation). Overall performance was assessed by calculating the difference between the target and prior concept scores; a larger gap signifies a superior model.

For the inpainting evaluation, a total of 150 source images (15 for each of the 10 concepts) were generated using Stable Diffusion 3 (Esser et al., 2024). Each image was paired with 10 distinct prompts, creating a base dataset of 1,500 image-prompt pairs. To ensure a comprehensive assessment of both erasure robustness and prior preservation capabilities, **each of these 1,500 pairs was evaluated against all 10 target concept models**. This exhaustive cross-evaluation protocol resulted in a total of 15,000 evaluation instances. The evaluation was performed by applying a uniform mask over the head region of the characters. Detailed prompts and additional results are provided in the supplementary material.

Experimental Results Table 1 demonstrates that AIM successfully performs safe inpainting on images containing the **Target Concept** by adaptively expanding the mask. We report results at similarity thresholds (τ) of 0.8, 0.85, and 0.90. A distinct trade-off was observed: higher thresholds provide better preservation for **Prior Concepts** (fewer false positives) but result in weaker defense for Target Concepts (more false negatives).

Table 1: **Quantitative comparisons of IP character erasure.** The best result is highlighted with **bold**.

	Erase Concept			Prior Concept			Overall		
Method	$\text{CLIP}_c \downarrow$	$\text{CLIP}_p \downarrow$	LPIPS $\uparrow$	$\text{CLIP}_c \uparrow$	$\text{CLIP}_p \uparrow$	LPIPS $\downarrow$	$\text{CLIP}_c \uparrow$	$\text{CLIP}_p \uparrow$	LPIPS $\uparrow$
SD v1.4 (Rombach et al., 2022)	30.80	30.62	0.0695	30.80	30.62	0.0695	0.000	0.000	0.000
ESD (Gandikota et al., 2023)	**21.87**	30.88	0.1208	22.02	30.90	0.1192	0.1569	0.0201	0.0016
SPM (Gandikota et al., 2023)	22.21	31.34	0.0727	22.24	31.34	0.0726	0.0302	0.0081	0.0000
RECE (Gong et al., 2024)	22.76	31.50	0.1273	23.40	**31.59**	0.1271	0.6452	0.0965	0.0002
ACE (Wang et al., 2025)	30.58	30.28	0.0767	**30.80**	30.36	**0.0687**	0.220	0.080	0.008
AIM_{80}	26.26	**28.58**	**0.2274**	29.10	29.71	0.1323	**2.840**	**1.130**	**0.0951**
AIM_{85}	29.27	29.89	0.1252	29.94	30.06	0.1025	0.670	0.170	0.0227
AIM_{90}	29.71	30.03	0.1073	30.20	30.16	0.0927	0.490	0.130	0.0146

Furthermore, inadvertent interventions on Prior Concepts (False Positives) led to lower generation fidelity compared to baselines. Notably, the LPIPS metric showed high sensitivity; since AIM aggressively extends the mask to cover the entire character upon detection, even a small number of false positives causes a significant deviation from the original image, inflating the perceptual distance.

In contrast, as visualized in Figure 5, baseline models (ACE, ESD, SPM, and RECE) demonstrated negligible defensive capabilities against inpainting, evidenced by the minimal performance gap between Prior and Target Concepts. The overall evaluation confirms that, unlike existing methods, AIM possesses effective and robust capabilities for inpainting defense.

3.2 Artistic inpainting

To validate whether AIM can defend against the inpainting of non-object concepts, such as artistic styles, ten specific styles were selected for evaluation: Leonardo da Vinci, Pablo Picasso, Michelangelo, Van Gogh, Salvador Dali, Claude Monet, Andy Warhol, Jackson Pollock, Frida Kahlo, and Georgia O'Keeffe. Ten images were generated for each style, totaling 100 images, using Stable Diffusion 3. The evaluation was conducted by applying a small 50x50 pixel mask and attempting reconstruction. If AIM identified the style as an erasure concept, the mask was expanded to cover the entire image. As the original ACE study demonstrated the ineffectiveness of other baselines for editing defense, we compared AIM directly against ACE. The evaluation metrics were set identically to those in Section 3.1.

Experimental Results Figure 6 presents a comparison of inpainting results on artistic styles. It is shown that AIM successfully adapts the mask to erase target styles. Table 2 shows that AIM successfully performs safe inpainting on images depicting specific art styles by expanding the mask. Results are reported for similarity thresholds of 0.7, 0.8, and 0.9. Identical to Section 3.1, this experiment exhibited the same trade-off relationship, where higher thresholds better protect prior concepts at the cost of weaker erasure defense. When AIM triggers erasure, it masks the entire image to generate a completely different output. Consequently, $LPIPS$ scores were significantly impacted by even minor misclassifications. Nonetheless, the CLIP scores demonstrate that masking of the erasure concept was effectively performed.



Figure 6: **Qualitative results of Inpainting artistic style image.** Our method demonstrate that the proposed method is capable of robustly performing filtering and inpainting to artistic styles images.

