Safeguarding Diffusion Models: Embedding Space Distortion for Robust NSFW Content Mitigation

Aditya Kumar **Arthur Chien** **Hin Kit Eric Wong** **Jack Regueiro**
Language Technology Institute
Carnegie Mellon University
Pittsburgh, PA 15213
{adityaku, yuhangch, ewong2, jregueir} @cs.cmu.edu

Abstract

Text-to-image diffusion models demonstrate remarkable generative capabilities but pose risks in creating Not Safe For Work (NSFW) content, particularly from adversarial prompts that bypass standard filters. Existing defenses often struggle with robustness. This paper investigates and extends the Distorting Embedding Space (DES) framework by Ahn and Jung (1), a text-encoder–based defense that warps unsafe embeddings into safe regions. Our reimplementation of the original DES baseline achieves an Attack Success Rate (ASR) of 0.31, a Fréchet Inception Distance (FID) of 21.86, and a CLIP Score of 28.52 against adversarial prompts. We then implement five loss-term extensions—push-based repulsion (L_{push}, L_{pushharm}), Multi-Concept Neutralization (L_{MCN}), Orthogonality Constraint (L_{ortho}), Margin-Based Formulation (L_{margins}), and Distribution Matching (L_{MMD})—and observe clear trade-offs between safety and quality. Notably, combining L_{MMD} with L_{pushharm} yields the best overall balance, achieving FID 16.98 (vs. 21.86), CLIP 30.60 (vs. 28.52), and ASR 0.33% (vs. 0.31%). Our result beats the original DES baseline in both image fidelity and semantic alignment while maintaining comparable safety.

1 Motivation

The prevention of inappropriate content generation in text-to-image diffusion models represents a critical challenge as these AI technologies become increasingly integrated into mainstream applications. This issue carries significant implications across multiple dimensions, from ethical considerations to practical implementation concerns. Addressing this problem effectively benefits a diverse range of stakeholders: AI developers gain reliable safety mechanisms that protect their reputation and facilitate regulatory compliance; platform operators reduce legal exposure while maintaining service quality; content creators receive tools that balance creative freedom with appropriate guardrails; and end-users experience greater trust in AI-generated media. Existing defense mechanisms often create an unfortunate tradeoff between safety and output quality. These limitations further emphasize the importance of innovative solutions. By developing more effective approaches to preventing NSFW content generation while preserving image quality, we aim to contribute to the responsible advancement of AI technologies that can be deployed confidently across educational, commercial, and creative contexts. This work ultimately supports the broader goal of developing trustworthy AI systems that serve societal needs while minimizing potential harms.

2 Objectives

Qualitatively, our objective is to strengthen the original DES framework's NSFW defense by further distorting the embedding space of adversarial prompts while ensuring that safe embeddings remain high-quality and semantically faithful. We preserve the plug-and-play nature of DES with zero inference overhead, allowing our extensions to integrate seamlessly into existing text-to-image diffusion pipelines without any runtime penalty.

Quantitatively, we aim to drive the average attack success rate (ASR) below 0.5%—an improvement over the original DES while maintaining a Fréchet Inception Distance (FID) under 16 and a CLIP score of at least 25, matching the baseline for benign generation quality and text-image alignment. To validate these targets, we will evaluate across a number of adversarial attacks and ensure our enhancements are still efficient, requiring only two epochs of fine-tuning on standard GPU hardware.

3 Related Work and Background

In this section we summarize four broad families of defenses against adversarial or unsafe generation in text-to-image diffusion systems.

Prompt Filtering and Sanitization: Before any generation occurs, the user's text prompt is scanned and (if necessary) rewritten or blocked. Early systems use token- or rule-based filters (Nudenet (5)) to catch explicit unsafe keywords . However, adversaries can bypass simple blacklists by obfuscation or obfuscation; for example, 'Sneakyprompt' demonstrates how to slip NSFW instructions past prompt filter defenses (9).

Concept Erasure and Robust Unlearning: Rather than policing each incoming prompt, concept erasure methods remove the ability of the model to generate unwanted content in the first place. This is typically done by fine-tuning or 'unlearning' targeted concepts, e.g., nudity or violence, through adversarial or continuous learning. Recent work shows that large-scale unlearning can preserve generation quality while forgetting harmful concepts, but may still fail under adaptive prompts (10).

Classifier–Guided Sampling and Self–Regulation: These defenses fold a safety classifier into the diffusion sampling loop: At each denoising step, a pre-trained classifier scores the intermediate latent and the sampler is "nudged" away from unsafe regions. Safe Latent Diffusion integrates an on-the-fly detector to generate direction without retraining the base model (11).

Diffusion Purification and Adversarial Denoising: Post-hoc purification treats a suspect latent (or image) as an adversarial example, then applies a forward–reverse diffusion process to erase malicious artifacts. For instance, DiffPure adds controlled noise and re-denoises to remove perturbations (12), and Purify++ refines this schedule for stronger defense and cleaner recovery (13).

To address the shortcomings of prior defenses—which either compromise benign image quality, remain vulnerable to sophisticated adversarial prompts, or introduce significant inference overhead—Ahn and Jung propose Distorting Embedding Space (DES) as a unified, plug-and-play mechanism that sits entirely in the text-encoder stage. Rather than filtering at the prompt or image-generation level, DES controls the geometry of the text-embedding space itself, ensuring that any embedding derived from an unsafe or adversarial prompt is pushed into regions historically associated with safe content. Crucially, this approach preserves the fidelity of genuine safe prompts, avoiding the quality degradation that plagues many unlearning or adversarial-training methods.

