The Illusion of High Utility in Safety Alignment of Text-to-Image Diffusion Models

Adeel Yousaf, Soumik Ghosh, James Beetham,
Amrit Singh Bedi, and Mubarak Shah

Institute of Artificial Intelligence
University of Central Florida, Orlando, United States
{adeel.yousaf,soumik.ghosh,james.beetham,amritbedi}@ucf.edu,
shah@crcv.ucf.edu

Abstract. Safety alignment of text-to-image (T2I) diffusion models aims to suppress harmful generations while preserving utility on benign prompts. Recent methods often appear to deliver high safety with high utility, but this conclusion rests largely on coarse global utility metrics (e.g., FID, CLIPScore) that are insensitive to fine-grained semantic correctness, creating an *illusion* of high utility. We show that when utility is measured with structured evaluation, this illusion breaks: on TIFA (Text-to-Image Faithfulness evaluation with Question Answering), safety-aligned models suffer substantial drops in semantic fidelity, including failures in object counts, attributes, and relationships. To diagnose the source of this gap, we analyze the text-encoder prompt embedding space and uncover *semantic collapse*, a contraction of embedding spread coupled with distortion of inter-prompt similarity structure, which strongly correlates with structured utility loss. Guided by this insight, we propose Structure-Aware Geometric Regularization (SAGE)[1], a safety alignment objective that explicitly preserves embedding spread and inter-prompt relational structure during adaptation. Our method restores structured utility (TIFA +5.0% over prior state-of-the-art) while maintaining strong safety performance and competitive coarse-grained utility scores.

Keywords: Text-to-Image · Safety Alignment · Embedding Analysis

1 Introduction

Text-to-image (T2I) diffusion models such as stable diffusion (SD) [19] and DALL-E [16] can generate highly realistic images from natural language prompts, but their training on large web-scale datasets also exposes them to unsafe concepts, enabling the generation of Not-Safe-For-Work (NSFW) content [20, 29]. As these models become widely accessible, mitigating unsafe generations has become a critical requirement for responsible deployment. At the same time, safety modifications must preserve the model's core capability: generating images that

[1] https://adeelyousaf.github.io/SAGE_ECCV26_Project_Page/

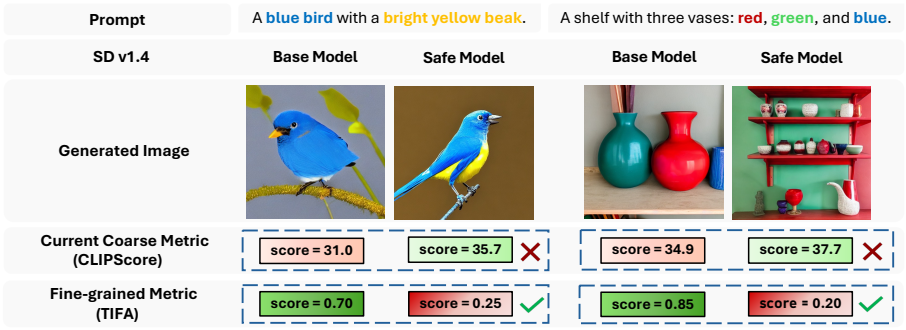


Fig. 1: The illusion of high utility under coarse evaluation. Comparison between the **base model** and a **safe model** (unlearned) across fine-grained prompts. While the safe model fails to generate specific attributes (e.g., the *yellow beak* or the *correct vase colors/count*), the standard **CLIPScore** provides misleadingly higher scores for the incorrect images (✗). In contrast, the fine-grained metric, **TIFA**, accurately captures the utility degradation (✓), properly penalizing the safe model for failing to satisfy the detailed visual requirements of the prompt.

follow the prompt. If enforcing safety significantly degrades model capability, aligned models may become less attractive than their unrestricted counterparts.

Recently, several works have proposed safety alignment methods for T2I models to suppress harmful generations [1, 12, 17, 20, 22, 27–29]. These approaches typically evaluate safety using attack success rates (ASR), while utility is assessed using coarse metrics such as Fréchet Inception Distance (FID) [9] and CLIPScore. In early works, improving safety often came at the expense of model performance, suggesting a safety-utility tradeoff in practice. However, recent methods such as DES [1] appear to achieve strong safety while maintaining nearly identical performance to the base model under these metrics, suggesting that the safety–utility tradeoff may be largely resolved.

An illusion of high utility. We argue that this conclusion is incomplete because the utility metrics most commonly reported are too coarse-grained to capture compositional instruction-following failures, as summarized in Tab. 1. This issue is illustrated in Fig. 1: despite a competitive CLIPScore, the safety-aligned model violates prompt constraints such as object attributes and counts. For T2I generation, utility is not only about overall visual quality or global image–text similarity, but also about whether the model correctly renders the objects, attributes, counts, and relationships specified in the prompt. When evaluated with structured benchmarks such as Text-to-Image Faithfulness Evaluation with Question Answering (TIFA) [10], a different picture emerges: despite appearing competitive under FID and CLIPScore, state-of-the-art safety alignment methods consistently underperform the base model in compositional and semantic fidelity.