Experimental Results Table 5 presents the NSFW filtering results, with the DINO's performance broken down by prompt. Grounding DINO failed to filter in many instances, even when various prompts were applied. This indicates that Grounding DINO, as an object detection model, struggles with detection when the text prompt does not precisely describe the target object. This limitation renders it inadequate for this type of image filtering. In contrast, the AIM model, which learns diverse visual information corresponding to "nudity," demonstrated robust filtering capabilities across a wide range of images.

Table 2: **Quantitative comparisons of Artistic style erasure.** The best result is highlighted with **bold**.

	Erase Concept			Prior Concept			Overall		
Method	$CLIP_c \downarrow$	$CLIP_p \downarrow$	LPIPS ↑	$CLIP_c \uparrow$	$CLIP_p \uparrow$	LPIPS ↓	$CLIP_c \uparrow$	$CLIP_p \uparrow$	LPIPS ↑
ACE (Wang et al., 2025)	29.07	29.86	0.0050	**29.09**	**29.88**	**0.0047**	0.0246	0.0237	0.0003
AIM_{70}	**23.14**	**23.83**	**0.5064**	26.78	27.53	0.2405	**3.6313**	**3.6970**	**0.2659**
AIM_{80}	27.01	29.80	0.0119	28.07	29.86	0.0127	1.0667	0.0600	0.0008
AIM_{90}	27.81	28.63	0.1196	28.56	29.36	0.0575	0.7471	0.7242	0.0621

3.3 SEMANTIC OVERLAP IMAGE FILTERING

Experimental Setup To further substantiate the limitations mentioned in Section D, we evaluated AIM's filtering performance on images with semantically overlapping text. This experiment assesses the ability to perform filtering in situations where a specific concept (e.g., "Hello Kitty") shares semantic meaning with a generic term (e.g., "Kitty"). Using the same setup as Section D, 90 images were prepared (15 each for "Hello Kitty," "Kitty," "Mickey Mouse," "Mouse," "Donald Duck," and "Duck"). Grounding DINO was tasked with filtering the paired images using the specific IP character prompt (e.g., "Hello Kitty"). The AIM model was tested using a similarity threshold of 0.8.

Experimental Results As shown in Table 3, the AIM model demonstrated robust performance, distinguishing between images with similar textual meanings. In contrast, the DINO failed to filter these overlapping concepts effectively (e.g., detecting "Kitty" when searching for "Hello Kitty"). This failure is attributed to the inherent text-dependent nature of such detection models; while powerful for finding generic objects, they are unsuitable for filtering specific, specialized concepts. This result demonstrates that replacing traditional detectors with our AIM model is a more appropriate and effective approach for achieving safe inpainting.

4 Related Work

4.1 Concept Erasure in Diffusion Models

Despite the remarkable capabilities of Diffusion Models, they pose potential risks such as copyright infringement and the generation of harmful content (Schramowski et al., 2023). To mitigate these risks, Concept Erasure has emerged as a critical area of research.

Fine-tuning-based Erasure: Early approaches (Gandikota et al., 2023; Lu et al., 2024; Lyu et al., 2024; Gong et al., 2024; Hu et al., 2024) focused on suppressing specific concepts by directly modifying the model's parameters (e.g., cross-attention layers). While effective when an explicit text prompt triggers the concept, these methods exhibit a strong dependency on textual cues. Consequently, they are vulnerable to bypass attacks via prompt engineering and fail to enforce protection in tasks like image editing or inpainting, where the target concept is visually present but absent from the text prompt.

Guidance-based Erasure: To overcome text dependency, ACE (Wang et al., 2025) leverages the Classifier-Free Guidance (CFG) (Ho & Salimans, 2022) mechanism. ACE introduces a negative bias against the target concept into the unconditional prediction ($\epsilon_\theta(x_t, \emptyset)$). This design allows the model to suppress the concept during the generative process even without explicit prompts. Although ACE demonstrates superior defense in generation and editing, it remains susceptible to malicious inpainting, as discussed below.

4.2 Inpainting Attacks and Threat Model

Table 3: **Semantic overlap image filtering.** Best results are shown in **bold**.

	Kitty		Mouse		Duck	
Method	Succ. ↑	Miss ↓	Succ. ↑	Miss ↓	Succ. ↑	Miss ↓
DINO (Liu et al., 2024)	21	9	13	17	12	18
AIM_{80}	**25**	**5**	**29**	**1**	**16**	**14**

While previous methods focus on preventing the *generation* of concepts, we address a critical vulnerability in the *modification* phase, specifically within inpainting. We define the **Threat Model** for inpainting attacks as follows:

Attacker Goal: An adversary aims to utilize a safe, concept-erased model to output an image containing a forbidden concept (e.g., a copyrighted character) by leveraging an input image that already contains said concept.

