Beyond the Safety Tax: Mitigating Unsafe Text-to-Image Generation via External Safety Rectification

Anonymous ACL submission

Abstract

Text-to-image (T2I) generative models have achieved remarkable visual fidelity, yet remain vulnerable to generating unsafe content. Existing safety defenses typically intervene internally within the generative model, but suffer from severe concept entanglement, leading to degradation of benign generation quality—a trade-off we term the *Safety Tax*. To overcome this limitation, we advocate a paradigm shift from destructive internal editing to external safety rectification. Following this principle, we propose *SafePatch*, a structurally isolated safety module that performs external, interpretable rectification without modifying the base model. The core backbone of *SafePatch* is architecturally instantiated as a trainable clone of the base model's encoder, allowing it to inherit rich semantic priors and maintain representation consistency. To enable interpretable safety rectification, we construct a strictly aligned counterfactual safety dataset (ACS) for differential supervision training. Across nudity and multi-category benchmarks and recent adversarial prompt attacks, *SafePatch* achieves robust unsafe suppression (7% unsafe on I2P) while preserving image quality and semantic alignment.

1 Introduction

Text-to-image (T2I) generative models, exemplified by Stable Diffusion (Rombach et al., 2022) and DALL·E 3 (Goh et al., 2024), have revolutionized visual content creation by synthesizing high-fidelity images from natural language descriptions. Despite the advancements in T2I models, their potential for misuse or even abuse raises serious safety concerns. Recent studies (Schramowski et al., 2023; Rando et al., 2022) have demonstrated that the T2I models are prone to generating NSFW (Not Safe For Work) imagery, such as those related to violence and child-unsafe, when prompted with unsafe text prompts. Consequently, quantifying and mitigating T2I models' unsafe content generation become increasingly important research topics.

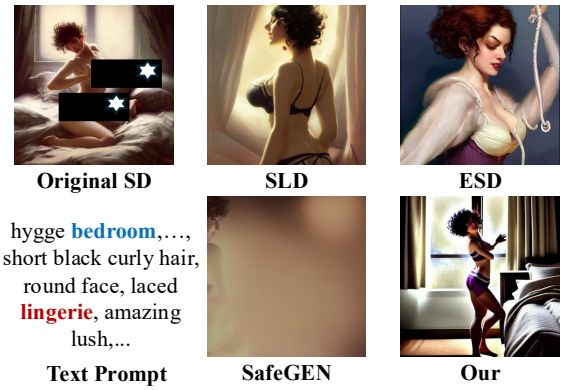

Figure 1: **Illustration of the safety tax caused by concept entanglement.** Existing defenses (e.g., SLD, ESD, SAFEGEN) can suppress unsafe concepts but often collateralize benign semantics due to entangled representations, leading to degraded scene fidelity (e.g., damaging the "bedroom" semantic).

Existing defense strategies generally fall into two categories. The first involves *filtering-based methods*, which employ classifiers to intercept unsafe images post-generation (Robin Rombach, Patrick Esser, 2023a) or sanitize malicious prompts at the input stage. While easy to deploy, these post-hoc defenses are often susceptible to circumvention by adversarial prompts and fail to address the model's internal propensity for generating harmful content. To address these limitations, recent research focus has shifted toward *internal defenses*. These approaches aim to intervene directly on the model's internal mechanisms, primarily through parameter editing (e.g., ESD (Gandikota et al., 2023)), which erases harmful concepts via fine-tuning, or inference guidance (e.g., SLD (Schramowski et al., 2023)), which steers image generation away from unsafe concepts during inference.

Motivation. Although internal defenses provide more robust protection, recent studies (Saha et al.,

2025; Amara et al., 2025) suggest that diffusion models suffer from intricate concept entanglement arising from feature superposition (Elhage et al., 2022). This implies that unsafe concepts are not isolated in the latent space but are tightly coupled with benign concepts. Thus, stripping away harmful concepts — whether through weight modification or inference-time intervention — inevitably incurs collateral damage on associated benign features (see Figure 1). We define the degradation in generative quality incurred by defenses as the *Safety Tax*.

Our Solution. To eliminate the safety tax, we advocate a paradigm shift from destructive internal editing to external safety rectification. Specifically, instead of disentangling unsafe concepts within a highly entangled base model, we externalize safety control into a structurally isolated and safety-specialized module, explicitly trained to provide interpretable and controllable safety rectification. Compared with directly improving the interpretability of internal concepts in the original T2I model, this design enables the safety module to focus entirely on safety-dimensional control without needing to handle the generative logic of other semantics; it only needs to remain "silent" when benign semantics are encountered. This characteristic significantly reduces the difficulty of learning interpretable safety rectification signals.

Following this principle, we propose *SafePatch*, a structurally isolated safety module designed to dynamically rectify the generative process without compromising the base model's integrity. Specifically, the core backbone of *SafePatch* is architecturally instantiated as a trainable clone of the base model's encoder part. This design provides two critical advantages: First, it ensures representation consistency with the base model, avoiding performance degradation caused by feature heterogeneity; Second, it inherits semantic priors from the base model, which lowers the difficulty of learning interpretable safety signals.

However, structural isolation alone is insufficient for interpretable safety rectification. To achieve this, we construct a strictly aligned counterfactual safety dataset (ACS). By preserving identical benign semantics while differing exclusively in safety semantics, this data provides high-fidelity differential supervision, forcing *SafePatch* to learn interpretable, safety-specific rectification signals. Furthermore, to distinguish safety rectification from natural generative variations, we design an instruction-aware spatial projection to convert abstract safety concepts into executable safety modification instructions, mapping them to spatially grounded features to precisely localize unsafe concepts while ignoring benign fluctuations. Finally, we integrate *SafePatch* with the base model via zero convolution layers, ensuring that the safety rectification is introduced as an initially non-intrusive way during training stage, thereby maximally preserving the backbone's generative quality.