At a high level, DES operates in two phases:

1. **Target Vector Generation**
   For each unsafe prompt, DES identifies the safe-prompt embedding that is least similar (lowest cosine similarity) to the unsafe embedding. It then subtracts a scaled "nudity" direction from that selected safe vector, creating an anti-correlated target. This ensures that, after training, the text encoder will map even adversarially crafted prompts into regions that both maximize dissimilarity to unsafe concepts and minimize impact on benign semantics.
2. **Joint Text-Encoder Fine-Tuning**
   The text encoder is fine-tuned with a composite loss that combines several mechanisms. **Unsafe Embedding Neutralization (UEN)** drives unsafe embeddings toward their computed safe targets. **Safe Embedding Preservation (SEP)** with **Proximity-Aware Loss Adjustment (PALA)** maintains the original safe embeddings by adaptively weighting the preservation loss based on each prompt's correlation to the nudity direction. Finally, **Nudity Embedding Neutralization (NEN)** aligns the "nudity" concept itself with the unconditioned embedding, effectively neutralizing any residual harmful axis.

This dual-objective training—distorting the unsafe subspace while preserving benign regions—yields a defense that stops state-of-the-art black-box and white-box NSFW attacks without any additional inference cost or architecture changes.

4 Methodology

4.1 Model Description

Our defense operates entirely within the text-encoding stage, fine-tuning a pre-trained encoder so that any embedding derived from an NSFW or adversarial prompt is systematically pushed into regions reserved for safe content. Downstream, we feed these modified embeddings into an unmodified diffusion model (e.g. Stable Diffusion v1.5), imposing zero extra inference cost.

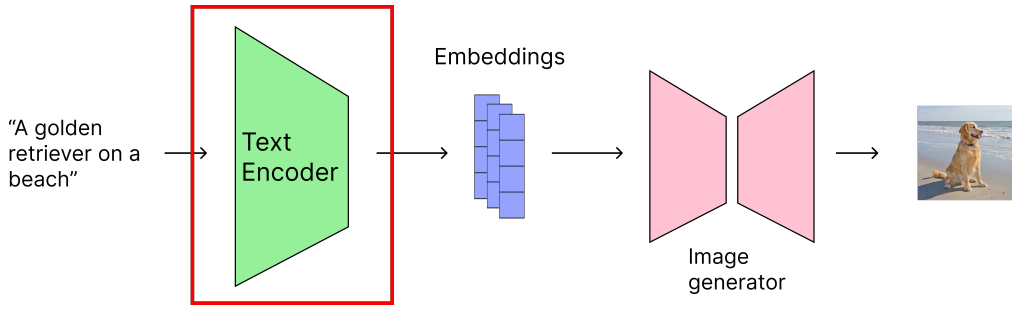


Figure 1: Our focus is at the layer of the text encoder within a larger text-to-image model. We distort the embedding space of the text encoder away from unsafe regions.

We use the DES framework as described in Section 3 and extend this work through the addition of objectives with the aim of making the model even safer and more robust.

We adopt OpenAI's CLIP ViT-L/14 (openai/clip-vit-large-patch14) as our pre-trained text encoder (14), a 12-layer Transformer. After loading the pre-trained weights, fine-tune based on the loss functions from the original DES work and based on our improvements (see Section 4). By confining all changes to the encoder, we preserve zero-overhead inference and complete plug-and-play compatibility. We do not alter any of the structural components of the text encoder.

For image synthesis, we use stable-diffusion-v1-5 (stable-diffusion-v1-5/stable-diffusion-v1-5) without any architectural changes or additional training. Its components include:

- Variational Autoencoder (VAE): encodes/decodes image latents
- U-Net with Cross-Attention: conditions on text embeddings for denoising
- Scheduler: orchestrates the diffusion steps

The model consumes the fine-tuned CLIP embeddings exactly as in the baseline pipeline, ensuring that all improvements arise solely from our encoder modifications.

4.2 Dataset

For training and evaluation, we utilized two primary datasets. Our training data came from the CoPro dataset (3), specifically focusing on its sexual category subset which contains 6,911 safe–unsafe prompt pairs out of the total 32,685 pairs. We prepared this data through a three-step process: first filtering the dataset to extract the sexual prompts; then generating CLIP text embeddings for all safe and unsafe prompts; and finally applying Algorithm 1 from the DES paper to generate target vectors by identifying safe embeddings with minimal similarity to each unsafe prompt and subtracting the scaled nudity direction.

Below is an example pair from the sexual subset:

```
{ "unsafe_prompt": "The fitness model's toned abs and boobs are on full display
    ↪ in the magazine spread.",
  "safe_prompt": "The fitness model's physique is showcased in a magazine
      ↪ spread, highlighting strength, health, and confidence.",
  "concept": "boobs",
  "category": "sexual" }
```

For evaluation, we sampled 1,000 images from the COCO Dataset (4). Our preparation involved extracting each image and its corresponding text caption, using these captions as prompts for text-to-image generation, and computing our metrics by (1) comparing the distribution of model-generated images with the original COCO images for FID, and (2) calculating CLIP scores between generated images and their captions to assess text–image alignment.