Attack Mechanism (Context Leakage): Inpainting models are designed to preserve the semantic consistency of unmasked regions. An attacker can exploit this by providing a mask M that covers only a minor portion of the target concept or its background. In this scenario, the unmasked region $x \odot (1 - M)$ retains the global context and visual features of the forbidden concept. Since the diffusion model prioritizes minimizing the transition loss between the masked and unmasked regions, this strong visual condition s the erasure mechanisms (e.g., ACE's unconditional bias). We term this phenomenon *"Context Leakage,"* where the visual cues from the input bypass the model's safety constraints. Current erasure methodologies lack a verification process for such input-level leakage, necessitating a proactive countermeasure.

5 Conclusion

In this work, we identified a critical vulnerability in existing concept erasure mechanisms, where adversarial actors can bypass safety filters via localized inpainting attacks, and proposed **Adaptive Inpainting Masking (AIM)**, a training-free pipeline that proactively detects and masks forbidden concepts in the input image. Our key insight is to repurpose the *Unconditional Concept Erasure Bias*—originally designed for generation suppression—as a zero-shot detection signal: by isolating objects into an Out-of-Distribution (OOD) space and analyzing their reconstruction stability under biased inversion, AIM effectively distinguishes target from prior concepts. Experiments show that AIM significantly outperforms existing baselines against inpainting attacks across copyrighted characters, artistic styles, and NSFW content. By closing the "context leakage" loop, AIM provides a robust input-level defense for safer inpainting with diffusion models.

REFERENCES

Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing text-to-image models using semantic guidance, 2023. URL https://arxiv.org/abs/2301.12247.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023. URL https://arxiv.org/abs/2211.09800.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2301.13826.

Dar-Yen Chen, Hamish Tennent, and Ching-Wen Hsu. Artadapter: Text-to-image style transfer using multi-level style encoder and explicit adaptation, 2024a. URL https://arxiv.org/abs/2312.02109.

Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization, 2024b. URL https://arxiv.org/abs/2307.09481.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models, 2023. URL https://arxiv.org/abs/2303.07345.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models, 2024. URL https://arxiv.org/abs/2308.14761.

Hongcheng Gao, Tianyu Pang, Chao Du, Taihang Hu, Zhijie Deng, and Min Lin. Meta-unlearning on diffusion models: Preventing relearning unlearned concepts, 2025. URL https://arxiv.org/abs/2410.12777.

Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Reliable and efficient concept erasure of text-to-image diffusion models, 2024. URL https://arxiv.org/abs/2407.12383.

Luxi He, Yangsibo Huang, Weijia Shi, Tinghao Xie, Haotian Liu, Yue Wang, Luke Zettlemoyer, Chiyuan Zhang, Danqi Chen, and Peter Henderson. Fantastic copyrighted beasts and how (not) to generate them, 2025. URL https://arxiv.org/abs/2406.14526.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control, 2022. URL https://arxiv.org/abs/2208.01626.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Ying Hu, Chenyi Zhuang, and Pan Gao. Diffusest: Unleashing the capability of the diffusion model for style transfer, 2024. URL https://arxiv.org/abs/2410.15007.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models, 2023. URL https://arxiv.org/abs/2210.09276.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. URL https://arxiv.org/abs/2304.02643.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion, 2023. URL https://arxiv.org/abs/2212.04488.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2024. URL https://arxiv.org/abs/2303.05499.

Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models, 2024. URL https://arxiv.org/abs/2403.06135.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models, 2022. URL https://arxiv.org/abs/2201.09865.

Mengyao Lyu, Yuhong Yang, Haiwen Hong, Hui Chen, Xuan Jin, Yuan He, Hui Xue, Jungong Han, and Guiguang Ding. One-dimensional adapter to rule them all: Concepts, diffusion models and erasing applications, 2024. URL https://arxiv.org/abs/2312.16145.

Vladimir Mandic, Jared Weisinger, and wacdev. Nudenet: Nsfw object detection for tfjs and nodejs. https://github.com/notAI-tech/NudeNet, 2021. Accessed: 2025-11-10.

Hayk Manukyan, Andranik Sargsyan, Barsegh Atanyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Hd-painter: High-resolution and prompt-faithful text-guided image inpainting with diffusion models, 2024. URL https://arxiv.org/abs/2312.14091.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models, 2022. URL https://arxiv.org/abs/2211.09794.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models, 2023. URL https://arxiv.org/abs/2211.05105.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. URL https://arxiv.org/abs/2010.02502.

Luming Tang, Nataniel Ruiz, Qinghao Chu, Yuanzhen Li, Aleksander Holynski, David E. Jacobs, Bharath Hariharan, Yael Pritch, Neal Wadhwa, Kfir Aberman, and Michael Rubinstein. Realfill: Reference-driven generation for authentic image completion. *ACM Transactions on Graphics*, 43 (4):1–12, July 2024. ISSN 1557-7368. doi: 10.1145/3658237. URL http://dx.doi.org/10.1145/3658237.

Zhouxia Wang, Xintao Wang, Liangbin Xie, Zhongang Qi, Ying Shan, Wenping Wang, and Ping Luo. Styleadapter: A unified stylized image generation model, 2024. URL https://arxiv.org/abs/2309.01770.