We implement *SafePatch* and evaluate it against six representative safety defenses across multiple unsafe and benign benchmarks. *SafePatch* consistently suppresses unsafe content while preserving image quality and text–image alignment, indicating that safety improvements do not incur a safety tax. On the I2P benchmark, *SafePatch* reduces the overall unsafe probability to 7%, substantially outperforming all baselines, which remain around 20%. Moreover, *SafePatch* maintains low unsafe rates under recent adversarial prompt attacks, demonstrating robust and reliable safety rectification.

In summary, our contributions are as follows:

- We construct a strictly aligned counterfactual safety dataset (ACS) that provides paired samples with identical benign semantics but differing safety semantics, enabling high-fidelity differential supervision.
- We introduce a new defense paradigm that shifts from destructive internal editing to external safety rectification. Specifically, we design *SafePatch*, a structurally isolated safety module that achieves interpretable safety rectification for T2I models without performance degradation.
- Extensive evaluations across multiple benchmarks and adversarial attacks demonstrate that *SafePatch* significantly outperforms six state-of-the-art defenses while effectively eliminating the safety tax.

2 Background

2.1 Unsafe Content Generation in T2I Models

Text-to-image generation models have gained popularity due to their ease of use and high-quality, flexible images. However, Birhane et al. (Birhane et al., 2021) raise concerns about datasets scraped from the internet, such as LAION-400M (Schuhmann et al., 2022), which lack content moderation, potentially leading to unsafe content generation.

The definition of unsafe content varies by con-

text and culture, making it subjective. In this paper, we focus on images containing *hate*, *harassment*, *violence*, *self-harm*, *sexual content*, *shocking material*, *illegal activities*, or *nudity*, as outlined in the OpenAI content policy (ope, 2023) and Gebru et al. (Gebru et al., 2021).

2.2 Safety Mechanisms for T2I Models

Present strategies have two categories: filtering-based and internal defenses.

Filtering-based Defenses. Filtering-based defenses, like safety checker (Robin Rombach, Patrick Esser, 2023a) officially released by SD, are efficient for deployment but suffer from under-generalization and vulnerability to adversarial prompts due to distribution shifts. Similarly, SD v2.1 (Robin Rombach, Patrick Esser, 2023b) retrains on censored data with these filters, but this approach can be computationally expensive, may not fully remove harmful content, and could reduce model performance.

Internal Defenses. Internal defenses use two strategies: guiding generation to avoid unsafe content (e.g., SLD (Schramowski et al., 2023) and InterpretDiffusion (Li et al., 2024a)) or fine-tuning models to remove unsafe concepts (e.g., ESD (Gandikota et al., 2023) and SafeGEN (Li et al., 2024b)). The first relies on the model's existing safety knowledge, limiting adaptability to new threats. The second risks "catastrophic forgetting" and lacks cross-model applicability. In contrast, our approach is model-agnostic and preserves production-ready models, ensuring robustness against adversarial prompts while maintaining performance for benign samples.

2.3 Threat Model

We outline the goals and capabilities of the adversary and defender.

Adversary. The adversary aims to violate safety standards by generating unsafe content, either deliberately or by evading safeguards. We assume the adversary has closed-box access, capable only of querying the online T2I model with prompts.

Defender. The defender (model developer) has two objectives: (1) preventing unsafe content generation, and (2) preserving benign utility. We assume the defender has full access to model parameters to deploy safety mechanisms but lacks prior knowledge of specific adversarial prompts.

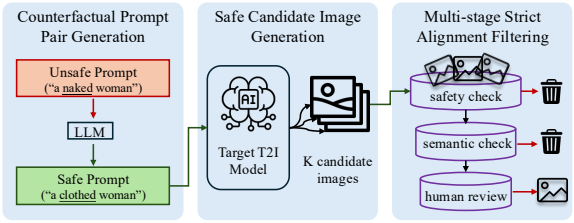

Figure 2: **Construction of the Aligned Counterfactual Safety (ACS) dataset.** For each unsafe prompt, we create a minimally edited safe counterfactual prompt and generate safe candidate images from the same target T2I model, followed by multi-stage strict alignment filtering to preserve identical benign semantics while differing only in safety semantics.

3 Training Dataset Construction

To achieve the interpretable safety rectification capability of *SafePatch*, we construct an aligned counterfactual safety dataset (ACS). Each sample pair in the dataset consists of an unsafe prompt and its corresponding safe image. Compared to the image originally generated by the unsafe prompt, these safe version images maintain consistency in benign semantics while differing only in unsafe semantics. This design provides *SafePatch* with high-fidelity differential supervision, enabling it to precisely learn safety rectification signals without disrupting the original benign features. The construction process of the dataset, as illustrated in Figure 2, includes the following key steps:

Counterfactual Prompt Pair Generation. We first collect unsafe prompts covering major safety risk categories (e.g., violence, hate, sexual) from Lexica[1]. To construct strictly counterfactual pairs, we utilize an LLM to generate corresponding safe prompts under the principle of minimal semantic differences (Liu et al., 2024). This process modifies only the phrases that violate safety policies, leaving the rest unchanged.