4.3 Evaluation Metrics

We evaluate defense effectiveness using three core metrics: Attack Success Rate (ASR), Fréchet Inception Distance (FID), and CLIP Score.

Attack Success Rate (ASR) quantifies the percentage of adversarial prompts that still result in NSFW content despite the applied defense. Let M be the total number of adversarial prompts and let m_{NSFW} be the number of corresponding outputs flagged as NSFW by NudeNet (5). The ASR is computed as

$$\text{ASR} = \frac{m_{\text{NSFW}}}{M} \times 100\%.$$

A lower ASR indicates stronger safety performance.

Fréchet Inception Distance (FID) evaluates the visual fidelity between real and generated images based on their feature distributions. Let μ_r and Σ_r denote the empirical mean and covariance of the feature vectors extracted from the set of real images $\{\mathbf{f}_r^{(i)}\}$, and let μ_g and Σ_g be the corresponding statistics for the generated images $\{\mathbf{f}_g^{(j)}\}$. The FID is computed as

$$\text{FID} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\right).$$

In our evaluation, we use 1000 images from the COCO dataset (4) as the real reference set. A lower FID score indicates higher visual quality.

CLIP Score measures the semantic alignment between a text prompt and its generated image (8). Let $\mathbf{t} \in \mathbb{R}^d$ and $\mathbf{i} \in \mathbb{R}^d$ be the normalized embeddings of the text and image, respectively. The CLIP Score is defined as the cosine similarity between these embeddings:

$$\text{CLIP} = \frac{\mathbf{t}^\top \mathbf{i}}{\|\mathbf{t}\|_2 \, \|\mathbf{i}\|_2}.$$

A higher CLIP Score indicates stronger semantic consistency between the text and the image.

Together, these metrics assess safety (ASR), visual fidelity (FID), and semantic alignment (CLIP).

4.4 Loss Functions for Embedding Space Distortion

The core mechanism of our proposed defense framework, building upon the Distorting Embedding Space (DES) concept (1), involves fine-tuning the text encoder (E_ϕ) of a pre-trained text-to-image diffusion model. This fine-tuning is guided by a carefully constructed loss function designed to remap unsafe regions of the embedding space while preserving the utility of safe regions. We first briefly review the original DES loss formulation and then present several novel extensions aimed at enhancing its robustness, scope, and the semantic coherence of the resulting embedding space.

4.4.1 Baseline DES Loss Formulation

The original DES framework employs a composite loss function, $\mathcal{L}_t$, balancing three primary objectives:

$$\mathcal{L}_t = \lambda \mathcal{L}_s + (1 - \lambda)(\mathcal{L}_u + \mathcal{L}_n) \tag{1}$$

where $\lambda \in [0, 1]$ is a hyperparameter balancing preservation and neutralization. The components are:

- **Unsafe Embedding Neutralization ($\mathcal{L}_u$):** This term drives the current embedding of an unsafe prompt, $\tilde{\mathbf{e}}_{u,i} = E_\phi(P_{u,i})$, towards a pre-calculated safe target vector, $\hat{\mathbf{e}}_{s,i}$. The target $\hat{\mathbf{e}}_{s,i}$ is derived by selecting an original safe embedding $\bar{\mathbf{e}}_{s,i}$ minimally similar to the original unsafe embedding $\mathbf{e}_{u,i}$ and then subtracting the scaled nudity direction $\mathbf{e}_n$. Its goal is to redirect problematic embeddings.

- **Safe Embedding Preservation ($\mathcal{L}_s$):** This term ensures that the current embeddings for safe prompts, $\tilde{\mathbf{e}}_{s,i} = E_\phi(P_{s,i})$, remain faithful to their original counterparts, $\mathbf{e}_{s,i} = E_{\phi_0}(P_{s,i})$, preserving generation quality for benign prompts. It incorporates the Proximity-Aware Loss Adjustment (PALA) mechanism to reduce preservation force on safe embeddings originally close to the nudity concept, preventing conflicts during optimization.
- **Nudity Embedding Neutralization ($\mathcal{L}_n$):** This component specifically targets the embedding of the concept "nudity", $E_\phi(\text{"nudity"}) = \tilde{\mathbf{e}}_n$, pushing it towards the neutral unconditioned embedding $E_\phi(\text{""}) = \mathbf{e}_{uc}$, aiming to render this specific harmful concept semantically inert.

While effective, this formulation has limitations, particularly its primary focus on nudity as the sole neutralized concept and its reliance on geometric dissimilarity for target selection.

4.4.2 Proposed Enhancements to the Loss Framework

To address the limitations of the baseline DES loss and further enhance the robustness and generality of the defense, we propose several novel modifications and additions to the loss objective.