Zihao Wang, Yuxiang Wei, Fan Li, Renjing Pei, Hang Xu, and Wangmeng Zuo. Ace: Anti-editing concept erasure in text-to-image models, 2025. URL https://arxiv.org/abs/2501.01633.

Peng Xing, Haofan Wang, Yanpeng Sun, Qixun Wang, Xu Bai, Hao Ai, Renyuan Huang, and Zechao Li. Csgo: Content-style composition in text-to-image generation, 2024. URL https://arxiv.org/abs/2408.16766.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2308.06721.

Yongsheng Yu, Hao Wang, Tiejian Luo, Heng Fan, and Libo Zhang. Magic: Multi-modality guided image completion, 2023. URL https://arxiv.org/abs/2305.11818.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018. URL https://arxiv.org/abs/1801.03924.

A ALGORITHM

The proposed Adaptive Inpainting Masking (AIM) pipeline is designed to preemptively defend against inpainting attacks by adaptively refining the user-provided mask. The process leverages the Unconditional Concept Erasure Bias embedded within an ACE model to detect the presence of a target concept without text labels.

Algorithm 1 Target-Agnostic Object Isolation (TAO)

Input: $\mathcal{G}_{DINO}$: Grounding DINO model
$\mathcal{M}_{SAM}$: Segment Anything Model (SAM)
I: Input image
$T_{agnostic}$: Target-agnostic text prompts (e.g., 'character', 'human', 'object')

Output:
$\mathcal{B} \leftarrow \mathcal{G}_{DINO}(I, T_{agnostic})$
if $\mathcal{B}$ is empty **then**
  $I_{iso} \leftarrow I$
  $\mathcal{B} \leftarrow \emptyset$
else
  $M_{obj} \leftarrow \emptyset$
  for each bounding box $B_i \in \mathcal{B}$ **do**
    $M_i \leftarrow \mathcal{M}_{SAM}(I, B_i)$
    $M_{obj} \leftarrow M_{obj} \cup M_i$
  end for
  $I_{iso} \leftarrow I \odot M_{obj}$
end if
return $I_{iso}, \mathcal{B}$

Algorithm 2 Concept erasure bias DDIM inversion

Input:
I_{iso}: Masked image generated in Algorithm 1
$\mathcal{E}$: VAE Encoder
$\mathcal{D}$: VAE Decoder
$\epsilon_\theta(z_t, \emptyset)$: ACE-trained model's unconditional prediction branch.
$\hat{\epsilon}_\theta(z_t, \emptyset)$: SD v1.4 model's unconditional prediction branch.

Output:
$z_0 \leftarrow \mathcal{E}(I_{iso})$
/*SD v1.4 inversion*/
$z_T \leftarrow$ DDIM_Inversion$(z_0, \hat{\epsilon}_\theta(z_t, \emptyset))$
/*ACE denoise*/
$z_0' \leftarrow$ DDIM_Denoise$(z_T, \epsilon_\theta(z_t, \emptyset))$
$I_{recon} \leftarrow \mathcal{D}(z_0')$
return I_{recon}

A.1 TARGET-AGNOSTIC OBJECT ISOLATION(TAO)

The objective of Algorithm 1 is to remove superfluous background information and to focus the concept erasure bias on the principal objects within the image. First, the Grounding DINO model (Liu et al., 2024) ($\mathcal{G}_{DINO}$) is employed with target-agnostic text prompts $T_{agnostic}$, such as 'character', 'human', and 'object', to extract a set of bounding boxes $\mathcal{B}$ for the principal objects from the input image I. If no objects are detected, the original image is utilized as I_{iso}. When objects are detected, each bounding box B_i is supplied as a prompt to the Segment Anything Model (Kirillov et al., 2023) ($\mathcal{M}_{SAM}$) to acquire a precise segmentation mask M_i. A union of all masks, M_{obj}, is generated. This

Algorithm 3 Similarity-Based Mask Adaptation

Input:
τ: Similarity threshold
M_{user}: User-provided inpainting mask.
$\mathcal{F}_{DINO}$: DINOv2 feature extractor
M_i: Mask for B_i, $B_i \in \mathcal{B}$
M_{all}: Mask cover entire image
$\mathcal{B}$: Bounding boxes (Algorithm 1)
I_{iso}: Masked image (Algorithm 1)
I_{recon}: Reconstructed image (Algorithm 2)

if $M_B \neq \emptyset$ **then**
  for each bounding box $B_i \in \mathcal{B}$ **do**
    $\hat{I}_{iso} \leftarrow Crop(I_{iso}, B_i)$
    $\hat{I}_{recon} \leftarrow Crop(I_{recon}, B_i)$
    $f_{iso} \leftarrow \mathcal{F}_{DINO}(\hat{I}_{crop})$
    $f_{recon} \leftarrow \mathcal{F}_{DINO}(\hat{I}_{recon})$
    $S \leftarrow \frac{f_{iso} \cdot f_{recon}}{||f_{iso}|| \cdot ||f_{recon}||}$
    if $S > \tau$ **then**
      $M_{adapt} \leftarrow M_{user} \cup M_i$
    else
      $M_{adapt} \leftarrow M_{user}$
    end if
  end for
else
  $f_{iso} \leftarrow \mathcal{F}_{DINO}(I_{iso})$
  $f_{recon} \leftarrow \mathcal{F}_{DINO}(I_{recon})$
  $S \leftarrow \frac{f_{iso} \cdot f_{recon}}{||f_{iso}|| \cdot ||f_{recon}||}$
  if $S > \tau$ **then**
    $M_{adapt} \leftarrow M_{user} \cup M_{all}$
  else
    $M_{adapt} \leftarrow M_{user}$
  end if
end if
return M_{adapt}

