arXiv:2508.11904v1 [cs.CV] 16 Aug 2025

SafeCtrl: Region-Based Safety Control for Text-to-Image Diffusion via Detect-Then-Suppress

Lingyun Zhang[1], **Yu Xie**[2*], **Yanwei Fu**[1], **Ping Chen**[1]
[1]Fudan University
[2]Purple Mountain Laboratories
lyzhang22@m.fudan.edu.cn, xieyu9332@gmail.com, yanweifu@fudan.edu.cn, pchen@fudan.edu.cn

Abstract

The widespread deployment of text-to-image models is challenged by their potential to generate harmful content. While existing safety methods, such as prompt rewriting or model fine-tuning, provide valuable interventions, they often introduce a trade-off between safety and fidelity. Recent localization-based approaches have shown promise, yet their reliance on explicit "concept replacement" can sometimes lead to semantic incongruity. To address these limitations, we explore a more flexible detect-then-suppress paradigm. We introduce SafeCtrl, a lightweight, non-intrusive plugin that first precisely localizes unsafe content. Instead of performing a hard A-to-B substitution, SafeCtrl then suppresses the harmful semantics, allowing the generative process to naturally and coherently resolve into a safe, context-aware alternative. A key aspect of our work is a novel training strategy using Direct Preference Optimization (DPO). We leverage readily available, image-level preference data to train our module, enabling it to learn nuanced suppression behaviors and perform region-guided interventions at inference without requiring costly, pixel-level annotations. Extensive experiments show that SafeCtrl significantly outperforms state-of-the-art methods in both safety efficacy and fidelity preservation. Our findings suggest that decoupled, suppression-based control is a highly effective and scalable direction for building more responsible generative models.

Introduction

Text-to-image diffusion models (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2021; Rombach et al. 2022a; Saharia et al. 2022) have revolutionized visual content generation, yet their widespread deployment is critically hampered by their propensity to generate harmful content, such as nudity, violence, or biased imagery. Unsafe Diffusion(Qu et al. 2023) shows that a non-negligible proportion of outputs from popular diffusion models (e.g., up to 18.92% for Stable Diffusion) contain harmful elements. This challenge (Birhane, Prabhu, and Kahembwe 2021) largely stems from unsafe data inherited from web-scale training corpora (Schuhmann et al. 2022, 2021).

To address this, prevailing safety interventions have emerged, but they typically fall into two categories, both facing a fundamental dilemma. The first category, operating on the input/output level (eg, prompt rewriting(Yang et al. 2024; Wu et al. 2024) or rejection sampling (Rando et al. 2022)), acts as a brittle outer shell, often sacrificing the user's creative intent and being easily circumvented. The second, more robust category, intervenes at the model level (eg, fine-tuning (Gandikota et al. 2023; Schramowski et al. 2023; Wu et al. 2025)). While more effective, these intrusive methods often result in a degradation of image quality or a loss of semantic coherence. This recurring trade-off, we believe, is a direct consequence of the tight cou-

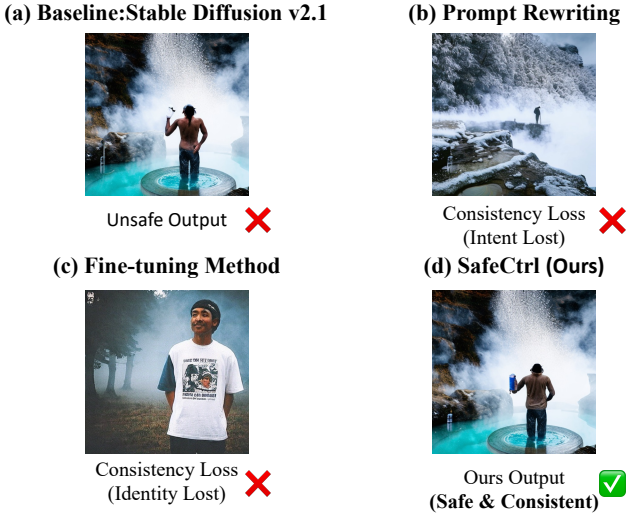


Figure 1: **Comparison of safety interventions on an artistic prompt.** The original prompt was *"An artist rinsing off sweat in a snowy mountain hot spring, with mist swirling around"*. **(a)** The baseline generates **explicit content** despite capturing the artistic intent. **(b) Prompt Rewriting** adds overhead and rewrites the prompt to "...taking pictures...fully clothed...", causing a catastrophic loss of the original scene and intent. **(c) Fine-tuning** fails to maintain the subject's identity. **(d)** In contrast, our non-intrusive **SafeCtrl** is the only method shown to produce a safe, consistent, and faithful output, demonstrating a superior solution to the safety-fidelity trade-off.

*Corresponding author.

pling between the safety mechanism and the core generative process. Recent modular approaches, like Concept Replacer (CR) (Zhang et al. 2025), have taken a promising step towards decoupling by localizing and replacing concepts. While effective, their "hard replacement" paradigm can still introduce semantic artifacts and lacks adaptability. This suggests that a more elegant and effective paradigm may lie not in replacement, but in suppression.