Our diagnosis: semantic collapse. To understand the above phenomenon, we analyze how existing text-based T2I safety methods reshape the prompt embed-

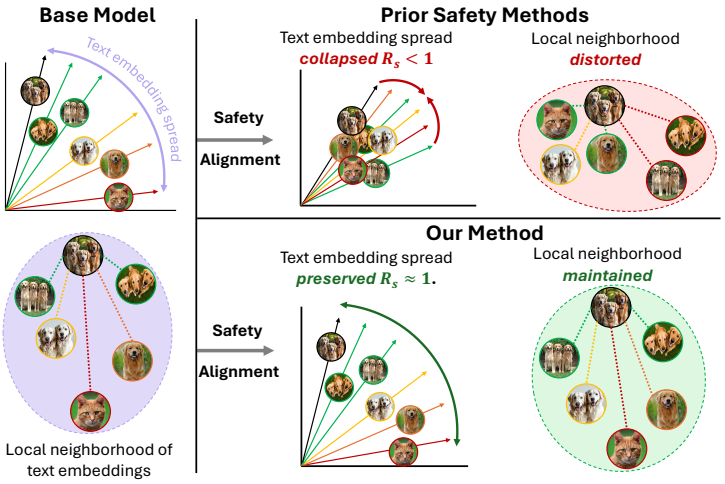


Fig. 2: Embedding geometry under safety alignment. The figure illustrates how safety alignment alters the structure of the text-encoder embedding space. **Left: Base Model.** The prompt "Three golden retrievers" is used as a reference. In the base model, semantically related prompts are arranged according to their similarity: "Three dogs" lies closest to the reference, followed by "Two golden retrievers", while an unrelated concept ("cat") appears far away. The arrows visualize the spread of prompt embeddings, and the circular region highlights the local semantic neighborhood around the reference prompt. **Top-right: Prior safety alignment methods.** Safety tuning often alters this structure in two ways. First, the embedding spread can decrease, causing prompt embeddings to become more concentrated. Second, the local semantic neighborhood becomes distorted: prompts that were previously unrelated may move closer to the reference prompt, causing unrelated concepts (e.g., "cat") to appear within the neighborhood of the reference prompt. **Bottom-right: Our method.** Our alignment objective preserves both properties of the original embedding space. The embedding spread remains comparable to that of the base model, while the local semantic neighborhood around the reference prompt is better retained.

ding space. This embedding space is not merely an auxiliary representation; it organizes semantic relationships among prompts and, therefore, plays a central role in how objects, attributes, counts, and relationships are preserved during generation. Intuitively, safety tuning can pull many prompts into a tighter region of the embedding space and reshuffle which prompts are considered similar, which may preserve coarse global similarity scores while breaking fine-grained constraint satisfaction. Our analysis reveals two consistent structural effects of safety alignment: (1) *embedding contraction*, where the overall spread of prompt embeddings decreases and the embeddings become more concentrated, and (2) *neighborhood distortion*, where the local similarity structure among prompts shifts, causing different prompts to become nearest neighbors relative to the base model, as shown in Fig. 2. We refer to this phenomenon as *semantic collapse*.

Proposed approach. Motivated by this diagnosis, we propose a geometry-aware safety alignment objective designed to counteract the two failure modes underlying semantic collapse. Here, the *geometry* of the prompt embedding space refers to (i) its overall spread and (ii) the local neighborhood structure induced by pairwise similarities, both of which shape how semantic constraints are preserved during generation. Accordingly, our objective aims to preserve the embedding geometry that supports instruction-following while still allowing the model to adapt for safety. First, we introduce an embedding spread regularization term that penalizes contraction in total embedding spread relative to the base model. Second, we introduce a local structural correlation loss that preserves pairwise similarity relationships among semantically close prompts, thereby maintaining the fine-grained structure of the embedding manifold. Empirically, our proposed alignment restores fine-grained semantic fidelity while maintaining strong safety performance and competitive global utility metrics. Our contributions are:

1. **Illusion of high utility under coarse evaluation.** We show that widely used global utility metrics (FID, CLIPScore) can *mislead* by suggesting little or no utility loss after safety alignment, whereas structured evaluation with TIFA reveals substantial degradations in compositional instruction-following (counts, attributes, relationships).
2. **Diagnosing semantic degradation in embedding space.** We identify *semantic collapse*, an embedding spread contraction, and local neighborhood distortion in the prompt embedding space, and show that it strongly correlates with structured utility loss.
3. **Our fix: geometry-aware safety alignment.** We introduce a geometry-aware alignment objective that regularizes the spread of embeddings and their relational structure. Our method restores structured fidelity (TIFA 75.4 vs. 76.3 base) while maintaining strong safety (average ASR 1.2% vs. 67.6% base) and competitive global alignment (CLIPScore 26.4 vs. 26.5 base).

2 The Illusion of High Utility

What coarse metrics miss. A central goal of T2I safety alignment is to reduce unsafe generations while retaining utility on benign prompts. In practice, utility in prior safety work is predominantly reported using coarse global metrics such as FID and CLIPScore (Tab. 1). These metrics are convenient and widely adopted, but they do not directly test whether a model follows the *structured constraints* expressed in many prompts (e.g., object counts, attributes, and relations) [7,10]. As a result, a safety-aligned model can appear to preserve utility under standard protocols even when it fails to satisfy prompt constraints Fig. 1 illustrates a representative failure mode: despite competitive global scores, the safety-aligned model violates prompt-specific requirements such as color consistency and object count. This mismatch is not visible when evaluation focuses on distributional realism (FID) or coarse image-text similarity (CLIPScore), motivating the need for structured utility benchmarks.