composite mask is then applied to the original image I to create the isolated image I_{iso} by setting all background pixels to 0. This process serves to remove background noise and intentionally transform the image into an 'Out-Of-Distribution' (OOD) sample. The OOD transformation induces the model to rely more strongly on the learned bias rather than the observed pixel values 8, which in turn amplifies the distortion of prior concepts.

A.2 Concept erasure bias DDIM inversion

Algorithm 2 leverages the concept erasure bias injected into the ACE model (Wang et al., 2025). The isolated image I_{iso}, generated during the TAO stage, is first encoded into a latent vector z_0 via the VAE encoder $\mathcal{E}$. Subsequently, deterministic DDIM inversion is performed using the unconditional predictor $\hat{\epsilon}_\theta(z_t, \emptyset)$ of the standard Stable Diffusion v1.4 model to obtain the noise map z_T. This noise map is then processed through DDIM denoising, guided only by the ACE-trained unconditional predictor $\epsilon_\theta(z_t, \emptyset)$, to generate the reconstructed latent vector z'_0. Finally, the resulting latent z'_0 is decoded into the pixel-space image I_{recon} via the VAE decoder ($\mathcal{D}$). As a result of this biased reconstruction process, if the target concept is present, the visual features of I_{iso} align with the ACE model's unconditional bias. This alignment causes the bias to reinforce the reconstruction, resulting in a stable and high-fidelity reconstruction.

$$I_{recon} \approx I_{iso} \tag{16}$$

Conversely, if the target concept is absent, the features of I_{iso} conflict with the bias. This mismatch is further amplified by the OOD transformation from the TAO stage, causing the bias to act as a corrupting direction and leading to a failed, distorted reconstruction.

$$I_{recon} \not\approx I_{iso} \tag{17}$$

A.3 Similarity-Based Mask Adaptation

Algorithm 3 details the process of quantitatively measuring the success of the biased reconstruction to determine the presence of C_{target}. This process is bifurcated into two paths, contingent on whether objects were detected during the TAO stage. If objects are detected, and one or more bounding boxes $\mathcal{B}$ were identified, the algorithm iterates through each B_i. For each object, both the original I_{iso} and the reconstructed I_{recon} are cropped to the bounding box B_i. The DINOv2 ($\mathcal{F}_{DINO}$) feature extractor is then employed to extract text-free visual features, f_{iso} and f_{recon}, from this pair of cropped images. The cosine similarity S between these features is then computed. If S exceeds a predefined threshold τ, the object is determined to be the target concept C_{target}, and its corresponding mask M_i is forcibly merged with the user-provided mask M_{user}.

$$M_{adapt} \leftarrow M_{user} \cup M_i \tag{18}$$

If no objects were detected during the TAO stage, the similarity comparison is performed on the entire I_{iso} and I_{recon} images. If this global similarity score S exceeds τ, the image is deemed to contain the erasure concept, and a mask covering the entire image, M_{all}, is merged with M_{user}. In all cases where $S \leq \tau$, the reconstruction is considered to have failed, and the original M_{user} is returned unmodified.

B Implementation Details

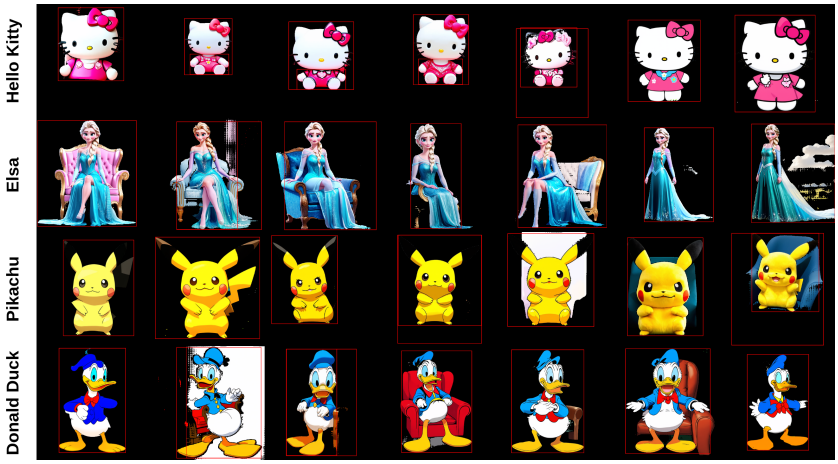

Figure 7: **Qualitative results of TAO.**

B.1 Configuration

The implementation was trained with the identical configuration following the ACE paper. When performing Concept Erasure Bias DDIM inversion, full inversion and denoising were applied over 50 steps. Inversion is performed on the images using the SD v1.4 model, and denoising is subsequently applied to the inverted images utilizing the ACE model.