Therefore, we introduce **SafeCtrl**, a framework that pioneers a detect-then-suppress paradigm for safety control. SafeCtrl operates as a lightweight, non-intrusive plugin that works in synergy with a frozen, pretrained diffusion model. It first *detects* and localizes unsafe content using a novel unsafe attention module. Then, instead of performing a hard A-to-B substitution, it *suppresses* the harmful semantics, allowing the generative process to naturally and coherently resolve into a safe alternative that respects the original context. This is achieved by dynamically modulating the input text embedding, effectively neutralizing the influence of unsafe concepts only in the localized regions.

A key innovation lies in our training and inference strategy for this nuanced control. While prior works like Safety-DPO (Liu et al. 2024) have applied DPO to globally fine-tune the entire model, this can inadvertently affect safe regions of an image. We introduce a novel approach: we leverage Direct Preference Optimization (DPO) on readily available, image-level preference data to train our lightweight, external control module. Crucially, the fine-grained, region-guided suppression is applied only at inference time. This unique decoupling of global training from local intervention allows SafeCtrl to learn nuanced safety behaviors without requiring costly, pixel-level annotations, and ensures that modifications are precisely targeted, leaving safe areas entirely untouched.

Our contributions are summarized as follows:

- We identify the "tight coupling" of safety and generation as a core problem and propose a novel detect-then-suppress paradigm as a more flexible, decoupled solution.
- We instantiate this paradigm in SafeCtrl, a lightweight, non-intrusive plugin featuring a novel unsafe attention module for real-time detection and a guidance mechanism for targeted suppression.
- We propose a novel training strategy that leverages image-level DPO to learn region-guided control, eliminating the need for expensive pixel-level supervision.
- We demonstrate through extensive experiments that SafeCtrl significantly outperforms a wide range of state-of-the-art methods in both safety efficacy and fidelity preservation.

Related Work

Safety Control via Localization and Intervention

Controlling harmful content generation(Somepalli et al. 2023) in diffusion models(Rombach et al. 2022b; Ramesh et al. 2021) is a critical research area. Early approaches relied on brittle input/output filters (Rando et al. 2022). A more robust line of work intervenes directly in the generative process. One major category involves **global model editing**, where methods like ESD (Gandikota et al. 2023), UCE (Gandikota et al. 2024), and SLD (Schramowski et al. 2023) attempt to "erase" or steer away from unsafe concepts by fine-tuning the entire model, particular neurons(Yang, Cao, and Xu 2024) or lora(Hu et al. 2022). While often effective, these intrusive techniques risk degrading the model's general capabilities and content consistency.

Many studies(Chen, Tennent, and Hsu 2024; Zhang, Rao, and Agrawala 2023; Wang et al. 2023; Wu et al. 2023; Mou et al. 2024; Ye et al. 2023) have focused on how to control model generation. To enable more precise interventions, another category focuses on **region-based control**. These methods first localize a concept and then apply a targeted modification. In the context of general-purpose editing, works like DIFFEDIT (Couairon et al. 2022) have shown the potential of using attention maps for localization. For safety, Concept Replacer (CR) (Zhang et al. 2025) adopted this idea, using a heavyweight, duplicated U-Net(Ronneberger, Fischer, and Brox 2015) for segmentation and then performing a "hard replacement" of the concept.

Our approach builds upon the insights of region-based control but introduces a fundamentally different paradigm. Instead of performing intrusive global edits or relying on a costly external segmentation model, SafeCtrl operates as a lightweight, non-intrusive plugin. More importantly, we shift the intervention paradigm from CR's "hard replacement" to a more flexible "semantic suppression". This allows the model to naturally generate a safe and coherent alternative, rather than being forced to insert a potentially incongruous concept.

Preference Alignment for Generative Models

Aligning generative models with human preferences, rather than relying on explicit rules, has become a dominant paradigm, popularized by RLHF (Ouyang et al. 2022) and more recently, Direct Preference Optimization (DPO) (Rafailov et al. 2023). DPO has been successfully applied to align diffusion models with global properties like aesthetics (Wallace et al. 2024). The most relevant work, Safety-DPO (Liu et al. 2024), uses image-level preferences to globally fine-tune the entire model for safety.

Our approach proposes a novel application of DPO that bridges the gap between global supervision and local control. While Safety-DPO performs a global update that can inadvertently affect safe image regions, we are the first to use image-level DPO to train a decoupled, region-specific control module. Our framework learns a general safety alignment from holistic image preferences, while the precise, spatially-targeted suppression is applied only at inference time. This strategy bypasses the need for costly, pixel-level safety annotations, offering a more scalable and precise path to safety alignment.