Table 1: Utility evaluation setups used in prior T2I safety methods, including the datasets and coarse metrics (FID and CLIPScore) reported by each method.

Method	Utility Dataset	FID	CLIPScore
DES [1]	COCO	✓	✓
STEREO [22]	I2P Unsafe Prompts	✓	✓
ADV-Unlearn [28]	COCO	✓	✓
ESD [6]	COCO	✓	✓
SALUN [5]	Non-forget Classes	✓	✗
RECE [8]	COCO	✓	✓
RACE [11]	COCO	✓	✓
MACE [15]	COCO	✓	✓
SLD [21]	COCO, Human Study	✓	✓
AlignGuard [13]	COCO	✓	✓
Safe-CLIP [17]	COCO, LAION-400M	✓	✓
SafeR-CLIP [27]	Parti-Prompts	✗	✓

Structured utility evaluation with TIFA. To directly evaluate compositional instruction-following, we benchmark safety-aligned models using TIFA [10], a structured protocol that verifies whether objects, attributes, counts, and relations specified in the prompt are correctly instantiated in the generated image. Unlike coarse metrics, TIFA explicitly targets semantic correctness and constraint satisfaction.

Quantifying the illusion of high utility. Re-evaluating representative safety alignment methods under TIFA reveals substantial semantic degradation that is not reflected by coarse metrics. For example, DES [1] improves FID (16.23 vs. 17.23 base) and largely maintains CLIPScore (25.5 vs. 26.5), suggesting minimal degradation under standard evaluation. However, TIFA shows a 6.2% overall drop relative to the base model, with non-uniform category-level degradation, including a 13.0% decrease on food-related prompts. These failures remain largely invisible under FID and CLIPScore. We refer to this systematic gap: *high coarse-metric utility alongside degraded structured semantics* as the **illusion of high utility**.

2.1 From Illusion to Cause: Measuring Semantic Collapse.

The TIFA gap raises a natural question: *what changes in the model lead to these compositional failures*? Since many safety methods operate by fine-tuning the text encoder, we analyze how safety alignment reshapes the prompt embedding space. We quantify *Semantic Collapse* as a geometric shift characterized by (i) embedding spread contraction and (ii) neighborhood distortion.

Embedding spread. We first measure how the spread of embeddings changes after safety fine-tuning. Given B benign prompts with ℓ_2-normalized embeddings $\mathbf{z}^{(i)}$ and mean embedding $\bar{\mathbf{z}}$, we define the embedding spread as the average

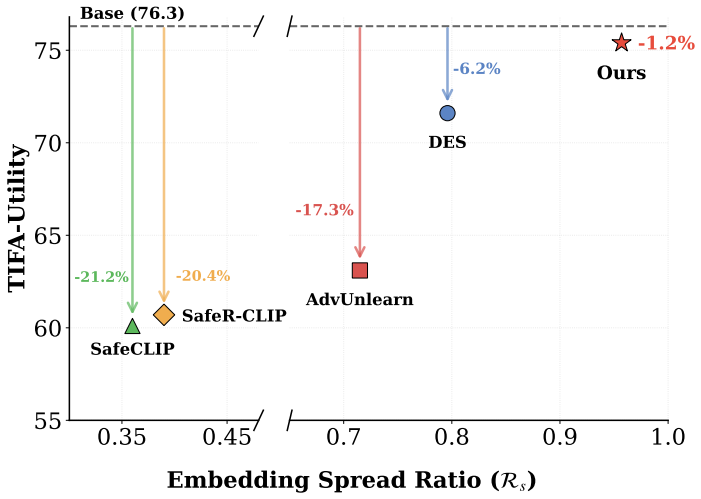


Fig. 3: Relationship between spread ratio ($\mathcal{R}_s$) and structured utility (TIFA). Methods with larger reductions in overall embedding spread exhibit larger TIFA drops, indicating that embedding compression is closely associated with compositional degradation.

squared distance of the embeddings from their batch mean:

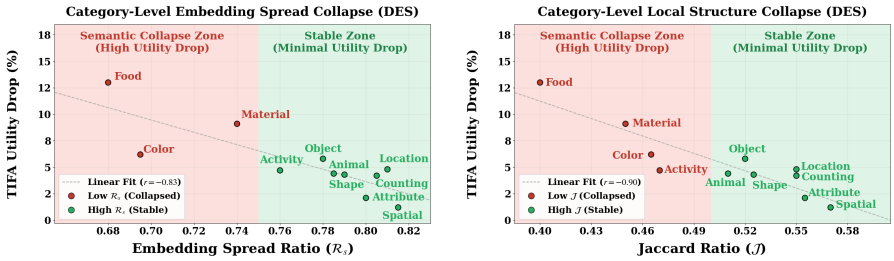
$$\mathcal{S} = \frac{1}{B}\sum_{i=1}^{B}\left\|\mathbf{z}^{(i)} - \bar{\mathbf{z}}\right\|_2^2. \tag{1}$$

We compute this quantity for both the safety-aligned model and the base model, yielding $\mathcal{S}_\theta$ and $\mathcal{S}_0$. The embedding spread ratio is defined as $\mathcal{R}_s = \frac{\mathcal{S}_\theta}{\mathcal{S}_0}$. A value $\mathcal{R}_s < 1$ indicates a reduction in embedding spread, $\mathcal{R}_s \approx 1$ indicates a spread comparable to that of the base model, and $\mathcal{R}_s > 1$ indicates an increase in embedding spread.