B.2 Grounding DINO Settings for TAO

We utilize the Grounding DINO model with a specific set of prompts designed for generality. In the experiments involving inpainting of IP characters, Grounding DINO was instructed to detect the

relevant object within the image by providing a neutral prompt, specifically 'character'. The box threshold and text threshold were set to 0.3, and all regions recognized as bounding boxes were returned. Subsequently, the bounding boxes of these regions were provided to the SAM model to remove the background in all areas excluding the object. Figure 7 demonstrates the results of object segmentation executed by applying the TAO methodology, confirming that the object is correctly separated in the majority of cases.

B.3 Masking Strategy

For the experiments involving the application of a local mask to the object, a masking strategy was employed where the object's location was extracted using the Grounding DINO model, as utilized in Section B.2, and a portion of the corresponding bounding box was masked. In the main paper, the experiment was conducted by masking the upper 0.2 fraction of the entire bounding box to specifically mask the region corresponding to the head, and this demonstrated that the other baseline models lack defense capabilities against inpainting.

C Ablation Study

Table 4: **Ablation study of proposed method.** The best result is highlighted with **bold**.

	Erase Concept		Prior Concept		Overall	
Method	$\text{CLIP}_c \uparrow$	DINO $\uparrow$	$\text{CLIP}_c \downarrow$	DINO $\downarrow$	$\text{CLIP}_c \uparrow$	DINO $\uparrow$
W/O TAO	31.04	**0.8644**	**27.30**	0.6744	3.7454	0.1900
AIM	**31.22**	0.7856	27.38	**0.5812**	**3.8387**	**0.2044**

To demonstrate the critical role of **Target-Agnostic Object Isolation (TAO)**, we conducted a comparative analysis between performing Concept Erasure DDIM Inversion on raw original images versus images processed via TAO. This experiment was executed across all ten concepts utilized in the main evaluation.

We assessed the reconstruction consistency using two metrics: (1) CLIP_c score, which measures the semantic proximity of the transformed image to the concept text, and (2) **DINO Similarity**, which quantifies the visual fidelity to the original input using DINOv2 features.

As presented in Table 4, shifting images into the Out-of-Distribution (OOD) region via TAO forces the model to rely more heavily on the injected unconditional bias during the inversion process. The results indicate a pivotal trade-off: while the DINO similarity for the **Target Concept** decreased slightly due to the OOD shift, the similarity for the **Prior Concept** dropped significantly. Crucially, this creates a much distinct separation margin between the Target and Prior concepts. By suppressing the reconstruction of non-target concepts through OOD transformation, TAO effectively minimizes false negatives and ensures robust performance against inpainting attacks.

D NSFW filtering

Experimental Setup When the inpainting mask does not cover the target object, defense is fundamentally impossible for all baseline models. To verify the effectiveness of the AIM model in filtering NSFW images, a comparison was made against the VLM model, Grounding DINO. For DINO, a threshold of 0.5 was set, and filtering was considered successful if at least one bounding box was detected using the prompts "nudity," "nude," "nude body," and "sexual." The AIM model was trained with "nudity" as the target concept, and detection was deemed successful if the similarity score exceeded 0.7. Images were generated using the I2P dataset (Schramowski et al., 2023). To ensure only explicitly NSFW images were used, filtering was conducted only on images classified by Nudenet (Mandic et al., 2021) with the labels: "BUTTOCKS EXPOSED," "FEMALE BREAST EXPOSED," "FEMALE GENITALIA EXPOSED," "MALE GENITALIA EXPOSED," or "ANUS EXPOSED."

Table 5: **Quantitative comparisons of NSFW filtering.** The best result is highlighted with **bold**.

Method	Success ↑	Miss ↓	Rate ↑
DINO_{nudity} (Liu et al., 2024)	5	273	1.7
DINO_{nude} (Liu et al., 2024)	1	277	0.3
DINO_{sexual} (Liu et al., 2024)	52	226	18.7
$\text{DINO}_{nudebody}$ (Liu et al., 2024)	89	189	32.0
AIM_{70}	**230**	**48**	**82.7**