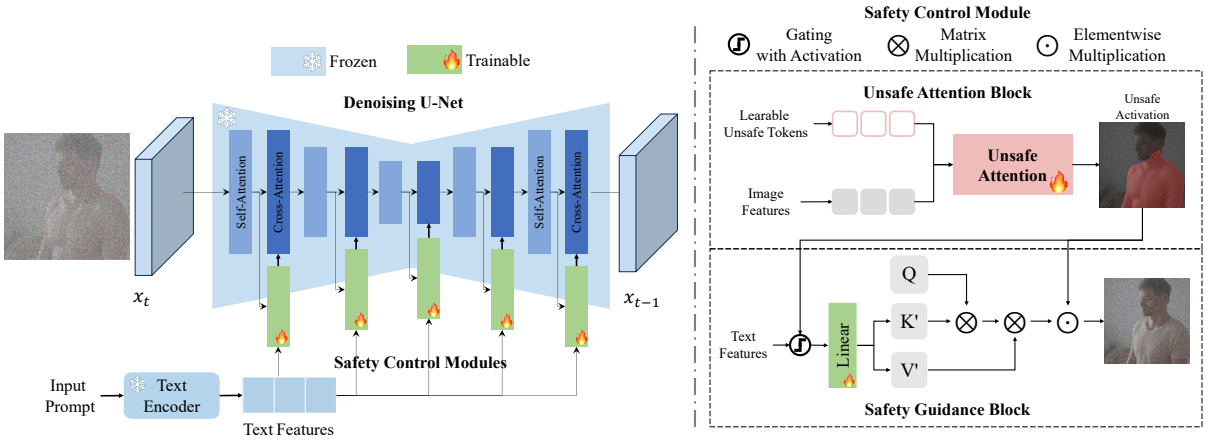


Figure 2: The overall architecture of SafeCtrl. SafeCtrl operates as a set of trainable, non-intrusive modules integrated into a frozen Denoising U-Net. At each step t, it receives the noisy latent x_t, text features, and the input prompt. The Safety Control Module is detailed in the Right. The *Unsafe Attention Block* uses learnable tokens to detect potentially unsafe regions, producing a spatial mask. Concurrently, the *Safety Guidance Block* processes text features and the localization information to generate a guidance signal, which modulates hidden states to steer the generation towards a safe and consistent output x_{t-1}.

Method

Preliminaries: Text-to-Image Diffusion Models

Text-to-image diffusion models(Dhariwal and Nichol 2021), such as Stable Diffusion (Rombach et al. 2022a), have demonstrated remarkable capabilities in generating high-fidelity images from textual prompts. These models operate by iteratively denoising a latent variable z_t over a series of timesteps $t = T, \ldots, 1$, starting from pure Gaussian noise $z_T \sim \mathcal{N}(0, I)$. At each step, a U-Net-based denoising network $\epsilon_\theta(z_t, t, c)$ predicts the noise to be removed from z_t, conditioned on the timestep t and an external context c.

For text-to-image generation, the context c is derived from the input prompt via a text encoder, eg, CLIP (Radford et al. 2021). The semantic information from c is then injected into the U-Net's intermediate layers, typically through cross-attention mechanisms (Vaswani et al. 2017). This allows the model to align the visual generation with the textual description. Formally, the cross-attention output for a set of visual feature queries Q and text feature keys K and values V is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \qquad (1)$$

where d_k is the dimension of queries. Despite their success, the reliance on massive, often unfiltered web-scale datasets for training makes these models susceptible to generating harmful or unsafe content that is implicitly associated with certain concepts in the data. Our work aims to address this challenge by introducing an external safety control mechanism that operates on this very cross-attention process, which we detail in the following sections.

The Detect-Then-Suppress Paradigm

To address the safety challenges outlined in §?? without compromising the integrity of the pretrained model, we introduce a novel safety control framework named **SafeCtrl**. The core philosophy of our approach is a paradigm shift from direct concept removal or replacement to a more flexible, two-stage process: **Detect-Then-Suppress**. As illustrated in Figure 2, SafeCtrl operates as an external, lightweight control system that works in synergy with the frozen U-Net. Our framework executes the following two stages at denoising:

1. **Detection.** An *Unsafe Attention Block* first analyzes the intermediate image features from the U-Net to identify and localize potentially harmful semantic regions. This produces a spatial risk map indicating "what" to suppress and "where".
2. **Suppression.** A *Safety Guidance Block* then takes this risk map and the original text features as input to compute a modified, "safe" conditioning signal. This signal is then used to modulate the U-Net's cross-attention process, effectively steering the generation away from the unsafe concept only in the identified regions.

This entire control mechanism is realized through a set of compact, trainable modules that are placed in parallel with the U-Net's attention layers, rather than being injected into them. This design ensures that our safety interventions are both precise and non-destructive. The detailed implementations of the detection and suppression stages are presented in following sections.

Unsafe Region Detection

The first stage of our framework is to accurately and efficiently localize potentially unsafe regions within the image as it is being generated. To achieve this, we introduce a lightweight **Unsafe Attention Block** that operates externally, in parallel with the frozen U-Net layers. Our key insight is that we can repurpose the rich, multi-level features from the U-Net as a powerful input to our specialized detec-

Method	Safety (on I2P)	Fidelity & Alignment (on I2P)		Fidelity & Alignment (on COCO-30k)	
	NudeNet Detections ↓	FID ↓	CLIP ↑	FID ↓	CLIP ↑
Base Model SD v2.1	586	-	31.65	14.87	31.53
FMN	424	23.74	28.89	13.52	30.39
AC	838	129.3	20.84	14.13	31.37
UCE	182	14.82	31.32	16.09	31.29
SLD-M	212	20.45	30.19	20.92	30.38
ESD-u	180	21.93	29.62	14.11	30.34
MACE	111	33.43	23.64	**13.42**	29.41
CR	126	17.88	30.68	15.15	30.67
Ours (SafeCtrl)	**55**	**12.14**	**31.54**	15.18	**31.30**