Neighborhood distortion. Embedding spread alone does not capture whether *relative relationships* among prompts are preserved. For each prompt i, let $\mathcal{N}_i^{(0)}$ denote its top-K nearest neighbors under the base embeddings and $\mathcal{N}_i^{(\theta)}$ the corresponding set under the safety-aligned embeddings. We quantify neighborhood overlap using the Jaccard similarity:

$$J_i = \frac{|\mathcal{N}_i^{(0)} \cap \mathcal{N}_i^{(\theta)}|}{|\mathcal{N}_i^{(0)} \cup \mathcal{N}_i^{(\theta)}|} \tag{2}$$

where $\mathcal{N}_i^{(0)}$ and $\mathcal{N}_i^{(\theta)}$ are the top-K neighbors under the base and safe-aligned models, respectively. A high Jaccard score indicates that the model retains its relational logic and subject-attribute binding, ensuring that related concepts remain clustered together as they were in the base model.



(a) (b)

Fig. 4: Geometric characterization of semantic collapse under safety alignment. (a) Category-level analysis for DES showing that a lower Embedding Spread Ratio ($\mathcal{R}_s$) corresponds to a higher TIFA utility drop (Pearson $r = -0.86$). Categories such as *Food* and *Material* fall into the semantic collapse region (low spread, high utility drop), while categories retaining higher spread ratios remain comparatively stable. **(b)** Category-level analysis for DES showing that a lower Jaccard Ratio (J) corresponds to a higher TIFA utility drop (Pearson $r = -0.90$). Categories such as *Food* fall into the semantic collapse region (low Jaccard, high utility drop), while categories with higher Jaccard ratios remain comparatively stable.

2.2 Utility Degradation Through the Lens of Semantic Collapse

We evaluate prior safety-alignment methods by computing the embedding spread ratio and local semantic structure preservation metrics over the full set of TIFA prompts.

Method-level evidence. Fig. 3 plots embedding spread against fine-grained utility (TIFA), revealing a strong positive correlation between embedding spread and semantic performance. Methods that induce greater spread contraction consistently exhibit larger drops in utility. For example, DES [1] attains a spread ratio of 0.80 with a TIFA score of 71.6, whereas Adv-Unlearn [28] shows a lower spread ratio of 0.72 and a correspondingly lower TIFA score of 63.1. Similar trends are observed across other text-encoder-based safety methods, including SafeCLIP [17] and SafeRCLIP [27].

Category-level evidence (within a single method). A category-wise analysis further strengthens this observation. Fig. 4a shows that categories with greater spread contraction exhibit larger semantic degradation. For example, under DES [1], the *Food* category has the lowest spread ratio ($\mathcal{R}_s = 0.68$) and a 13.0% drop in TIFA accuracy. In contrast, categories with spread ratios closer to 1 experience substantially smaller utility declines. This consistent relationship across semantic categories indicates that embedding spread contraction is closely tied to fine-grained performance loss, motivating the need to explicitly preserve embedding spread during safety alignment.

Neighborhood distortion completes the picture. Local semantic structure analysis for DES further explains the observed utility degradation. As shown in Fig. 4b (K=5), categories with lower Jaccard ratios exhibit substantially larger TIFA drops. For instance, *Food* lies in the semantic collapse region, characterized

by low neighborhood overlap and high utility degradation, whereas categories such as *Spatial* retain higher Jaccard ratios and experience minimal performance decline. The strong negative correlation ($r = -0.90$) indicates that disruption of local semantic neighborhoods closely tracks fine-grained utility loss.

Takeaway. Across methods and semantic categories, structured utility loss is consistently associated with *semantic collapse*, an embedding spread contraction together with neighborhood distortion in the benign prompt embedding space. This motivates explicitly preserving both components of embedding geometry during safety fine-tuning, which we address next with a **geometry-aware safety alignment** objective.

3 Method

Key Notations and Problem Setup. Following existing safety alignment works [1,17,27,28], we consider a pre-trained text-to-image (T2I) diffusion model $\mathcal{F}_\phi$, where $\phi = \{T_\theta, U_\psi\}$. Here, T_θ denotes the text encoder parameterized by θ, and U_ψ denotes the conditional denoising network (UNet) parameterized by ψ. During safety alignment, U_ψ remains frozen, and only T_θ is updated. We denote T_0 as the frozen, original text encoder and T_θ as its trainable counterpart. We denote cosine similarity by $\cos(\cdot, \cdot)$, which measures the normalized inner product between two embeddings. Our training setup utilizes a dataset of paired captions $\mathcal{D} = \{(p_u, p_s)\}$, where p_u represents an unsafe prompt and p_s is its corresponding safe or neutral counterpart. The goal of safety alignment is to optimize T_θ such that unsafe prompts are mapped toward their corresponding safe embeddings, i.e., $T_\theta(p_u) \to T_0(p_s)$, thereby suppressing unsafe image generation while preserving utility on safe prompts.

Revisiting Existing Text-Based T2I Safety Methods. Existing text-based T2I safety alignment methods modify the text encoder of a pre-trained diffusion model while keeping the UNet fixed. These approaches mainly differ in how unsafe text representations are steered. For example, SafeCLIP [17] aligns unsafe embeddings with externally generated safe counterparts, while SafeRCLIP [27] instead maps them to nearby safe neighbors in the model's latent space. DES [1] moves unsafe representations toward a distant safe anchor while removing the target concept direction, and Adv-Unlearn [28] employs adversarial training to steer unsafe prompts toward neutral generation directions.