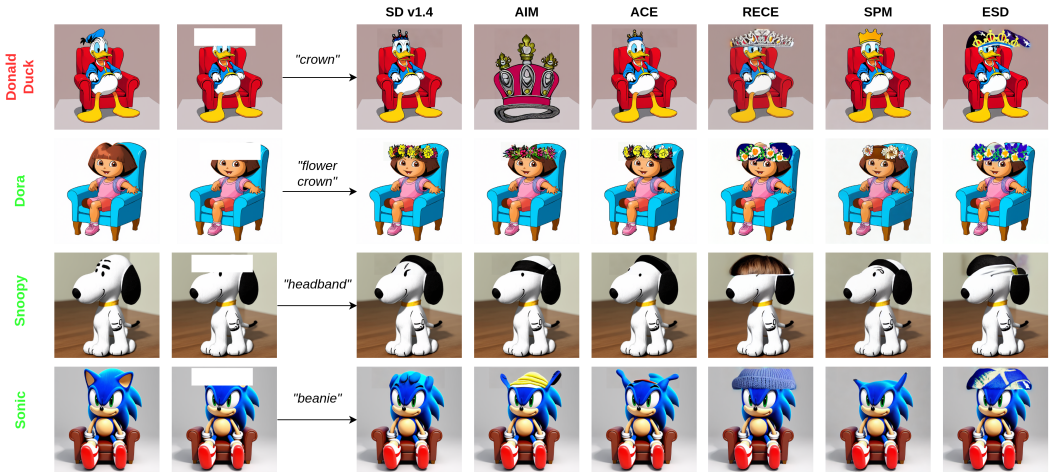

Figure 8: **Qualitative results of IP concept erasure.**

E More Experiments Results

E.1 Prompts and Results

The generative capability for IP concept erasure was evaluated in the main paper by inputting ten specific prompts. The specific prompts utilized were comprised of ”beanie,” ”beret,” ”crown,” ”flower crown,” ”goggles,” ”headband,” ”headphones,” ”helmet,” ”visor,” and ”wig,” through which the defense performance against inpainting was verified. Figures 8 to 10 present additional qualitative results concerning the IP character erasure performed in the main paper, and Figures 11 to 13 show the qualitative results obtained by applying the ”belt” prompt after masking the central region of the character for AIM, and those obtained by applying ”boots,” ”sandals,” ”sneakers,” and ”socks” after masking the lower region of the character.