Table 1: **Main quantitative comparison of safety and fidelity.** Our method, SafeCtrl, is evaluated against state-of-the-art baselines. We report safety via NudeNet detections on the I2P benchmark (lower is better), and fidelity/alignment via FID (lower is better) and CLIP Score (higher is better) on both I2P and COCO-30k. The results show that SafeCtrl achieves the **best safety performance** (lowest detection count) while simultaneously maintaining fidelity scores that are the **closest to the original SD v2.1 baseline**, demonstrating a superior balance between the two competing objectives. A detailed breakdown of NudeNet detection categories is provided in the Appendix.

tion module, thus avoiding the need for a separate, heavyweight segmentation model like in CR (Zhang et al. 2025).

Spatial Risk Map Estimation. Our Unsafe Attention Block computes a spatial risk map by matching image features against a set of learnable "unsafe concept" prototypes. Specifically, we introduce a small set of learnable tokens, $\mathcal{T}_{\text{unsafe}}$, which are trained to act as general-purpose semantic embeddings for various unsafe concepts (eg, one token might learn to represent "nudity", another "violence").

At each denoising step t, the block operates as follows: First, the input image features from the U-Net are passed through a learnable projection head to generate the query tensor Q. Simultaneously, our learnable unsafe tokens $\mathcal{T}_{\text{unsafe}}$ are passed through a separate projection head to produce the key K and value V tensors.

The core of our detection mechanism is then a cross-attention operation between the image queries and the unsafe concept keys. The resulting attention map, A_{cross}, effectively represents the spatial distribution of unsafe semantics by measuring the alignment between each image region (Q) and the learned unsafe prototypes (K). A self-attention map, A_{self}, is also computed from the image features themselves to capture spatial context. These two maps are then fused to produce the final spatial risk mask M, which robustly highlights the unsafe regions:

$$M = \text{vec}(A_{\text{cross}}) \cdot A_{\text{self}}. \tag{2}$$

where $\text{vec}(\cdot)$ denotes the operation of flattening the cross-attention map into a vector.

Training Objective. Our training process is designed to be highly parameter-efficient. Instead of fine-tuning the entire U-Net, we only train the parameters within our Unsafe Attention Block: the set of learnable unsafe tokens $\mathcal{T}_{\text{unsafe}}$ and the lightweight projection heads for Q and K. This allows for rapid, few-shot adaptation to new concepts with minimal annotation and computational cost.

To effectively teach our module to identify unsafe regions, we employ a composite loss function guided by ground-truth masks M'. The total loss $\mathcal{L}_{\text{detect}}$ is a weighted sum of a cross-entropy loss and a Mean Squared Error (MSE) loss:

$$\mathcal{L}_{\text{detect}} = \lambda_{\text{ce}} \underbrace{\text{CE}(A_{\text{cross}}, M')}_{\mathcal{L}_{\text{CE}}:\text{Semantic Alignment}} + \lambda_{\text{mse}} \underbrace{\|M - M'\|_2^2}_{\mathcal{L}_{\text{MSE}}:\text{Spatial Refinement}}. \tag{3}$$

The cross-entropy term aligns our learnable tokens with the unsafe semantics in labeled regions, while the MSE term refines the final fused mask's spatial accuracy.

Temporal Smoothing at Inference. To ensure stable localization across the diffusion process and reduce noisy fluctuations between timesteps, we apply exponential smoothing to the predicted masks at inference time. The smoothed mask $\bar{M}_t$ used for suppression is computed as:

$$\bar{M}_t = \alpha \cdot \bar{M}_{t-1} + (1 - \alpha) \cdot M_t, \tag{4}$$

where M_t is the raw risk mask at timestep t, and we set $\alpha = 0.3$. This temporal consistency is crucial for effective and stable guidance in the suppression stage.

Suppression via Preference-Aligned Guidance

Once an unsafe region is identified, the **Safety Guidance Block** performs a targeted intervention to steer the generation towards a safe outcome. Our core idea is to modulate the cross-attention mechanism, which is responsible for injecting textual semantics into the image features. Instead of altering the U-Net's architecture, we dynamically compute a "safe" conditioning signal and apply it only within the detected unsafe regions, ensuring a minimally invasive yet effective suppression.