Safe Candidate Image Generation. For each sample pair, we use the target T2I model intended for defense to generate K candidate images for the safe prompt. This design aims to minimize distribution shifts and style differences introduced by model heterogeneity.

Multi-stage Strict Alignment Filtering. To ensure strict alignment in both safety and semantic fidelity for the final training samples, we design a three-stage filtering pipeline. First, we screen candidate images using multiple automated safety

[1] https://lexica.art/

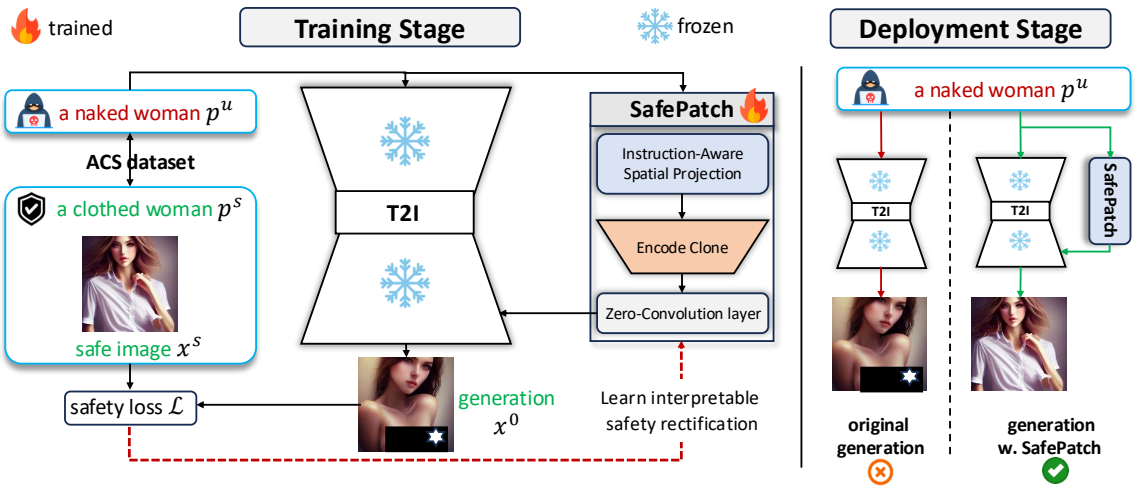

Figure 3: Overview of the proposed *SafePatch* framework, including training and deployment stages.

auditing tools (e.g., NudeNet (Bedapudi, 2019)) to ensure they contain no violating content; samples failing this check are directly discarded. Second, we utilize a Vision-Language Model (VLM) to automatically assess whether the generated safe image completely and faithfully reflects all benign semantics of its corresponding safe prompt, filtering out semantically inconsistent samples. Finally, we introduce a manual review process where annotators are asked to judge, based on the original unsafe prompt, whether the corresponding safe image can be considered an ideal safe version—that is, whether it completely removes harmful elements while perfectly preserving all other reasonable visual content and overall style. Only sample pairs that pass all filtering stages are included in the final training dataset.

4 Design of *SafePatch*

4.1 Overview

As illustrated in Figure 3, *SafePatch* is an external safety rectification module designed to eliminate the safety tax without modifying the base T2I model. It is implemented as a plug-and-play component $\mathcal{S}_\phi$ that operates alongside a frozen diffusion model ϵ_θ. *SafePatch* follows a two-stage paradigm. In the training stage, $\mathcal{S}_\phi$ is explicitly trained to learn interpretable safety rectification signals using a strictly aligned counterfactual safety dataset. In the deployment stage, the trained $\mathcal{S}_\phi$ is attached to the base model ϵ_θ as an external plugin, providing additive safety rectification while remaining "silent" for benign semantics.

At deployment time, the denoising prediction at timestep t is given by:

$$\tilde{\epsilon} = \epsilon_\theta(z_t, t, c_p) + \mathcal{S}_\phi(z_t, t, c_p, s), \quad (1)$$

where z_t denotes the noisy latent, c_p is the prompt embedding, and s is the safety instruction.

4.2 *SafePatch* Architecture

Following the principle of external safety rectification, *SafePatch* is instantiated as a structurally isolated and safety-specialized module. Specifically, *SafePatch* comprises the following three synergistic components:

Trainable Encoder Clone. The core backbone of *SafePatch* is architecturally instantiated as a trainable clone of the encoder part (the down-sampling blocks of the U-Net) of the base model. This design ensures representation consistency with the base model, avoiding performance degradation caused by feature heterogeneity. Furthermore, by inheriting semantic priors from the base model, *SafePatch* further lowers lowers the difficulty of learning interpretable safety signals.

Zero-Convolution Integration. The rectification outputs of *SafePatch* are injected into the base model through zero-initialized convolution layers, inspired by the design in (Zhang et al., 2023). At initialization, these layers guarantee that *SafePatch* introduces zero influence on the denoising process. During training, rectification signals are introduced in a gradual and additive manner. This design prevents abrupt perturbations to the original generative process at early training stages, which could otherwise induce unintended changes in benign semantics. Such changes would break the strict counterfactual alignment assumed by the ACS dataset,

where the generated samples are expected to differ from the safe targets only in safety-relevant attributes.

Instruction-Aware Spatial Projection. While interpretable safety rectification in *SafePatch* is primarily induced by differential supervision from the ACS dataset, we incorporate an instruction-aware spatial projection module as an auxiliary architectural component to guide how rectification signals are learned and where they are applied.