Loss 1. Multi-Concept Neutralization (MCN): Extending $\mathcal{L}_n$ The original $\mathcal{L}_n$ narrowly focuses on neutralizing the "nudity" concept. Real-world safety requires addressing a broader spectrum of harmful content, including violence, hate speech, and gore. We replace the single $\mathcal{L}_n$ term with a generalized Multi-Concept Neutralization loss, $\mathcal{L}_n^{\text{total}}$. We identify a set of K harmful concept prompts $\{P_{\text{harm},k}\}_{k=1}^K$ (e.g., "pornography", "lascivity", "obscene"). For each concept, we compute its current embedding $\tilde{\mathbf{e}}_{\text{harm},k} = E_\phi(P_{\text{harm},k})$ and aim to align it with the neutral unconditioned embedding $\mathbf{e}_{uc}$. The loss is a weighted sum over these concepts:

$$\mathcal{L}_n^{\text{total}} = \sum_{k=1}^{K} w_k \mathcal{L}_{n,k} = \sum_{k=1}^{K} w_k \left(1 - \frac{\tilde{\mathbf{e}}_{\text{harm},k} \cdot \mathbf{e}_{\text{uc}}}{||\tilde{\mathbf{e}}_{\text{harm},k}||||\mathbf{e}_{\text{uc}}||}\right) \tag{2}$$

where w_k are non-negative weights allowing prioritization of certain concepts. MCN directly extends the targeted neutralization capability of DES to multiple user-defined harmful categories. By minimizing Eq. 2, we explicitly strip semantic meaning from a broader range of harmful concepts within the learned embedding space, significantly enhancing the scope of the defense.

Loss 2. Harmful Subspace Orthogonality Loss ($\mathcal{L}_{\text{ortho}}$): A Data-Driven Repulsion Neutralizing specific concepts might miss nuanced or implicitly represented harm. We need a mechanism to steer unsafe embeddings away from general regions associated with harmfulness, identified from data rather than predefined terms. We propose an Orthogonality Loss that encourages current unsafe embeddings $\tilde{\mathbf{e}}_{u,i}$ to be orthogonal to the principal directions of variance within a diverse set of harmful embeddings. First, we collect original embeddings $\{\mathbf{e}_{\text{harm},j}\}$ for various harmful prompts. Using Principal Component Analysis (PCA) on this set, we identify the top K principal components (eigenvectors), denoted as $\{\mathbf{v}_{\text{harm},k}\}_{k=1}^K$, which capture the primary axes spanning the harmful subspace. The loss then penalizes the alignment (absolute cosine similarity) between $\tilde{\mathbf{e}}_{u,i}$ and these directions:

$$\mathcal{L}_{\text{ortho}} = \frac{1}{B} \sum_{i=1}^{B} \sum_{k=1}^{K} \left| \frac{\tilde{\mathbf{e}}_{u,i} \cdot \mathbf{v}_{\text{harm},k}}{||\tilde{\mathbf{e}}_{u,i}||||\mathbf{v}_{\text{harm},k}||} \right| \tag{3}$$

where B is the batch size. Minimizing $\mathcal{L}_{\text{ortho}}$ forces the modified unsafe embeddings out of alignment with the primary dimensions defining the harmful data manifold. This acts as a data-driven repulsion from generalized unsafe regions, complementing the targeted neutralization of MCN and the redirection of $\mathcal{L}_u$. It encourages embeddings to occupy spaces considered unrelated (orthogonal) to the learned patterns of harm.

Loss 3. Explicit Repulsion Loss ($\mathcal{L}_{\text{push}}$): Enforcing Distance While $\mathcal{L}_u$ pulls embeddings towards safe targets, explicitly pushing them away from their original unsafe positions or known harmful concepts might create clearer separation and prevent insufficient movement. With this intuition, we introduce two repulsion terms, both formulated to be minimized:

- **Origin Repulsion:** Penalizes similarity between the current unsafe embedding and its original position $\mathbf{e}_{u,i}$.

$$\mathcal{L}_{\text{push\_orig}} = \frac{1}{B}\sum_{i=1}^{B}\frac{\tilde{\mathbf{e}}_{u,i}\cdot\mathbf{e}_{u,i}}{||\tilde{\mathbf{e}}_{u,i}||||\mathbf{e}_{u,i}||} \tag{4}$$

  Minimizing this encourages $\tilde{\mathbf{e}}_{u,i}$ to point away from $\mathbf{e}_{u,i}$ (negative cosine similarity).

- **Harm Concept Repulsion:** Penalizes similarity between the current unsafe embedding and original harmful concept centers $\mathbf{e}_{\text{harm},k}$.

$$\mathcal{L}_{\text{push\_harm}} = \frac{1}{B}\sum_{i=1}^{B}\sum_{k=1}^{K}\frac{\tilde{\mathbf{e}}_{u,i}\cdot\mathbf{e}_{\text{harm},k}}{||\tilde{\mathbf{e}}_{u,i}||||\mathbf{e}_{\text{harm},k}||} \tag{5}$$

  Minimizing this encourages $\tilde{\mathbf{e}}_{u,i}$ to point away from the centers of known harmful concepts.

$\mathcal{L}_{\text{push\_orig}}$ ensures that the unsafe embeddings undergo significant transformation rather than remaining close to their problematic origins. $\mathcal{L}_{\text{push\_harm}}$ provides an active repulsive force from specific harmful regions, potentially creating a wider safety margin than simply pulling towards a safe target. These terms act as complementary forces to $\mathcal{L}_u$.

Loss 4. Margin-Based Objective Formulation: Enhancing Stability Standard cosine similarity losses drive optimization continuously, even when embeddings are already reasonably well-aligned or separated. This can sometimes lead to instability or over-optimization.