Inference-Time Suppression Mechanism. In a standard cross-attention layer (Eq. (1)), the image features f attend to the text features c to produce an output Z. Our guidance block introduces a parallel path. It first computes a set of

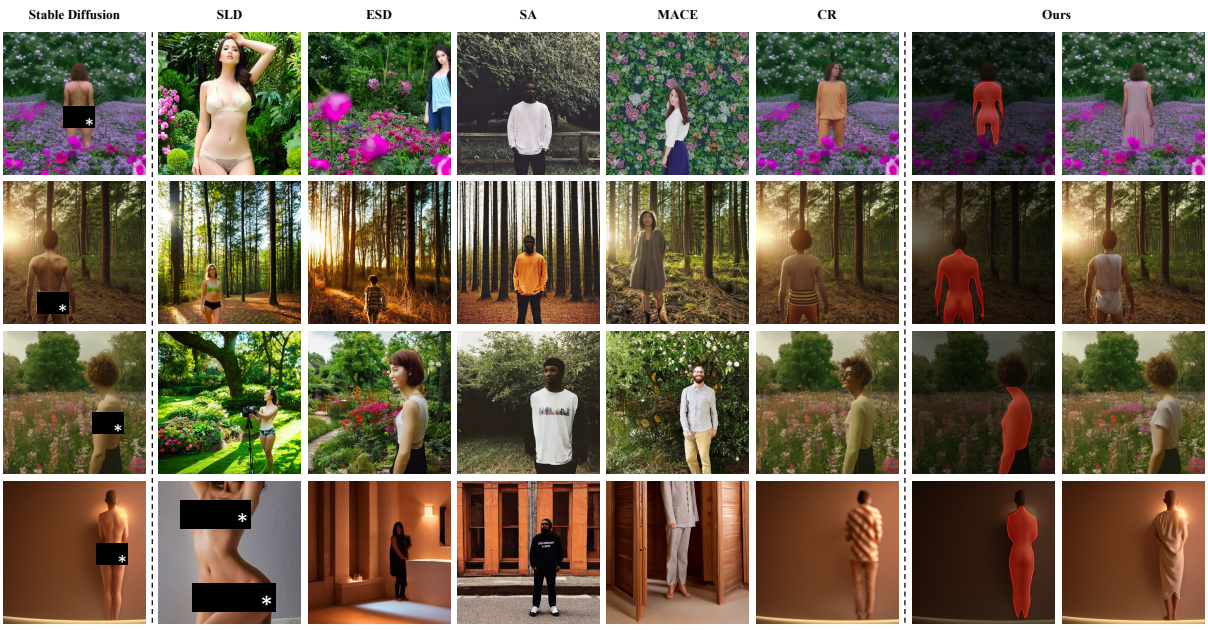


Figure 3: Qualitative comparison with state-of-the-art safety methods for nudity suppression. For each prompt, SafeCtrl is the only method that effectively removes the unsafe content while preserving high consistency with the original model's identity, background, and artistic style. In contrast, prior methods either fail to ensure safety or suffer from severe consistency loss and quality degradation.

"safe" text features by passing the original text embeddings c through a learnable projection layers, W'_K and W'_V, to get K' and V'. These are used to compute a parallel, safety-guided attention output Z':

$$Z' = \text{Softmax}\left(\frac{QK'^{\top}}{\sqrt{d}}\right)V', \tag{5}$$

where $Q = W_Q f$, $K' = W'_K c$, $V' = W'_V c$. Finally, we fuse the original attention output Z and our safety-guided output Z' using the smoothed risk map $\bar{M}$ from the detection stage. This fusion is performed via element-wise blending, confining the intervention only to the unsafe areas:

$$Z_{\text{new}} = Z \odot (1 - \bar{M}) \; + \; Z' \odot \bar{M}. \tag{6}$$

The resulting feature map Z_{new} is then passed to the next layer. This design ensures that the safety guidance is applied precisely where needed, preserving the original model's output in all safe regions.

Training with Preference Alignment via DPO. We train the lightweight projection layers W'_K and W'_V of the guidance block using Direct Preference Optimization (DPO) (Rafailov et al. 2023). Our key innovation is to leverage readily available, **image-level** preference pairs (y_w, y_l), where y_w is a preferred (safe) image and y_l is a rejected (unsafe) image. The DPO loss $\mathcal{L}_{\text{DPO}}$ trains our SafeCtrl-equipped model ϵ_θ to better denoise preferred examples and worse denoise rejected ones, compared to a frozen reference model ϵ_{ref} (see Appendix for the full loss equation).

This strategy is highly efficient as it decouples training from localization: the module learns a general-purpose safety alignment from global image preferences, while the precise, region-guided suppression (Eq. (6)) is applied only during inference. This avoids the need for costly, pixel-level safety annotations during preference tuning.

Experiments

We conduct a comprehensive set of experiments to rigorously evaluate our proposed SafeCtrl framework. Our evaluation is designed to demonstrate SafeCtrl's superiority in navigating the critical trade-off between Safety and Fidelity. To this end, we first present our main end-to-end results, then analyze the core detection module, and conclude with extensive ablation studies.

Experimental Setup

We implement SafeCtrl on top of the frozen Stable Diffusion v2.1 (Rombach et al. 2022a). Our evaluation is conducted on several standard benchmarks. **For localization accuracy**, we train our detection module in a few-shot setting (10 images per concept) and evaluate on CelebA-Mask-HQ, Pascal-Car, and Pascal-Horse test splits using mean Intersection over Union (mIoU). **For safety suppression**, we evaluate on the I2P benchmark (Schramowski et al. 2023) by measuring the detection rate of the NudeNet (Bedapudi 2019) classifier. **For fidelity and alignment**, we report Fréchet Inception Distance (FID) (Heusel et al. 2017) and CLIP Score (Radford et al. 2021) on both I2P and COCO-30k (Lin et al. 2014). We compare SafeCtrl against a comprehensive set of state-of-the-art methods, including SLD, ESD, CR, and others. **All implementation details, includ-**