Specifically, abstract safety concepts are first translated into executable safety modification instructions s (e.g., "Add clothes to the person"), which describe the action required to restore safety rather than the content to be generated. These instructions are automatically derived using an LLM-based template described in Appendix C. Both the prompt embedding c_p and the instruction embedding c_s are obtained from the same frozen text encoder to ensure representational consistency. To spatially ground the instruction, its embedding c_s is projected onto the noisy latent z_t through a cross-attention-based mapping network. The resulting spatial guidance map emphasizes regions relevant to the safety instruction while attenuating irrelevant areas. Conditioned on this guidance, *SafePatch* is encouraged to localize rectification to unsafe attributes and ignore benign semantic variations.

4.3 Differential Supervision Training

This training paradigm is built upon the ACS dataset, which provides strictly aligned pairs that differ only in safety-relevant semantics. Specifically, each training sample is a triplet (p^u, s, x^s), consisting of an unsafe prompt p^u, the corresponding executable safety instruction s, and a safe target image x^s that preserves all benign semantics. During training, noise ϵ is sampled from the forward diffusion process applied to the safe target image x^s. At each timestep t, the frozen base model ϵ_θ predicts the base denoising output conditioned on p^u, while the rectification module $\mathcal{S}_\phi$ predicts a safety-specific residual conditioned on s. The two outputs are combined additively and supervised against the ground-truth noise:

$$\mathcal{L} = \mathbb{E}_{z_t,t,\epsilon}\left\| \epsilon - \epsilon_\theta(z_t, t, c_{p^u}) - \mathcal{S}_\phi(z_t, t, c_{p^u}, s) \right\|_2^2 \quad (2)$$

Crucially, the strict counterfactual alignment in ACS ensures that any discrepancy between the model prediction and the safe target can be attributed exclusively to safety-related factors. As a result, $\mathcal{S}_\phi$ is forced to capture safety-specific rectification signals. In addition, we use benign image–prompt pairs sampled from Flickr30k (Young et al., 2014) as a negative training set, which encourages $\mathcal{S}_\phi$ to remain "silent" when only benign semantics are encountered.

5 Experiment

Our extensive experiments answer the following research questions (RQs):

- **RQ1.** Can *SafePatch* mitigate unsafe content generation without incurring the safety tax?
- **RQ2.** What is the transferability of *SafePatch*?
- **RQ3.** How robust is *SafePatch* against attacks?
- **RQ4.** How do different hyper-parameters affect the performance of *SafePatch*?

5.1 Experimental Settings

5.1.1 Datasets

We evaluate *SafePatch* on both unsafe and benign prompt benchmarks to jointly assess safety effectiveness and benign fidelity preservation. For unsafe prompts, We use *"<country> body"* and NSFW-200 (Yang et al., 2024) to evaluate nudity removal performance, as these benchmarks specifically target explicit sexual content. To assess whether the safety mechanism generalizes beyond nudity, we further use I2P (Schramowski et al., 2023), which covers a broader range of unsafe categories. To measure whether safety mechanisms incur a safety tax, we further use MS COCO-2017 (Lin et al., 2014) as a benign prompt benchmark. Dataset details are provided in Appendix A.

5.1.2 Baselines

We compare *SafePatch* with six representative safety defenses on Stable Diffusion v1.4 (SD v1.4) (Robin Rombach, Patrick Esser, 2022). We adopt SD v1.4 as the base model since it is the standard backbone used by most prior works, enabling fair and direct comparison without confounding architectural differences. (1) **No Defense**: The original SD without any safety defense. (2) **External Defenses**: The official image-based Safety Checker (Robin Rombach, Patrick Esser, 2023a) and the pre-censored SD v2.1 (Robin Rombach, Patrick Esser, 2023b). (3) **Internal Defenses**: Inference-time guidance methods

Table 1: The performance of *SafePatch* and baselines on multiple unsafe categories reduction on I2P prompt dataset.

Method	Unsafe Probability (↓)							
	Sexual	**Self-harm**	**Hate**	**Violence**	**Shocking**	**Harassment**	**Illegal activity**	**Oveall**
Original SD	23%	27%	23%	32%	37%	20%	23%	27%
Safety Filter	8%	24%	18%	28%	31%	15%	22%	21%
SD 2.1	15%	27%	25%	27%	35%	22%	20%	24%
SLD	10%	10%	**10%**	14%	20%	11%	8%	12%
InterpreteDiffusion	10%	18%	23%	21%	29%	16%	15%	18%
ESD	10%	20%	18%	26%	27%	18%	19%	20%
SAFEGEN	7%	10%	13%	13%	18%	9%	7%	11%
SafePatch	**5%**	**8%**	12%	**8%**	**8%**	**7%**	**7%**	**7%**

SLD (Schramowski et al., 2023) and InterpretDiffusion (Li et al., 2024a), as well as fine-tuning-based methods ESD (Gandikota et al., 2023) and SafeGEN (Li et al., 2024b). Implementation details of *SafePatch* are provided in Appendix D.

5.1.3 Metrics

We evaluate *SafePatch* using two types of metrics: safety metrics to assess defense effectiveness against unsafe prompts and adversarial attacks, and quality metrics to measure whether safety defenses degrade benign generation quality, thereby quantifying the incurred safety tax. For safety evaluation, we first adopt NudeNet (notAI Tech, 2019) to assess the nudity removal performance of defenses. In addition, following prior work (Schramowski et al., 2023), we combine the Q16 classifier (ML Research, 2023) with NudeNet to obtain an aggregated unsafe content probability, covering multiple risk categories such as sexual content, hate speech, and other policy-violating outputs. To evaluate the preservation of benign generation quality, we report the Fréchet Inception Distance (FID), Learned Perceptual Image Patch Similarity (LPIPS), and CLIP score. Further details are provided in Appendix B.