Despite these differences, most methods preserve utility using *point-wise alignment*, where each benign prompt is matched independently to its base-model representation. Concretely, for B benign prompts $\{p_i\}_{i=1}^{B}$, the utility objective is

$$\mathcal{L}_{\text{util}} = 1 - \frac{1}{B} \sum_{i=1}^{B} \cos\big(T_\theta(p_i),\, T_0(p_i)\big), \tag{3}$$

which encourages the adapted encoder to remain close to the frozen base encoder on a per-prompt basis. However, this objective constrains prompts independently and does not preserve the overall distribution or relational structure of the embedding space.

3.1 Structure-Aware Geometric Regularization (SAGE)

We propose Structure-Aware Geometric Regularization (SAGE), a geometry-preserving framework for T2I safety alignment. SAGE augments DES [1] with two regularization terms that prevent embedding collapse and local semantic distortion. First, we introduce the *Embedding Spread Preservation (ESP)* loss, which maintains the overall embedding spread of the latent space by preventing the embedding distribution from contracting relative to the base model. Second, we propose *Local Structure Alignment (LSA)*, which preserves semantic relationships among nearby prompts. Instead of independent point-wise alignment, LSA matches local similarity patterns between embeddings, ensuring that prompts close in the base model remain similarly related after safety adaptation.

Embedding Spread Preservation (ESP): As discussed in Sec. 2, existing text-based T2I safety alignment methods lead to *embedding spread contraction*, where the spread of text embeddings shrinks relative to the base model. This reduces discriminative capacity and harms compositional utility. To address this, we introduce the *Embedding Spread Preservation (ESP)* loss, which maintains the overall embedding spread of the trainable encoder T_θ relative to the frozen base encoder T_0. For B prompts $\{p_i\}_{i=1}^{B}$, let $\mathbf{z}_\theta^{(i)} = T_\theta(p_i)$ and $\mathbf{z}_0^{(i)} = T_0(p_i)$ denote the ℓ_2-normalized embeddings produced by the current encoder T_θ and the frozen base encoder T_0, respectively. We quantify the embedding spread as the average squared deviation from the batch mean, noted S_θ and S_0, as defined in equation 1. To prevent the embedding space from collapsing, we enforce a lower bound on the embedding spread:

$$\mathcal{L}_{\text{ESP}} = \max\big(0, \text{sg}(\text{S}_0) - \text{S}_\theta\big), \tag{4}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator, preventing gradients from flowing through the frozen base encoder T_0. This one-sided penalty ensures that safety alignment does not reduce the embedding spread below that of the base model, while avoiding unnecessary expansion of the embedding space.

Local Structure Alignment (LSA): While preserving safe embeddings, as noted in Equation 3, does keep the embeddings near their original base model locations, it does not preserve relative similarities between embeddings very well. Consequently, small independent shifts can distort similarity patterns among semantically related prompts, disrupting the local organization of the embedding space as noted in Section 2. To address this limitation, we introduce the *Local Structure Alignment (LSA)* loss. For B benign prompts $\{p_i\}_{i=1}^{B}$, we compute pairwise cosine similarities using the adapted and frozen encoders:

$$S_\theta(i,j) = \cos\big(T_\theta(p_i), T_\theta(p_j)\big), \quad S_0(i,j) = \cos\big(T_0(p_i), T_0(p_j)\big). \tag{5}$$

Rather than enforcing consistency across all prompt pairs, we preserve the local structure defined by the base model. For each prompt p_i, we identify its Top-K most similar prompts under S_0 and compute the alignment only over these local pairs, defined as:

$$\mathcal{L}_{\text{LSA}} = 1 - \frac{1}{|\mathcal{K}|} \sum_{(i,j)\in\mathcal{K}} S_\theta(i,j)\ \text{sg}\big(S_0(i,j)\big), \tag{6}$$

where $\mathcal{K}$ denotes the set of prompt pairs corresponding to the Top-K neighbors under the base encoder. Here, $S_\theta(i,j)$ and $S_0(i,j)$ are standardized over the pairs in $\mathcal{K}$ to have zero mean and unit variance. The resulting objective encourages the adapted encoder to retain the relative similarity relationships among the local pairs identified by the base encoder. However, applying LSA directly to safe embeddings can improve utility but may weaken safety. Because LSA preserves the base model's local similarity structure, it may also restore geometric patterns correlated with unsafe concepts such as "nudity." Prior work [1] shows that even benign prompts can exhibit non-trivial correlation with the nudity direction. Consequently, preserving the original local geometry may inadvertently recover unsafe semantic alignment. To mitigate this, we introduce a concept-perturbed variant of LSA. Instead of enforcing structural consistency on the original embeddings, we perturb the adapted safe embeddings along the unsafe concept direction and enforce the local structure constraint under this perturbation. Specifically, given a concept direction ("nudity" in our case) , we construct perturbed embeddings as

$$\tilde{T}_\theta(p_i) = T_\theta(p_i) + \alpha T_0(\text{“nudity”}), \tag{7}$$

α is set to 1 in our experiments. Enforcing this objective under the perturbation encourages prompt pairs identified as local neighbors by the base encoder to retain their relative similarity relationships, even when the adapted embeddings are shifted along the unsafe concept direction. The updated LSA objective is defined as:

$$\mathcal{L}_{\text{LSA}}^{\text{pert}} = 1 - \frac{1}{|\mathcal{K}|} \sum_{(i,j)\in\mathcal{K}} \tilde{S}_\theta(i,j)\ \text{sg}\big(S_0(i,j)\big), \tag{8}$$

where $\tilde{S}_\theta(i,j)$ denotes cosine similarity computed from the perturbed embeddings $\tilde{T}_\theta(p_i)$, $S_0(i,j)$ is computed from the base encoder T_0, and $\mathcal{K}$ is the set of prompt pairs corresponding to the Top-K neighbors under the base encoder. As in Eq. 6, $\tilde{S}_\theta(i,j)$ and $S_0(i,j)$ are standardized over the pairs in $\mathcal{K}$.