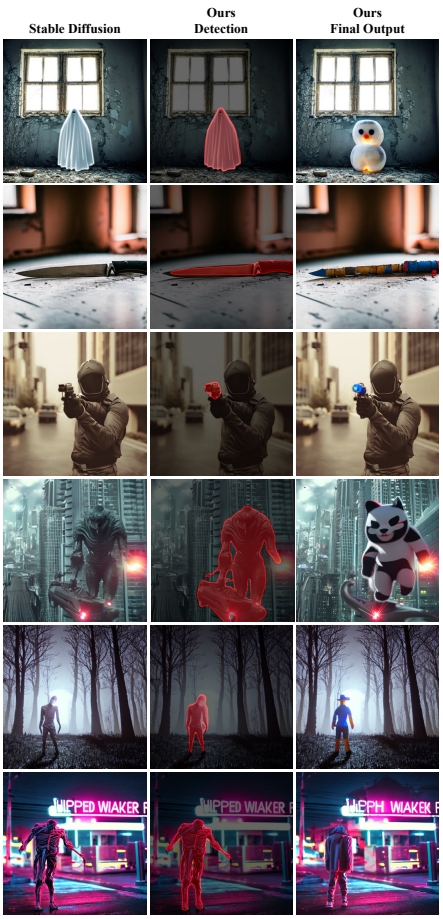


Figure 4: Qualitative results to other sensitive concepts. The final outputs highlight our method's ability to perform context-aware and plausible concept modification. By suppressing the unsafe semantics, SafeCtrl guides the generation towards coherent and creative alternatives (e.g., a ghost pivoting to a snowman, a monster concept being steered towards a cat, or a gun transforming into a toy).

ing dataset construction, hyperparameters, and the full list of baselines, are provided in the Appendix.

Main Results: Safety and Fidelity Comparison

We now present our main results, evaluating SafeCtrl against a comprehensive suite of state-of-the-art safety methods on the critical trade-off between safety and fidelity. Baselines include global fine-tuning methods (ESD, UCE, SA, MACE), guidance-based approaches (SLD), and Concept Replacer (CR). A detailed description of each baseline is provided in the Appendix.

Quantitative Analysis. Table 1 summarizes our quantitative results. **For safety**, we evaluate on the I2P benchmark, measuring the number of unsafe images generated (lower is better). SafeCtrl achieves the lowest detection count by a significant margin, demonstrating its superior suppression capabilities compared to all baselines. **For fidelity and alignment**, we report FID and CLIP scores on both I2P and the general-purpose COCO-30k benchmark. As shown, SafeCtrl consistently maintains scores that are closest to the original Stable Diffusion v2.1, indicating minimal impact on the model's generative distribution. This quantitatively validates that our method ensures safety without sacrificing image quality or semantic coherence.

Method	Pascal-Car	CelebA-HQ	Pascal-Horse
Supervised Baselines			
CNN+CRF	57.0	-	-
SegDDPM	-	78.0	58.7
Few-Shot / Unsupervised Methods			
ReGAN	52.2	69.9	54.3
SLiMe	68.3	75.7	63.3
CR	69.3	78.1	64.3
Ours (SafeCtrl)	**72.1**	**78.3**	**65.7**

Table 2: **Quantitative evaluation of localization accuracy (mIoU ↑).** We evaluate our detection module on three standard benchmarks under a 10-shot setting: Pascal-Car, CelebA-Mask-HQ, and Pascal-Pascal-Horse. Our method consistently outperforms state-of-the-art unsupervised and few-shot localization approaches. Detailed per-category results are in the Appendix.

Qualitative Analysis. Figure 3 provides a direct qualitative comparison. The visual evidence corroborates our quantitative findings and reveals the distinct failure modes of prior methods. For instance, SLD often fails to suppress nudity, while global editing methods like ESD and SA suffer from severe consistency loss, drastically altering the subject's identity. Even the stronger region-based method, CR, while preserving identity, can introduce semantically incoherent elements (e.g., pajamas in a garden). In contrast, **SafeCtrl** is the only method that consistently produces outputs that are both safe and highly faithful to the original prompt's identity, background, and artistic style. This visual superiority underscores the effectiveness of our detect-then-suppress paradigm.

Generalization to Other Concepts. Beyond nudity, SafeCtrl demonstrates strong generalization to a wider array of sensitive concepts, including horror themes (ghosts, monsters) and dangerous objects (knives, guns) in Figure 4. It successfully localizes and suppresses these varied concepts while maintaining contextual coherence. Detailed qualitative results for these concepts are presented in the Appendix.

Analysis of the Detection Module

To validate the effectiveness of our Unsafe Attention Block, we evaluate its localization accuracy both quantitatively and qualitatively. Since large-scale, pixel-annotated unsafe datasets are unavailable, we follow prior works (Zhang et al. 2025; Khani et al. 2023; Simsar et al. 2025) and assess performance on proxy segmentation tasks using standard benchmarks.