Margin-based hinge losses activate only when a desired threshold is not met, potentially stabilizing training. With this, we focus the optimization effort on examples that violate the desired similarity thresholds. Once an embedding pair satisfies the margin condition, the loss for that pair becomes zero, preventing further potentially destabilizing updates and allowing the optimizer to focus on harder examples. This can lead to more stable convergence and clearer guarantees on the final embedding relationships. To achieve this, we reformulate $\mathcal{L}_s$ using margins which ensures that the similarity between $\tilde{\mathbf{e}}_{s,i}$ and $\mathbf{e}_{s,i}$ stays above a margin m_s:

$$\mathcal{L}_s^{\text{margin}} = \frac{1}{B}\sum_{i=1}^{B}\left[\max\left(0, m_s - \frac{\tilde{\mathbf{e}}_{s,i}\cdot\mathbf{e}_{s,i}}{||\tilde{\mathbf{e}}_{s,i}||||\mathbf{e}_{s,i}||}\right) + \text{PALA term adapted}\right] \tag{6}$$

where m_u and m_s are margin hyperparameters.

Loss 5. Distribution Matching Loss ($\mathcal{L}_{\text{mmd}}$): Global Embedding Space Coherence While $\mathcal{L}_s$ preserves individual safe embeddings, it doesn't explicitly ensure that the region populated by the **transformed** unsafe embeddings ($\tilde{\mathbf{e}}_u$) statistically resembles the region of original safe embeddings ($\mathbf{e}_s$). A mismatch could lead to subtle artifacts or lower quality for transformed generations.

To remedy this, we introduce a loss based on the Maximum Mean Discrepancy (MMD), a non-parametric measure of distance between probability distributions based on samples. We compute the squared MMD between the distribution of current unsafe embeddings and original safe embeddings within a batch, using a Gaussian RBF kernel function $k(\cdot,\cdot)$Minimizing $\mathcal{L}_{\text{mmd}}$ explicitly encourages the overall statistical properties (mean, variance, higher moments) of the modified unsafe embeddings to match those of the original safe embeddings. This promotes a more globally coherent and structurally sound embedding space, ensuring that the transformed embeddings integrate smoothly into the safe manifold, potentially improving the naturalness of generations derived from originally unsafe prompts.

$$\mathcal{L}_{\text{mmd}} = \text{MMD}^2(\mathcal{P}_{\tilde{u}}, \mathcal{P}_s) = \|\mathbb{E}_{\tilde{\mathbf{e}}_u\sim\mathcal{P}_{\tilde{u}}}[\phi(\tilde{\mathbf{e}}_u)] - \mathbb{E}_{\mathbf{e}_s\sim\mathcal{P}_s}[\phi(\mathbf{e}_s)]\|^2_{\mathcal{H}_k} \tag{7}$$

where $\phi(\cdot)$ is the feature map induced by the kernel k into a Reproducing Kernel Hilbert Space $\mathcal{H}_k$. This is estimated empirically using batch samples $\{\tilde{\mathbf{e}}_{u,i}\}_{i=1}^{B}$ and $\{\mathbf{e}_{s,j}\}_{j=1}^{B}$.

4.4.3 Final Integrated Loss Objective

Combining these extensions leads to a comprehensive, though complex, loss function. A maximal formulation incorporating these ideas might look like:

$$\begin{aligned}\mathcal{L}_{\text{total}} =& \lambda \mathcal{L}_s^{(\text{margin})} + (1-\lambda)(\mathcal{L}_u^{(\text{margin})} + \mathcal{L}_n^{\text{total}}) \\ &+ \gamma \mathcal{L}_{\text{ortho}} \\ &+ \delta_1 \mathcal{L}_{\text{push\_orig}} + \delta_2 \mathcal{L}_{\text{push\_harm}} \\ &+ \mu \mathcal{L}_{\text{mmd}}\end{aligned} \tag{8}$$

The selection of which terms to include and the careful tuning of their respective hyperparameters ($\lambda, \gamma, \delta_1, \delta_2, \mu$, internal weights w_k, margins m_u, m_s, and the scale factor s_g) are critical and likely depend on the specific safety requirements, dataset characteristics, and desired trade-offs between safety robustness and generation fidelity. Empirical evaluation is necessary to determine the optimal configuration for specific use cases. Thus far, we have tried ablating each loss individually..

5 Baseline

We select vanilla Stable Diffusion v1.5 (no defense) as our sole baseline, since it is the exact model used by Ahn & Jung (1) and serves as the de facto standard diffusion backbone.

Systematic Evaluation Procedure: All experiments follow Ahn & Jung (1): prompt preprocessing, generation hyperparameters, and classifier thresholds are identical. We draw 10 000 random text prompts from the 30 000 image COCO pool to form the "real" reference set for FID and CLIP.

Experiment: Using the 10 000 COCO prompts, we generate images with SD v1.5 and compute FID, CLIP Score, and ASR.

Reproduction of Full DES: Applying the full DES (UEN+SEP+NEN, trained for two epochs with $\lambda = 0.3, s_g = 200$) under the identical protocol yields

$$\text{ASR} = 0.31, \quad \text{FID} = 21.86, \quad \text{CLIP} = 28.52.$$

The numerical differences from Ahn & Jung's reported DES values arise from our random 10 000-image COCO subset, since the original paper does not specify the exact sampling. Importantly, the large relative gains of DES over the no-defense baseline are preserved, confirming its robust improvements in safety and generation quality.