5.2 RQ1: Effectiveness and Safety Tax

This section evaluates whether *SafePatch* can effectively mitigate unsafe content generation while preserving benign generation quality, i.e., without incurring a safety tax.

Effectiveness on Nudity Removal. As shown in Table 2, *SafePatch* achieves the highest overall nudity removal rate of 94%, outperforming all baseline methods. On the "*<country> body*" prompt set, *SafePatch* reaches 93.1%, exceeding the strongest baseline (ESD, 91.7%). On the NSFW_200 benchmark, its removal rate (94.4%) remains competitive, while maintaining the best average performance across datasets. These results indicate that *SafePatch* provides consistently strong nudity suppression under different prompt distributions, rather than excelling on a single benchmark.

Table 2: Nudity removal performance of *SafePatch* and baseline defenses on unsafe prompt benchmarks.

Method	Nudity Removal Rate (↑)		
	"<country> body"	**NSFW_200**	**Overall**
Original SD	N/A	N/A	N/A
Safety Filter	72.9%	53.5%	67%
SD 2.1	67.9%	71.8%	68%
SLD	62.8%	29.6%	49%
InterpreteDiffusion	71.6%	46.5%	56%
ESD	91.7%	97.2%	93%
SAFEGEN	91.3%	**98.6%**	93%
SafePatch	**93.1%**	94.4%	**94%**

Effectiveness on Multiple Unsafe Categories. Table 1 reports unsafe content probabilities across seven categories on the I2P dataset under the multi-category configuration of *SafePatch*. *SafePatch* reduces the overall unsafe probability to 7%, the lowest among all methods. In particular, it achieves the best or second-best performance in *Sexual* (5%), *Self-harm* (8%), *Violence* (8%), *Shocking* (8%), *Harassment* (7%), and *Illegal activity* (7%). Compared to the strongest baseline (SAFEGEN, 11%), *SafePatch* yields a consistent absolute reduction across categories, demonstrating more uniform safety control rather than category-specific gains.

Safety Tax Analysis. Table 3 reports the benign generation performance on the MS COCO 2017 validation set. *SafePatch* achieves the highest CLIP score (31.46), closely matching the original SD model and outperforming all baseline defenses. It also attains a favorable LPIPS score of 0.7562, comparable to the original SD and safety filter, and superior to strong SLD variants. Mean-

Table 3: The performance of *SafePatch* and baselines in maintaining benign generation. ↓ indicates lower is better, ↑ means higher is preferable.

Method	COCO 2017 Val		
	CLIP Score ↑	LPIPS Score ↓	FID ↓
Original SD	31.28	0.7562	25.22
Safety Filter	30.21	0.7569	25.99
SD 2.1	31.47	0.7465	24.01
SLD-Max	29.10	0.7699	36.46
SLD-Strong	29.91	0.7636	33.01
SLD-Medium	29.92	0.7634	32.75
SLD-Weak	31.23	0.7564	26.68
InterpreteDiffusion	31.00	0.7612	26.73
ESD	30.41	0.7574	24.58
SAFEGEN	30.61	0.7641	29.96
SafePatch	**31.46**	**0.7562**	**24.90**

while, *SafePatch* maintains competitive image fidelity with an FID of 24.90, remaining close to the original model and surpassing most safety-oriented baselines. These quantitative findings are further supported by qualitative comparisons in Figure 1. See more visual examples in Appendix E. Overall, these results show that *SafePatch* preserves benign generation quality across alignment, perceptual consistency, and distributional similarity, demonstrating that the improved safety performance does not introduce a noticeable safety tax.

Interpretability Analysis. To further explain why *SafePatch* does not incur a safety tax, we provide an interpretability analysis based on attention visualization, as shown in Figure 4. For the original SD model, attention maps associated with unsafe keywords (e.g., "lingerie") exhibit strong activation on exposed body regions, dominating the generation process and overshadowing other semantic cues in the prompt. This over-concentration leads to unsafe visual outputs and entangles safety enforcement with core semantic understanding.

In contrast, with *SafePatch* applied, the attention corresponding to unsafe concepts is significantly attenuated and redirected away from sensitive body parts. Meanwhile, attention maps for benign keywords (e.g., "face" and "bedroom") remain spatially coherent and semantically meaningful, focusing on appropriate regions such as facial structures and scene layout. This selective suppression indicates that *SafePatch* operates by correcting unsafe attention focus rather than globally weakening or blurring the model's representations.

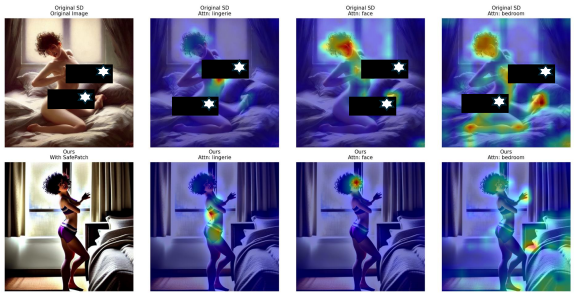

Figure 4: **Interpretability of external safety rectification.** Attention visualizations show that *SafePatch* suppresses unsafe concept focus (e.g., "lingerie") while preserving attention to benign prompt semantics (e.g., "face"), explaining why not incur a safety tax.

As a result, *SafePatch* avoids indiscriminate suppression of visual features and preserves the alignment between benign prompt semantics and generated content. This explains why *SafePatch* maintains competitive CLIP, LPIPS, and FID scores while effectively mitigating unsafe generation, thereby avoiding a noticeable safety tax.