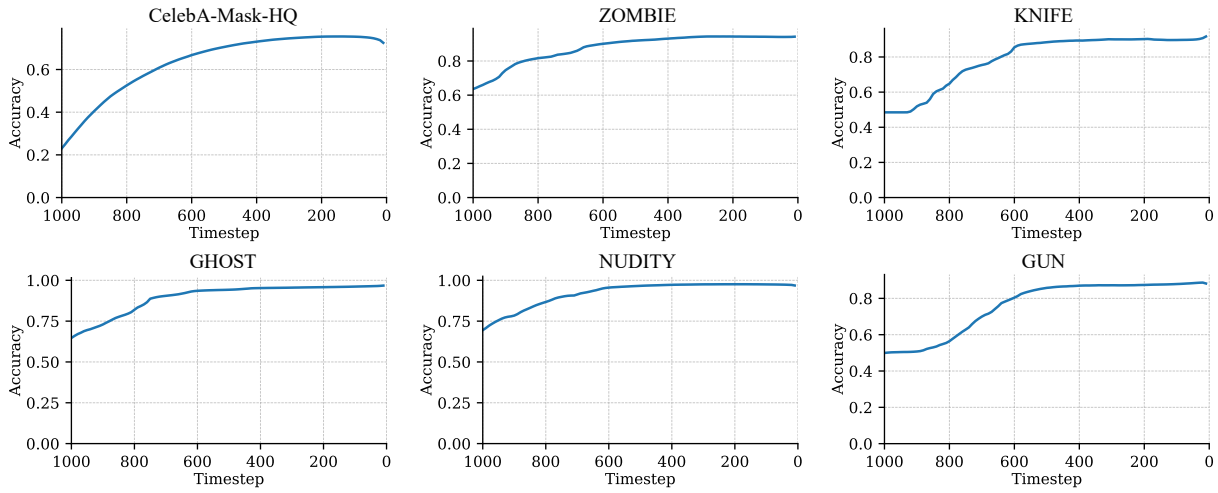


Figure 5: **Impact of denoising timestep on localization accuracy.** The mIoU for various concepts is plotted against the denoising timestep t. Accuracy consistently improves as the timestep t decreases, confirming that semantic concepts become more identifiable in the later, less noisy stages of generation.

Quantitative Accuracy. We compare SafeCtrl against state-of-the-art unsupervised and few-shot localization methods, including ReGAN(Tritrong et al. 2021), SLiMe(Khani et al. 2023), and the most relevant baseline, CR. All methods are trained in a consistent 10-shot setting for a fair comparison. The detailed setup and a full list of baselines are in the Appendix. As shown in Table 2, which reports the mean IoU, **SafeCtrl consistently outperforms all competing methods across all three benchmarks**: Pascal-Car, CelebA-Mask-HQ, and Pascal-Horse. This strong quantitative performance is particularly noteworthy given that our method achieves this with a lightweight, in contrast to CR which requires a heavyweight, duplicated U-Net. This validates our core hypothesis: repurposing the U-Net's rich features is a highly effective and parameter-efficient strategy for localization.

Qualitative Visualization. Fig 3 and Fig 4 provides qualitative evidence of our module's precision and generalization on actual sensitive concepts. The visualizations demonstrate that SafeCtrl can accurately localize a wide variety of targets, from large, deformable subjects like 'nudity' and 'monsters', to small, fine-grained objects like 'weapons'. This robust localization capability across diverse and challenging scenarios is the critical foundation that enables the success of our subsequent suppression stage. More examples are available in the Appendix.

Ablation Studies

We conduct a series of ablation studies to validate our key design choices. We first analyze the impact of the denoising timestep on detection accuracy, and then verify the effectiveness of our region-guided suppression mechanism.

Analysis of Detection Timestep. To justify our strategy of applying suppression primarily in the later stages of denoising, we first investigate at which timesteps unsafe concepts become semantically distinct. We evaluate the localization accuracy (mIoU) of our detection module across the entire denoising process ($t = 1000 \rightarrow 1$) on various concepts. As illustrated in Figure 5, for all concepts, the accuracy remains low during the initial noisy stages but rises sharply and stabilizes in the later stages (e.g., $t < 200$), where coherent image structures emerge. This result empirically validates our design choice to activate suppression in the later, more semantically stable part of the generation, ensuring that our interventions are based on reliable detections.

Effectiveness of Region-Guided Suppression. To isolate the benefit of our spatially-targeted intervention, we compare our full SafeCtrl model against a strong baseline variant: " w/o Region-Guidance". This variant uses the exact same DPO training on our plugin, but at inference time, the suppression is applied *globally* to the entire image, without using the spatial mask from our detection module. Table 3 shows the comparison. While the global suppression approach achieves a slightly lower NudeNet detection count, this comes at a significant cost to image fidelity, evidenced by its much worse FID and CLIP scores. Our full, region-guided SafeCtrl, in contrast, achieves a superior balance. It maintains high fidelity (FID 12.14, CLIP 31.54) by applying modifications only where necessary. This clearly demonstrates that our region-guided suppression is the critical component for preserving content consistency and avoiding the collateral damage often seen in global safety methods.