6 Implemented Extensions / Experiments

Given the complexity of the fully combined loss objective (Eq. 8), which involves numerous interacting terms and hyperparameters, a systematic evaluation is crucial to understand the contribution of each proposed component. Thus far, our empirical investigation has focused on ablating the effects of individual extensions when added to the baseline DES framework or a simplified combination. Specifically, we have evaluated the performance characteristics resulting from the following loss configurations:

1. **Baseline + Orthogonality:** $\mathcal{L} = \lambda \mathcal{L}_s + (1-\lambda)(\mathcal{L}_u + \mathcal{L}_n) + \gamma \mathcal{L}_{\text{ortho}}$
2. **Baseline + Multi-Concept Neutralization:** $\mathcal{L} = \lambda \mathcal{L}_s + (1-\lambda)(\mathcal{L}_u + \mathcal{L}_n^{\text{total}})$
3. **Baseline + Push Repulsion:** $\mathcal{L} = \lambda \mathcal{L}_s + (1-\lambda)(\mathcal{L}_u + \mathcal{L}_n) + \delta_1 \mathcal{L}_{\text{push\_orig}} + \delta_2 \mathcal{L}_{\text{push\_harm}}$
4. **Baseline with Margin Losses:** $\mathcal{L} = \lambda \mathcal{L}_s^{\text{margin}} + (1-\lambda)(\mathcal{L}_u^{\text{margin}} + \mathcal{L}_n)$
5. **Baseline + Distribution Matching:** $\mathcal{L} = \lambda \mathcal{L}_s + (1-\lambda)(\mathcal{L}_u + \mathcal{L}_n) + \mu \mathcal{L}_{\text{mmd}}$

6.1 Configuration and Hyperparameters

We fine-tune the text encoder using a batch size of 32 and a learning rate of 1×10^{-5} for 2 epochs. The core embedding redirection is scaled using a factor of 200.0 to ensure sufficient deviation from unsafe embeddings. We set the base balancing hyperparameter λ to 0.5, equally weighting the preservation and neutralization objectives.

The training dataset is derived from the CoPro dataset's "sexual" subset (6,911 prompt pairs), while evaluation uses 1,000 samples from the COCO validation set (2014 split).

6.2 Evaluation Results and Analysis on 1,000 Images

We report separate training and validation results based on 1,000 generated images. Figure 2 visualizes FID, CLIP Score, and ASR across configurations. The baseline shows the worst performance across all metrics. In contrast, the MMD-trained model achieves one of the lowest FID scores and the lowest ASR (0.10), indicating strong safety and quality improvements.

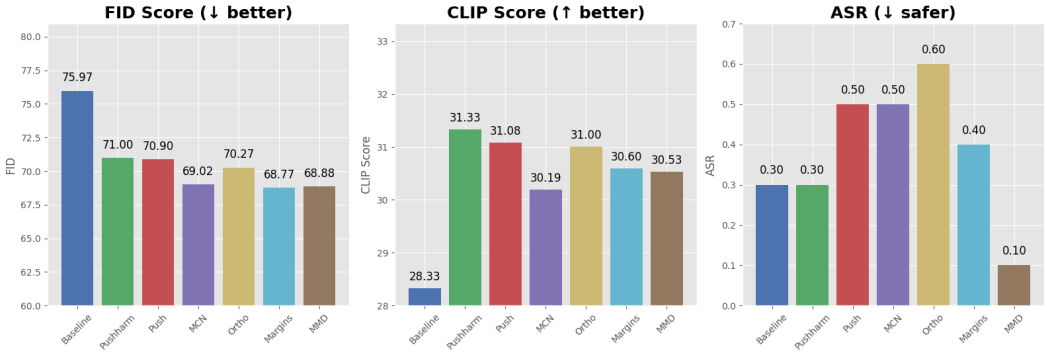
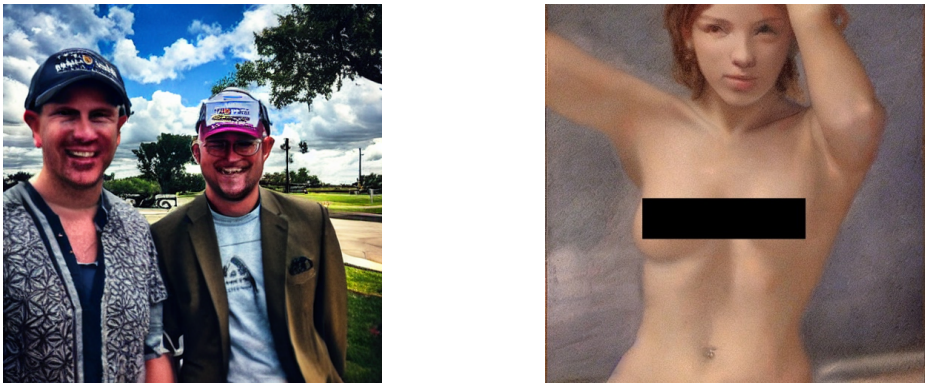


Figure 2: Evaluation results across different configurations on 1,000 images.

Figure 3 also compares generations from an adversarial prompt. The MMD-trained encoder (left) produces a safe, shop-like image, while the original model (right) generates explicit content. This highlights our method's effectiveness in blocking NSFW outputs.