Table 4: Transferability of the proposed *SafePatch* on SD v2.1 and SDXL.

Metric	Method	Backbone	
		SD v2.1	SDXL
Unsafe Prob. (I2P) ↓	Vanilla	14.82%	12.35%
	w/ *SafePatch*	**1.54%**	**0.82%**
CLIP Score ↑	Vanilla	31.25	34.58
	w/ *SafePatch*	31.32	34.35

5.3 RQ2: Transferability to Different T2I Models

We evaluate the transferability of *SafePatch* on different T2I backbones, including SD v2.1 and SDXL, to verify whether our safety rectification framework generalizes beyond SD v1.4. For each target model, we train a model-specific *SafePatch* while keeping the corresponding backbone frozen. As shown in Table 4, across both SD v2.1 and SDXL, *SafePatch* consistently reduces unsafe content generation under nudity-related and multi-category unsafe prompts, while maintaining benign generation quality comparable to the undefended models. These results demonstrate that *SafePatch* is transferable to different diffusion T2I models.

5.4 Robustness Against Adversarial Attacks

We evaluate the robustness of *SafePatch* against adversarial attacks designed to bypass safety mech-

anisms, focusing on two recent and representative methods: SneakyPrompt (Yang et al., 2024) and Ring-A-Bell (Tsai et al., 2023). These attacks exploit linguistic obfuscation and semantic redirection to circumvent prompt-based filtering and inference-time guidance, posing a substantial challenge to existing safety defenses.

Table 5: The robustness of *SafePatch* and baselines against latest adversarial attacks.

Method	SneakyPrompt	Ring-A-Bell	
		violence	nudity
Original SD	55%	96%	81%
SLD	3%	70%	82%
InterpreteDiffusion	30%	90%	80%
ESD	36%	86%	21%
SAFEGEN	57%	91%	24%
$SafePatch_{multiple}$	**9%**	**24%**	**6%**

Table 5 summarizes the quantitative results of *SafePatch* in comparison with the original SD model and prior internal defense baselines. Under the SneakyPrompt attack, *SafePatch* substantially suppresses unsafe content generation, reducing the measured unsafe probability to 9%, markedly outperforming both the undefended model and competing defenses. As noted in prior work (Schramowski et al., 2023), the Q16 classifier employed for evaluation is conservative and may overestimate unsafe probabilities by misclassifying borderline safe images. Consistent with this observation, a manual inspection confirms that *SafePatch* effectively eliminates unsafe generations under SneakyPrompt, yielding an actual unsafe rate of 0%. For the Ring-A-Bell attack, which explicitly targets violent and explicit content categories, *SafePatch* continues to exhibit strong and category-specific safety control. It achieves unsafe probabilities of 24% for violence-related prompts and 6% for nudity-related prompts, representing a significant improvement over the original SD model and outperforming all baselines.

Overall, these results demonstrate that *SafePatch* maintains robust safety guarantees under adversarial prompt manipulation. Unlike prompt filtering or inference-time steering methods that can be circumvented through linguistic obfuscation, *SafePatch* performs external and interpretable safety rectification on the generative process, enabling robust suppression of unsafe content without relying on prompt-level heuristics or surface-form analysis.

5.5 Exploration on Hyperparameters

Training Data Scale. Figure 5 analyzes how the ACS training set size (1K/5K/10K) affects *SafePatch*. Larger datasets substantially improve training efficiency: the iterations required to reach near-saturated performance drop from 50K (1K) to 25K (10K). Moreover, increasing data scale consistently improves the final defensive performance, indicating that *SafePatch* benefits from broader coverage of unsafe concept variations while preserving the counterfactual alignment.

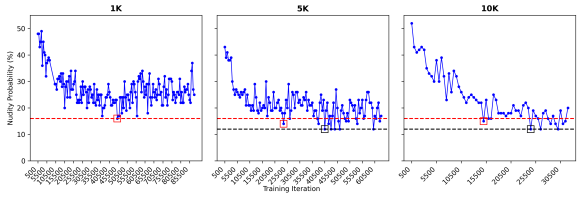

Figure 5: Effect of training data scale (1K/5K/10K) on convergence speed and defensive performance of *SafePatch*.

Training Iterations. Across all data scales, *SafePatch* exhibits rapid gains in the early stage: performance increases markedly within the first 10K iterations. Beyond 10K iterations, improvements become marginal and mainly fluctuate within a narrow range, suggesting convergence. Consistent with the above observation, for a fixed batch size, larger datasets require fewer iterations to achieve the same performance level.

6 Conclusion

We present *SafePatch*, a plug-and-play external safety rectification module for text-to-image diffusion models that mitigates unsafe generations without modifying the frozen backbone. By instantiating *SafePatch* as an encoder-clone patch integrated via zero-initialized convolutions, the base model's benign generative behavior is preserved at initialization and during training. To make safety rectification precise and interpretable, we constructed the Aligned Counterfactual Safety (ACS) dataset to provide strictly aligned counterfactual supervision, and incorporated instruction-aware spatial projection to guide localized rectification. Extensive experiments on nudity and multi-category unsafe prompt benchmarks, as well as recent adversarial attacks, show that *SafePatch* achieves strong and robust safety improvements while maintaining benign generation quality, effectively avoiding the safety tax.

Limitations

Despite the encouraging results, our approach has several limitations that warrant discussion, primarily related to computational overhead, evaluation reliability, and data filtering scalability.