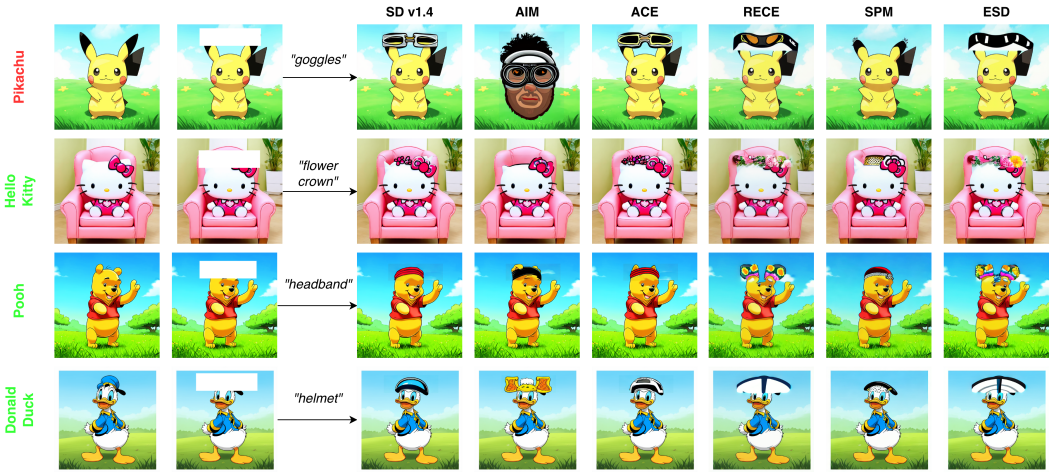

Figure 9: **Qualitative results of IP concept erasure.**

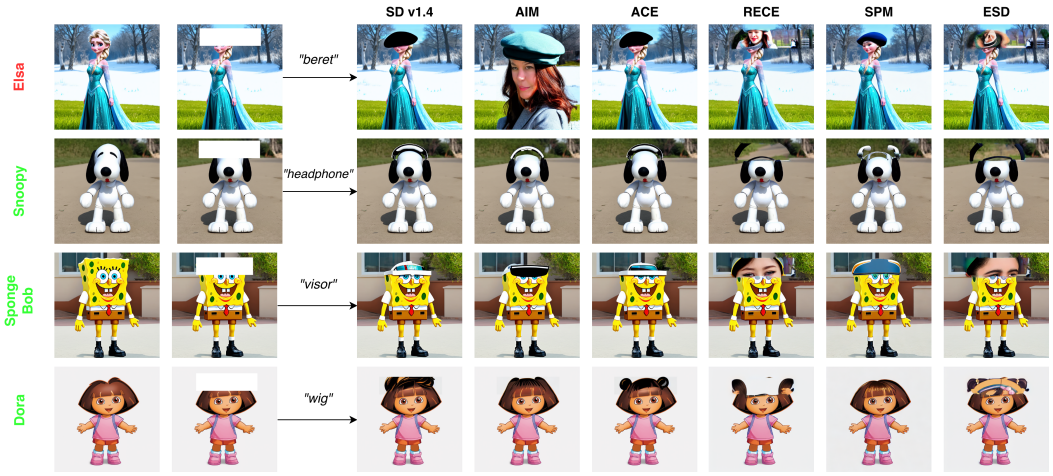

Figure 10: **Qualitative results of IP concept erasure.**

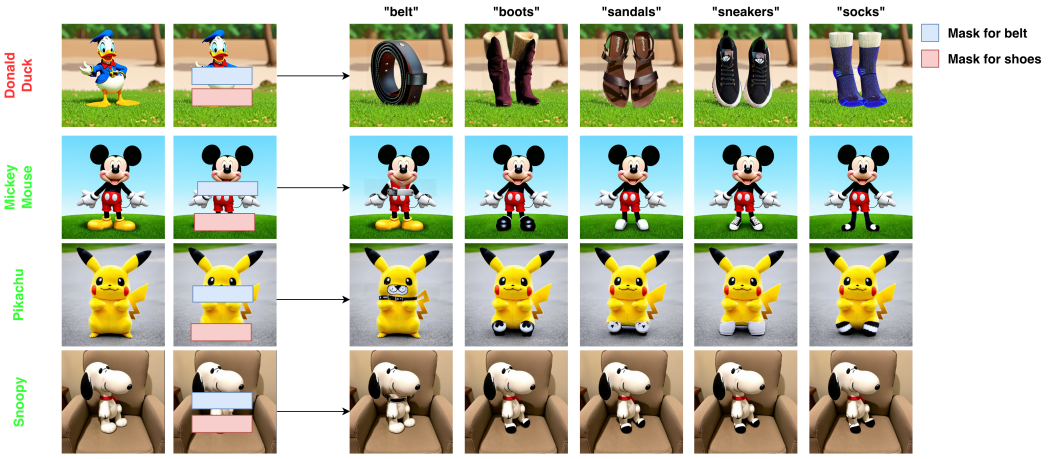

Figure 11: **Qualitative results of IP concept erasure.**

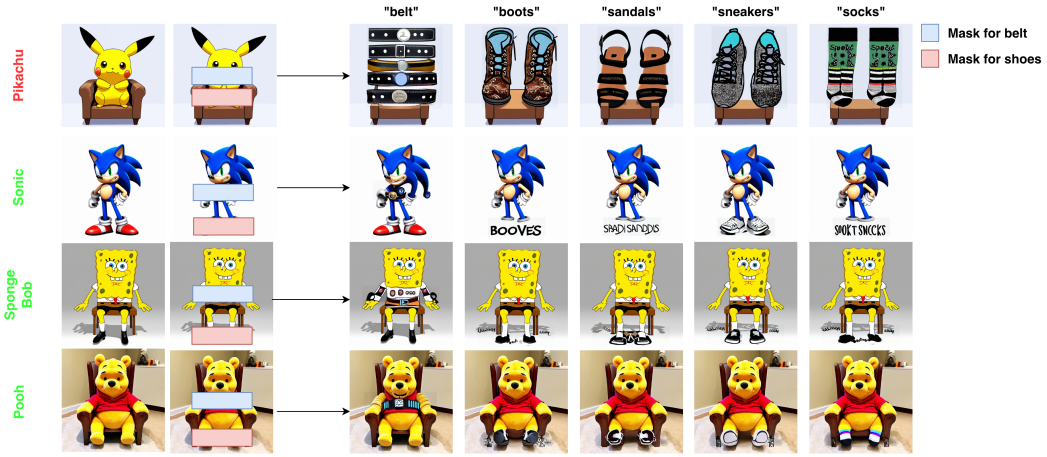

Figure 12: **Qualitative results of IP concept erasure.**

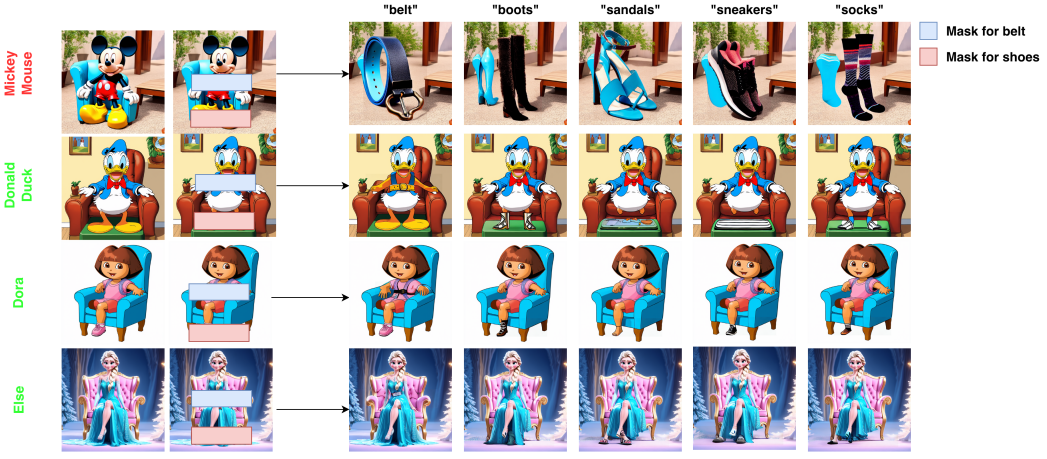

Figure 13: **Qualitative results of IP concept erasure.**