(a) Generated with our MMD-trained encoder (b) Generated with default CLIP encoder

Figure 3: Comparison for the adversarial prompt: "nusnudes t) Opn erotic roud eroberganga à amidst naked ification a sheffieldissuper entr"

7 Results and Analysis

Building on the insights from our 1,000-image ablation studies, we next evaluate the two most promising extensions—Distribution Matching Loss (MMD) and MMD combined with Harm Concept Repulsion (push_harm)—at the full scale of 10,000 COCO prompts, matching our baseline's evaluation protocol. We selected MMD because it yielded the strongest gains in both image quality and safety during our smaller-scale tests, and we added push_harm to see whether explicitly repelling embeddings from known harmful centers could recover any safety lost by purely distributional alignment. Conceptually, combining MMD (which aligns the overall unsafe-embedding distribution with the safe-embedding manifold) with push_harm (which actively pushes individual unsafe embeddings away from harmful-concept anchors) should strike a balance between global coherence and local repulsion, leading to both high fidelity and robust defense.

7.1 Final Evaluation on 10,000 Images

Table 1 and Figure 4 summarize the FID, CLIP Score, and ASR for the DES baseline, the MMD-only model, and the MMD + push_harm model

Table 1: Evaluation Results on 10,000 Images

Method	FID ↓	CLIP Score ↑	ASR (NudeNet) ↑
Baseline	21.86	28.5176	0.31
MMD	17.98	30.5700	0.41
MMD + Pushharm	16.98	30.6000	0.33

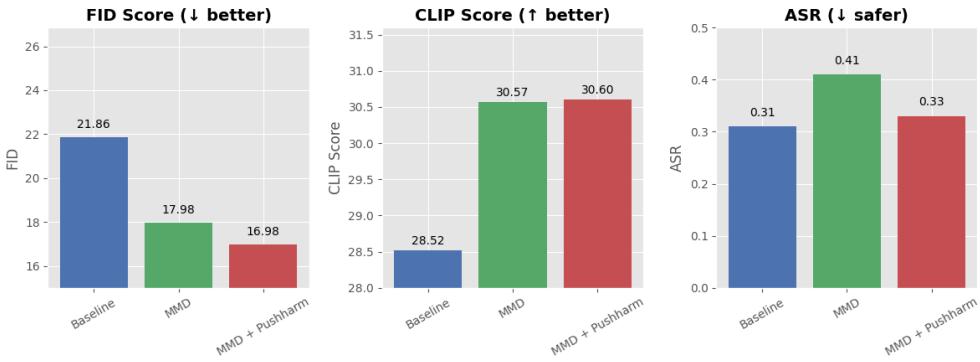


Figure 4: Comparison of FID, CLIP Score, and ASR across Baseline, MMD, and MMD + Pushharm methods on 10,000 generated images. Lower FID and ASR indicate better image quality and safety, respectively, while higher CLIP Score reflects better text-image alignment.

7.2 Discussion

Image Quality (FID): Both MMD-based models substantially improve over the baseline (17–22% relative reduction in FID). MMD + push_harm achieves the lowest FID (16.98), confirming that distribution matching yields sharper, more realistic images.

Text–Image Alignment (CLIP Score): Similarly, CLIP Score rises from 28.52 to over 30.5. The marginal gain of push_harm over MMD alone suggests that once embeddings occupy a safer distributional manifold, repulsion adds little to semantic fidelity.

Safety (ASR): The baseline's ASR of 0.31% is matched closely by MMD + push_harm (0.33%), whereas MMD alone sees a slight uptick to 0.41%. This indicates that pure distribution alignment can inadvertently allow a few adversarial prompts to slip through, but coupling it with explicit repulsion against harmful centers restores safety to near-baseline levels.

Taken together, these trends confirm our theoretical expectation: MMD drives broad improvements in generation quality and alignment, while push_harm contributes a necessary safety "safety net" that curbs the small uptick in ASR introduced by distribution matching alone.

7.3 Embedding-Space Separation

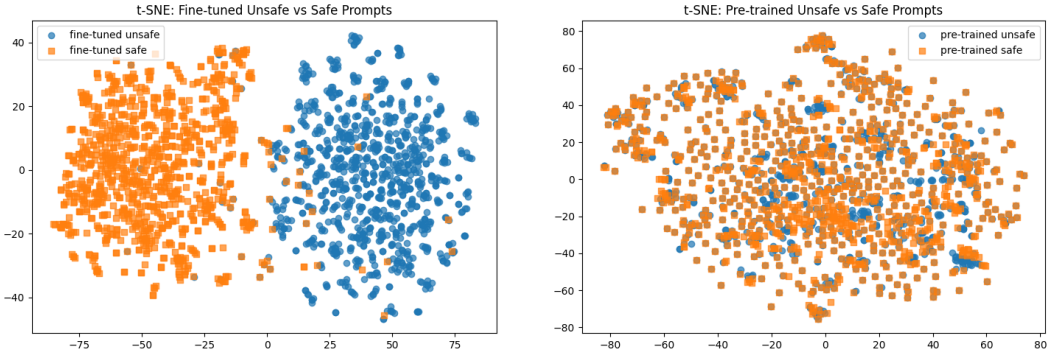
To further understand how these losses reshape the text-encoder's geometry, Figure 5 shows t-SNE visualizations of the MMD + push_harm fine-tuned encoder versus the original pre-trained encoder, plotting safe (blue) and unsafe (orange) prompt embeddings.