3.2 Full Training Objective

Our training objective combines safety alignment with geometry-preserving regularization. The loss consists of three components: (i) a safety loss $\mathcal{L}_{\text{safe}}$, (ii) a point-wise utility alignment loss $\mathcal{L}_{\text{util}}$, and (iii) geometric regularizers that preserve embedding spread and local semantic structure. The safety loss steers unsafe prompts toward designated safe targets, while the utility loss maintains per-prompt consistency with the frozen base encoder for benign inputs. We adopt the same formulations for $\mathcal{L}_{\text{safe}}$ and $\mathcal{L}_{\text{util}}$ as in prior safety alignment work [1], and provide their full definitions in the supplementary. The overall objective is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{safe}} + \lambda_u \mathcal{L}_{\text{util}} + \lambda_s \mathcal{L}_{\text{ESP}} + \lambda_l \mathcal{L}_{\text{LSA}}^{\text{pert}}, \tag{9}$$

where λ_u, λ_s, and λ_l control the contributions of utility preservation, embedding spread preservation, and local structure alignment, respectively.

4 Experiment

Implementation Details. We follow the experimental protocol used in prior text-based safety alignment methods, particularly DES [1], to ensure a fair comparison. All experiments use Stable Diffusion v1.4 [19] as the backbone. Consistent with text-level safety approaches, we fine-tune only the text encoder while keeping the remaining components frozen. For training, we use 6,911 safe–unsafe prompt pairs from the sexual category of the CoPro dataset [14]. Optimization is performed using AdamW with a learning rate of 1×10^{-5} for two epochs and a batch size of 128. For LSA, we use $K = 15$ nearest neighbors. Additional implementation details and results on other Stable Diffusion variants [19] are provided in the supplementary material.

Comparison Models. We compare against representative safety interventions for text-to-image generation spanning the main paradigms proposed in recent work. These include inference-time guidance methods such as SLD [20], embedding-space distortion approaches such as DES [1], concept erasure methods including MACE [15], AdvUnlearn [28], STEREO [22], and RECE [8], as well as CLIP-space alignment methods such as Safe-CLIP [17] and SaFeR-CLIP [27]. We also report results from the unmodified base model as a reference. All methods are evaluated using identical generation settings (50 DDIM steps, guidance scale 7.5) for fair comparison.

Evaluation Setup. We evaluate methods along four dimensions: structural fidelity, generative quality, safety robustness, and geometric consistency, For structural fidelity, we use both TIFA [10] and GenEval [7]. TIFA contains ∼4,000 prompts and over 25,000 automatically generated question–answer pairs spanning categories such as objects, attributes, colors, counting, actions, and spatial relations. TIFA decomposes each prompt into visual QA pairs and measures whether they can be correctly answered from the generated image. Following recent practice, we use the Qwen-3-32B MLLM [24] for evaluation. GenEval provides a complementary object-focused evaluation of compositional generation, covering capabilities such as object presence, counting, and color attribution. The benchmark contains 553 prompts. For generative quality, we report FID and CLIPScore [18], both computed on 10k generated samples using prompts from the COCO 30k dataset [3]. For safety robustness, we report Attack Success Rate (ASR), defined as the fraction of adversarial prompts that successfully induce unsafe outputs, detected using the NudeNet classifier [2]. We evaluate under multiple prompt-based attacks including MMA-Diffusion [25], Sneaky Prompt [26], P4D [4], and Ring-A-Bell [23], as well as the I2P sexual benchmark to measure residual unsafe content generation. We also report additional evaluations under white-box attacks in the supplementary material. We further analyze the geometric side effects of safety alignment in the text-embedding space in the supplementary material. Specifically, we report *Embedding Spread Ratio* and *Jaccard similarity* as structural diagnostics. Embedding Spread Ratio measures preservation of embedding spread relative to the base model, while Jaccard similarity measures preservation of local semantic neighborhood structure.

Table 2: Category-wise TIFA evaluation. Existing safety interventions degrade structural fidelity across multiple categories, while our method maintains strong performance (75.4 TIFA). Red highlights indicate the largest drop relative to the base model.