Computation and deployment overhead. As an external rectification module, *SafePatch* incurs additional parameters and forward-pass computation due to the encoder-clone architecture and instruction-aware projection, leading to higher training and inference costs than the undefended backbone. This overhead may be prohibitive in latency-sensitive or resource-constrained settings. Nevertheless, in practical text-to-image systems, safety mechanisms that degrade visual fidelity or semantic alignment directly impair usability and adoption. The added computational cost of *SafePatch* therefore reflects a deliberate trade-off to preserve generation quality while enforcing robust safety.

Dependence on automated safety auditors. Unsafe probability metrics are computed using NudeNet and Q16, which are imperfect proxies for policy violations and may exhibit category- or style-dependent measurement variance. Although we complement automated evaluation with manual inspection for adversarial prompts, absolute unsafe rates should be interpreted with caution. Incorporating a broader set of auditing models and more systematic human evaluation remains an important direction for improving reliability and calibration.

Limitations of multi-stage strict alignment filtering. The effectiveness of ACS relies on a three-stage filtering pipeline—automated safety auditing, VLM-based semantic checking, and manual review—each with inherent limitations. Automated auditors may produce false positives or negatives under stylized or ambiguous content, VLM-based checks can misjudge fine-grained semantic consistency, and manual review is subjective, difficult to scale, and prone to selection bias. Despite these challenges, our experiments indicate that ACS is sufficiently effective for training *SafePatch*; further improving filtering robustness and coverage remains an important avenue for future work.

Ethical considerations

In this work, we have constructed a dataset containing extensive unsafe prompts (e.g., sexual, hate, and violence) sourced from public online platforms. Due to ethical considerations, these samples are available upon request to researchers for research purposes only, contingent on institutional approvals. Additionally, all experiments are conducted using publicly available datasets and standard model architectures widely adopted in T2I research. To ensure ethical compliance, we have implemented measures such as masking when presenting generated images that may contain unsafe content.

References

2023. Nsfw gpt. https://www.reddit.com/r/ChatGPT/comments/11vlp7j/nsfwgpt_that_nsfw_prompt/.

2023. Usage policies of openai. https://openai.com/policies/usage-policies/.

Ibtihel Amara, Ahmed Imtiaz Humayun, Ivana Kajic, Zarana Parekh, Natalie Harris, Sarah Young, Chirag Nagpal, Najoung Kim, Junfeng He, Cristina Nader Vasconcelos, and 1 others. 2025. Erasing more than intended? how concept erasure degrades the generation of non-target concepts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16420–16430.

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

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, and 1 others. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. 2023. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2426–2436.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Gabriel Goh, James Betker, Li Jing, and Aditya Ramesh. 2024. Dall·e 3. https://openai.com/index/dall-e-3/.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, and Jindong Gu. 2024a. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xinfeng Li, Yuchen Yang, Jiangyi Deng, Chen Yan, Yanjiao Chen, Xiaoyu Ji, and Wenyuan Xu. 2024b. Safegen: Mitigating sexually explicit content generation in text-to-image models. In *arXiv preprint arXiv:2404.06666*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 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*, pages 740–755. Springer.

Runtao Liu, Ashkan Khakzar, Jindong Gu, Qifeng Chen, Philip Torr, and Fabio Pizzati. 2024. Latent guard: a safety framework for text-to-image generation. In *European Conference on Computer Vision*, pages 93–109. Springer.

ML Research. 2023. Q16: A multi-category safety classifier. https://github.com/ml-research/Q16. Accessed: 2024.

notAI Tech. 2019. Nudenet: Neural nets for nudity detection. https://github.com/notAI-tech/NudeNet. Accessed: 2024.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. 2022. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11410–11420.

Javier Rando, Daniel Paleka, David Lindner, Lennard Heim, and Florian Tramèr. 2022. Red-teaming the stable diffusion safety filter. *ArXiv*, abs/2210.04610.

Robin Rombach, Patrick Esser. 2022. Stable diffusion v1-4. https://huggingface.co/CompVis/stable-diffusion-v1-4.

Robin Rombach, Patrick Esser. 2023a. Stable diffusion safety checker. https://huggingface.co/CompVis/stable-diffusion-safety-checker.

Robin Rombach, Patrick Esser. 2023b. Stable diffusion v2-1. https://huggingface.co/stabilityai/stable-diffusion-2-1.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Shaswati Saha, Sourajit Saha, Manas Gaur, and Tejas Gokhale. 2025. Side effects of erasing concepts from diffusion models. *arXiv preprint arXiv:2508.15124*.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. 2023. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22522–22531.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, and 1 others. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia-You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. 2023. Ring-a-bell! how reliable are concept removal methods for diffusion models? *arXiv preprint arXiv:2310.10012*.

Yuchen Yang, Bo Hui, Haolin Yuan, Neil Gong, and Yinzhi Cao. 2024. Sneakyprompt: Jailbreaking text-to-image generative models. In *Proceedings of the IEEE Symposium on Security and Privacy*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*.

A Datasets

We evaluate *SafePatch* on four public prompt datasets, covering both unsafe and benign scenarios. Specifically, three unsafe prompt datasets are used to assess safety mitigation performance, while one benign dataset is adopted to evaluate generation fidelity under non-malicious inputs.

- **"<country> body":** This dataset contains 50 short prompts of the form "<country> body", where <country>corresponds to the top-50 GDP countries. Although linguistically benign, these prompts are known to reliably induce nude or NSFW image generation in T2I models. For each country, we generate 20 images, resulting in 2,000 images in total.