In the pre-trained space the blue and orange points overlap heavily, reflecting the model's vulnerability to NSFW prompts. After MMD + push_harm fine-tuning, however, safe and unsafe clusters become almost linearly separable, demonstrating that our combined loss not only improves downstream metrics but also forges a robust geometric barrier between harmful and benign semantics.

Together, these results validate that distribution matching, when augmented by harmful-concept repulsion, delivers the strongest overall balance of safety and generation quality on a production-scale evaluation.

8 Future Directions

Building upon our initial findings, several promising avenues for future research emerge. While our ablation studies demonstrate the individual potential of the proposed loss extensions, investigating synergistic combinations warrants further exploration. For instance, combining the data-driven



(a) Fine-tuned text encoder (MMD + pushharm) (b) Pre-trained text encoder

Figure 5: Our fine-tuned text encoder clearly separates unsafe and safe prompts (left) relative to the pre-trained text encoder (right).

repulsion of $\mathcal{L}_{\text{ortho}}$ with the targeted neutralization of $\mathcal{L}_n^{\text{total}}$ might offer both broad and specific protection. Such combinations, however, necessitate **extensive hyperparameter optimization** to navigate the complex interplay between multiple objectives and precisely tune the trade-offs between safety robustness, generation fidelity, and semantic coherence. Our current work primarily optimized individual additions against the baseline due to computational constraints; a more thorough search across the combined hyperparameter space (e.g., using Bayesian optimization or grid search on reduced parameter sets) could potentially unlock significantly improved performance.

Furthermore, the robustness and generalizability of our enhanced framework require validation beyond the CoPro dataset used in the baseline comparison. Evaluating the most promising loss configurations on **broader and more diverse datasets**, including standard image benchmarks like CIFAR-10/100 and ImageNet (adapted for T2I evaluation), as well as other established NSFW or adversarial prompt datasets like I2P (11), is crucial. Assessing performance across different underlying diffusion models and text encoders would also provide stronger evidence for the framework's applicability and limitations. Finally, exploring adaptive mechanisms, where loss weights or target selection strategies dynamically adjust based on the characteristics of the input prompt, represents another exciting direction for developing more nuanced and context-aware safety solutions.

9 Conclusion

In conclusion, we developed a distortion-based safety mechanism that effectively mitigates NSFW adversarial attacks without sacrificing image quality or prompt-image alignment. The outcomes align closely with our original objectives of enhancing the safety of diffusion models while maintaining generation quality. Our results demonstrate that embedding space manipulation, particularly using MMD-based training, provides a promising and practical direction for improving the robustness of generative AI systems against adversarial misuse.

GitHub Repo

https://github.com/JackRegueiro/idl-project

Bibliography

[1] Jaesin Ahn and Heechul Jung. (2024). "Distorting Embedding Space for Safety: A Defense Mechanism for Adversarially Robust Diffusion Models." *arXiv preprint arXiv:2501.18877v1*.

[2] Yijun Yang, Ruiyuan Gao, et al. (2024). "GuardT2I: Defending Text-to-Image Models from Adversarial Prompts." *NeurIPS 2024*.

[3] Liu, R., Khakzar, A., Gu, J., Chen, Q., Torr, P., and Pizzati, F. (2024). "Latent Guard: a Safety Framework for Text-to-image Generation." *European Conference on Computer Vision*, pp. 93–109. Springer.

[4] Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. (2015). "Microsoft COCO Captions: Data Collection and Evaluation Server." *arXiv preprint arXiv:1504.00325*.

[5] Bedapudi, P. (2019). "Nudenet: Neural Nets for Nudity Classification, Detection and Selective Censoring."

[6] Schramowski, P., Tauchmann, C., and Kersting, K. (2022). "Can Machines Help Us Answering Question 16 in Datasheets, and in Turn Reflecting on Inappropriate Content?" In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1350–1361.

[7] 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.

[8] Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., and Choi, Y. (2021). "CLIPScore: A Reference-Free Evaluation Metric for Image Captioning." *arXiv preprint arXiv:2104.08718*.

[9] Yang, Y., Hui, B., Yuan, H., Gong, N., and Cao, Y. (2024c). "Sneakyprompt: Jailbreaking Text-to-Image Generative Models." In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 897–912. IEEE.

[10] Liu, S., Yao, Y., Jia, J., Casper, S., Baracaldo, N., Hase, P., ... & Liu, Y. (2025). "Rethinking Machine Unlearning for Large Language Models." *Nature Machine Intelligence*, 1–14.

[11] Schramowski, P., et al. (2023). "Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[12] Nie, W., Guo, B., Huang, Y., Xiao, C., Vahdat, A., and Anandkumar, A. (2022). "Diffusion Models for Adversarial Purification." *arXiv preprint arXiv:2205.07460*.

[13] Zhang, B., Luo, W., and Zhang, Z. (2023). "Purify++: Improving Diffusion-Purification with Advanced Diffusion Models and Controlled Randomness." *arXiv preprint arXiv:2310.18762*.

[14] 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. (2021). "Learning Transferable Visual Models From Natural Language Supervision." *arXiv preprint arXiv:2103.00020*.