Method	Obj.	Ani.	Loc.	Col.	Food	Mat.	Att.	Cnt.	Sha.	Act.	Spa.	TIFA Avg↑	CLIP↑	FID↓
Base (SD v1.4)	78.9	82.8	89.8	79.9	84.1	83.7	79.6	63.6	58.0	68.2	52.9	76.3	26.5	17.23
State-of-the-Art T2I Safety Methods														
DES	73.2	78.5	85.4	73.7	71.1	74.6	77.4	59.3	53.6	63.5	51.7	71.6 ↓6.2%	25.5	16.23
AdvUnlearn	67.1	69.6	79.7	64.6	68.7	74.6	68.0	44.5	37.7	49.9	41.9	63.1 ↓17.3%	23.9	20.67
STEREO	71.7	75.4	86.9	73.7	79.4	77.5	74.2	**61.9**	52.2	58.3	48.2	69.9 ↓8.4%	24.6	21.69
RECE	**78.4**	80.5	88.0	**79.9**	80.2	**85.7**	77.6	61.1	**66.7**	65.6	50.9	74.8 ↓2.0%	26.0	17.51
MACE	62.9	68.7	79.6	62.5	54.9	68.9	69.6	55.5	46.4	53.9	47.4	62.6 ↓18.0%	23.8	24.87
SafeCLIP	59.0	67.4	79.1	69.5	58.1	75.2	71.4	56.6	50.0	46.6	40.5	60.1 ↓21.2%	22.3	33.40
SafeRCLIP	58.6	66.1	78.2	69.1	58.5	74.6	71.6	55.9	50.7	46.1	43.7	60.7 ↓20.4%	22.4	32.31
SLD	77.3	80.0	88.0	77.8	82.3	81.3	76.8	63.0	52.2	63.2	50.4	73.9 ↓3.1%	25.5	21.85
Our Method														
Ours	77.6	**80.8**	**88.3**	79.6	**83.5**	83.7	**80.1**	61.1	58.0	**66.3**	**53.8**	**75.4** ↓1.2%	**26.4**	**15.93**

4.1 Utility Results

TIFA Utility. Tab. 2 reports structural fidelity results on the TIFA benchmark. Recent safety interventions introduce noticeable degradation in compositional understanding: DES [1] drops 6.2% from the base model while STEREO [22] incurs an even larger 8.4% reduction. These degradations are not uniform across categories. DES loses 13.0% on food-related prompts and STEREO drops 9.9% on activity prompts, highlighting category-specific semantic collapse. In contrast, our method achieves a TIFA score of 75.4, remaining close to the base model (76.3) while improving 5.0% over DES and 7.3% over STEREO. On food-related prompts, our method scores 83.5, nearly recovering base model performance (84.1) and improving substantially over DES (+12%). These results suggest that explicitly preserving the geometric structure of the text embedding space retains fine-grained compositional understanding.

FID and CLIPScore. Our method achieves the lowest FID score (15.93) among all compared approaches, improving over the base model (17.23), indicating stronger distributional alignment with real images. At the same time, it maintains a competitive CLIPScore (26.4), closely matching the base model (26.5), which reflects preserved global text–image alignment. These results indicate that our method improves image distribution quality while preserving text–image alignment.

4.2 Robustness against Adversarial and Unsafe Prompts

Tab. 3 reports robustness across multiple jailbreak attacks as well as direct unsafe prompts. The base model is highly vulnerable, with an average ASR of 67.6. Existing safety methods reduce ASR to varying degrees, but their robustness varies substantially across attacks. Under jailbreak attacks, several approaches remain vulnerable. For instance, SafeCLIP [17] and SaFeRCLIP [27] exhibit high

Table 3: Comparison of Attack Success Rate (ASR) and CLIP Score across different methods. Lower ASR and higher CLIP scores indicate better safety and utility preservation, respectively.

Method	Safety Benchmarks (ASR ↓)						Utility
	MMA	Sneaky	I2P-S	Ring	P4D	Avg. ASR	CLIPScore↑
Base (SD v1.4)	80.4	42.7	34.3	98.1	82.4	67.6	26.5
DES	**0.2**	0.8	1.2	2.8	**0.0**	1.0	25.5
Adv-Unlearn	0.3	0.8	**1.1**	**0.0**	**0.0**	**0.4**	23.9
SafeCLIP	25.2	17.7	24.0	65.4	57.7	38.1	22.3
SafeRCLIP	24.6	16.1	17.9	73.8	43.0	35.1	22.4
STEREO	2.2	3.2	**1.1**	2.8	3.3	2.5	24.6
SLD	74.3	31.5	20.8	98.1	74.3	59.8	25.5
RECE	36.1	6.5	6.0	15.9	26.1	18.1	26.0
MACE	8.6	2.4	6.3	9.4	10.3	7.4	23.8
Ours	0.4	**0.8**	1.2	2.8	1.0	1.2	**26.4**

ASR across multiple attack settings. Inference-time guidance such as SLD performs poorly, producing unsafe generations across most scenarios. Weight-editing approaches such as RECE [8] and MACE [15] reduce ASR further, but their robustness remains inconsistent depending on the attack type. For direct unsafe prompts (I2P-S), most methods suppress explicit content effectively. DES [1] and our method achieve comparable performance (1.2 ASR), indicating similarly strong suppression of direct unsafe prompts. However, methods that push ASR extremely low often incur substantial utility degradation. For example, Adv-Unlearn [28] achieves the lowest average ASR (0.4), but suffers large drops in both CLIP score and structural fidelity, including a 17.3% reduction on TIFA. STEREO [22] also shows higher ASR (2.5) together with degraded utility.

In contrast, our method achieves consistently low ASR across all jailbreak attacks while maintaining strong utility. With an average ASR of 1.2, our approach remains highly competitive with the strongest safety baselines while avoiding the severe utility degradation observed in several prior methods, demonstrating a favorable balance between robustness and utility.