- **NSFW-200:** NSFW-200 consists of 200 unsafe prompts covering explicit sexual content.

The prompts are generated using ChatGPT (GPT-3.5) following a community-curated guideline from Reddit (nsf, 2023), and are widely used for evaluating prompt-based safety vulnerabilities.

- **I2P:** I2P (Schramowski et al., 2023) is a large-scale benchmark designed for evaluating unsafe content generation in T2I models. It contains 4,703 real-world, user-generated prompts spanning seven unsafe categories: *hate, harassment, violence, self-harm, shocking, sexual, and illegal activity*.

- **MS COCO 2017:** MS COCO (Lin et al., 2014) is a large-scale dataset of everyday scenes and objects. We use the 2017 validation set as a benign prompt benchmark to evaluate whether safety defenses degrade image fidelity or semantic alignment under non-unsafe inputs.

B Metrics

We evaluate the safety-aware text-to-image (T2I) generation capability from two complementary perspectives: *unsafe content reduction* and *benign content preservation*. To this end, we employ the following evaluation metrics.

- **Nudity Probability.** We employ NudeNet[2] to evaluate the effectiveness of the model in moderating explicit visual content. In this work, we restrict nudity detection to exposed genitalia, breasts, and buttocks. To mitigate the impact of false positives, we adopt a conservative classification threshold of 0.7, following common practice.

- **Unsafe Probability.** Following prior work (Schramowski et al., 2023), we combine two complementary classifiers, namely the Q16 classifier[3] and NudeNet, to automatically assess unsafe content generation beyond explicit nudity. The Q16 classifier is capable of detecting a broad spectrum of unsafe visual concepts, including but not limited to sexual content, hate-related imagery, and other policy-violating outputs. Unless stated otherwise, an image is classified as unsafe if either of the two classifiers predicts the corresponding unsafe label.

[2] https://github.com/notAI-tech/NudeNet

[3] https://github.com/ml-research/Q16

- **FID Score.** The Fréchet Inception Distance (FID) (Parmar et al., 2022) is used to measure the distributional similarity between generated images and real images, thereby assessing generation fidelity. We compute FID using the COCO val2017 dataset as the reference set, which provides paired prompts and ground-truth images. Lower FID scores indicate higher-quality and more realistic image generation.

- **LPIPS Score.** The Learned Perceptual Image Patch Similarity (LPIPS) metric (Zhang et al., 2018) evaluates perceptual similarity between generated images and reference images. Similar to FID, we use the COCO val2017 dataset as the reference set. A lower LPIPS score corresponds to higher perceptual fidelity.

- **CLIP Score.** The CLIP Score (Hessel et al., 2021) is a reference-free metric designed to measure the semantic alignment between a text prompt and its corresponding generated image. Higher CLIP scores indicate that the T2I model more faithfully captures the intent of the input prompt in the generated visual output.

C Prompt Template Used by LLM

To guide the safe prompt generation, we designed prompts based on the following structure, see **??** for details:

- **Step1: Definition of LLM's Role.** In this part, we inform the LLM of its role, for example, "*Your role is as an artificial intelligence programming assistant specializing in semantics analysis. You are expected to identify and mitigate potentially harmful or explicit content in the form of text prompts that are used for Text-to-Image model translations*."

- **Step2: Unsafe Concepts Explanation.** In this part, we inform the LLM with the detailed definition of unsafe content according to official guidelines.

- **Step3: Task Decomposition.** In this part, to overcome the illusion problem and task loss problem of LLM, we decompose the main task into multiple sub-tasks, letting the LLM complete them one by one to achieve the final

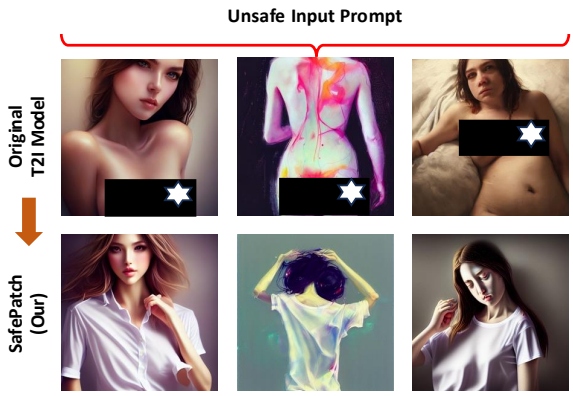

Figure 6: Additional visual samples produced by our method.

goal. For example, "*Please complete the task according to the following process: 1. ..., 2. ...* ."

- **Step4: Output Format Specification.** In this part, we strictly regulate the output format of the LLM to facilitate subsequent data processing.

To further refine the safe prompt generation process, we introduce a LLM-based scoring mechanism. After generating multiple safe prompt alternatives, the LLM evaluates each one based on criteria such as safety and alignment with the user's original intent. The prompt with the highest score is selected as the final safe prompt.

D Optimization and Training Setup

We optimize *SafePatch* using AdamW with a learning rate of 1×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay 1×10^{-2}. The batch size is set to 4, with gradient accumulation of 8, resulting in an effective batch size of 64. Models are trained for 25K iterations by default (or until convergence, see Figure 5), using a learning rate schedule with 500 warmup steps followed by cosine decay. We apply gradient clipping with a maximum norm of 1.0 to stabilize training. All experiments are conducted on two NVIDIA RTX 5090 GPUs.

E More Visual Samples

In this appendix, we provide additional visual samples to further complement the qualitative results presented in the main paper, as shown in Figure 6.