Conclusion

In this work, we introduced SafeCtrl, a novel framework designed to resolve the fundamental safety-fidelity trade-off in diffusion models. We identified the tight coupling of safety and generation in prior methods as the core problem and, in response, proposed a more flexible, decoupled detect-then-suppress paradigm. Our key innovation is a novel train-

Method	NudeNet ↓	FID (I2P) ↓	CLIP (I2P) ↑
w/o Region-Guidance	**48**	20.69	30.53
Ours (Full Model)	55	**12.14**	**31.54**

Table 3: **Ablation study on region-guided suppression.** While global suppression is slightly more aggressive on safety metrics, our region-guided approach achieves a significantly better trade-off, preserving much higher image fidelity and alignment.

ing strategy that leverages image-level Direct Preference Optimization to effectively train a region-specific control module, enabling precise, localized interventions without requiring costly pixel-level supervision. Extensive experiments validate that our lightweight plugin significantly outperforms state-of-the-art methods in achieving a superior balance between safety and fidelity. We believe our approach represents a significant and scalable step towards more controllable, interpretable, and trustworthy generative AI.

References

Bedapudi, P. 2019. Nudenet: Neural nets for nudity classification, detection and selective censoring.

Birhane, A.; Prabhu, V. U.; and Kahembwe, E. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*.

Chen, D.-Y.; Tennent, H.; and Hsu, C.-W. 2024. Artadapter: Text-to-image style transfer using multi-level style encoder and explicit adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8619–8628.

Couairon, G.; Verbeek, J.; Schwenk, H.; and Cord, M. 2022. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Gandikota, R.; Materzynska, J.; Fiotto-Kaufman, J.; and Bau, D. 2023. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2426–2436.

Gandikota, R.; Orgad, H.; Belinkov, Y.; Materzyńska, J.; and Bau, D. 2024. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5111–5120.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Khani, A.; Taghanaki, S. A.; Sanghi, A.; Amiri, A. M.; and Hamarneh, G. 2023. Slime: Segment like me. *arXiv preprint arXiv:2309.03179*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, R.; Chieh, C. I.; Gu, J.; Zhang, J.; Pi, R.; Chen, Q.; Torr, P.; Khakzar, A.; and Pizzati, F. 2024. SafetyDPO: Scalable Safety Alignment for Text-to-Image Generation. *arXiv preprint arXiv:2412.10493*.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; and Shan, Y. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 4296–4304.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Qu, Y.; Shen, X.; He, X.; Backes, M.; Zannettou, S.; and Zhang, Y. 2023. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. In *Proceedings of the 2023 ACM SIGSAC conference on computer and communications security*, 3403–3417.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36: 53728–53741.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831. Pmlr.

Rando, J.; Paleka, D.; Lindner, D.; Heim, L.; and Tramèr, F. 2022. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022a. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022b. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Schramowski, P.; Brack, M.; Deiseroth, B.; and Kersting, K. 2023. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22522–22531.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C. W.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; Schramowski, P.; Kundurthy, S. R.; Crowson, K.; Schmidt, L.; Kaczmarczyk, R.; and Jitsev, J. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Simsar, E.; Tonioni, A.; Xian, Y.; Hofmann, T.; and Tombari, F. 2025. Lime: localized image editing via attention regularization in diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 222–231. IEEE.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. pmlr.

Somepalli, G.; Singla, V.; Goldblum, M.; Geiping, J.; and Goldstein, T. 2023. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6048–6058.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Tritrong, N.; et al. 2021. Repurposing GANs for One-shot Semantic Part Segmentation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4473–4483. IEEE.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wallace, B.; Dang, M.; Rafailov, R.; Zhou, L.; Lou, A.; Purushwalkam, S.; Ermon, S.; Xiong, C.; Joty, S.; and Naik, N. 2024. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8228–8238.

Wang, Z.; Jiang, Y.; Lu, Y.; He, P.; Chen, W.; Wang, Z.; Zhou, M.; et al. 2023. In-context learning unlocked for diffusion models. *Advances in Neural Information Processing Systems*, 36: 8542–8562.

Wu, W.; Li, Z.; He, Y.; Shou, M. Z.; Shen, C.; Cheng, L.; Li, Y.; Gao, T.; Zhang, D.; and Wang, Z. 2023. Paragraph-to-image generation with information-enriched diffusion model. *arXiv preprint arXiv:2311.14284*.

Wu, Y.; Zhou, S.; Yang, M.; Wang, L.; Chang, H.; Zhu, W.; Hu, X.; Zhou, X.; and Yang, X. 2025. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 8496–8504.

Wu, Z.; Gao, H.; Wang, Y.; Zhang, X.; and Wang, S. 2024. Universal prompt optimizer for safe text-to-image generation. *arXiv preprint arXiv:2402.10882*.

Yang, T.; Cao, J.; and Xu, C. 2024. Pruning for robust concept erasing in diffusion models. *arXiv preprint arXiv:2405.16534*.

Yang, Y.; Gao, R.; Yang, X.; Zhong, J.; and Xu, Q. 2024. Guardt2i: Defending text-to-image models from adversarial prompts. *Advances in neural information processing systems*, 37: 76380–76403.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zhang, L.; Xie, Y.; Fu, Y.; and Chen, P. 2025. Concept replacer: Replacing sensitive concepts in diffusion models via precision localization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 8172–8181.