4.3 Structured Benchmark Evaluation on GenEval

Alongside TIFA [10], we also evaluate structured utility using GenEval [7], a benchmark designed to measure compositional generation abilities. This complementary evaluation allows us to examine whether the observed utility degradation generalizes beyond TIFA to another structured benchmark. Tab. 4 reports GenEval performance together with safety results measured using the average attack success rate (Avg. ASR) across safety benchmarks. We evaluate four compositional attributes: single-object generation, color binding, object count, and two-object composition. The Position and Attribute Binding metrics are omit-

Table 4: GenEval compositional generation performance for SD-v1.4 together with safety results. The Avg column reports the relative drop (↓) compared to the base model. The best and second-best scores are highlighted in red and blue, respectively.

Method	Single Object	Color	Count	Two Objects	Avg	Avg. ASR↓
SDv1.4	98.1	73.4	36.9	34.9	60.8	67.6
DES	96.6	68.9	31.3	30.1	56.7 ↓6.8%	1.0
STEREO	88.4	60.4	20.9	17.2	46.7 ↓23.2%	2.5
Adv-Unlearn	96.3	72.9	27.2	20.5	54.2 ↓10.9%	0.4
RECE	96.6	70.0	30.9	27.5	56.2 ↓7.6%	18.1
MACE	80.6	36.4	18.8	8.8	36.2 ↓40.5%	7.4
SafeCLIP	76.4	49.7	21.2	9.2	39.1 ↓35.7%	38.1
SafeRCLIP	76.6	50.8	21.3	9.3	39.5 ↓35.0%	35.1
SLD	97.5	70.2	40.0	25.8	58.4 ↓3.9%	59.8
Ours	97.2	73.4	34.4	34.4	59.8 ↓1.6%	1.2

ted because the base model already achieves extremely low performance on these tasks, making comparisons between safety interventions less informative.

Consistent with the observations from TIFA, many existing safety interventions introduce noticeable degradation in compositional reasoning. For example, DES and RECE reduce the overall GenEval score by 6.8% and 7.6%, respectively, while stronger filtering approaches such as MACE and SafeRCLIP cause substantial drops of 40.5% and 35.0%. Even methods that achieve strong safety performance, such as Adv-Unlearn (Avg. ASR of 0.4), still incur a noticeable decline in compositional generation ability. In contrast, our method maintains strong compositional performance while significantly improving safety. It achieves a GenEval score of 59.8, corresponding to only a 1.6% drop from the base model, while reducing the average ASR to 1.2. These results reinforce the findings from TIFA and show that our approach preserves structured compositional capabilities across multiple evaluation benchmarks while improving safety.

4.4 Ablation Study

We analyze the contribution of each component in our framework by incrementally adding the proposed regularization terms on top of the DES [1] baseline. Tab. 5 reports results across safety benchmarks and utility metrics. The first row corresponds to the original DES model without our geometric regularizers. Adding the *Embedding Spread Preservation (ESP)* loss improves generative utility while maintaining comparable safety performance. In particular, the CoCO utility score increases from 25.5 to 26.2, indicating improved distributional quality without substantially affecting ASR. Applying the *Local Structure Alignment (LSA)* loss alone preserves local semantic relationships but slightly increases ASR across several attacks. Combining both components yields the best overall trade-off. The full model achieves the highest utility (26.4 CoCO)

Table 5: Ablation study of the proposed loss components.

$\mathcal{L}_{ESP}$	$\mathcal{L}_{LSA}^{pert}$	Safety Benchmarks (ASR ↓)						Utility		
		MMA	Sneaky	I2P-S	Ring	P4D	Avg. ASR	CLIPScore ↑	FID ↓	TIFA ↑
✗	✗	0.2	0.8	1.2	2.8	0.0	1.0	25.5	16.23	71.6
✓	✗	0.3	0.8	0.8	1.9	1.7	1.1	26.2	15.74	74.5
✗	✓	0.8	1.6	1.2	2.8	2.0	1.7	26.2	15.99	76.0
✓	✓	0.4	0.8	1.2	2.8	1.0	1.2	**26.4**	**15.93**	75.4

while maintaining competitive robustness with an average ASR of 1.2. These results indicate that ESP stabilizes the overall embedding geometry while LSA preserves local semantic relationships, and their combination effectively balances safety and utility. Additional ablations for different variants of the ESP and LSA losses are provided in the supplementary material.

5 Conclusion

In this work, we show that commonly used utility metrics for T2I safety alignment obscure substantial fine-grained utility degradation. Our analysis attributes this behavior to a contraction of the embedding space and distortions in local similarity structure, a phenomenon we term *Semantic Collapse*. To address this issue, we introduce SAGE (Structure-Aware Geometric Regularization), a geometry-aware alignment framework that regularizes embedding spread and preserves local relational structure. Our results demonstrate that SAGE restores structured utility while maintaining strong safety robustness. These findings highlight the importance of preserving inter-embedding geometric relationships when aligning models for safety.

Acknowledgment

A. S. Bedi acknowledges the support of the Defense Advanced Research Projects Agency (DARPA) under Cooperative Agreement No. HR0011262E011. The content of this information does not necessarily reflect the position or policy of the U.S. Government, and no official endorsement should be inferred.

References

1. Ahn, J., Jung, H.: Mitigating sexual content generation via embedding distortion in text-conditioned diffusion models (2025), https://arxiv.org/abs/2501.18877 2, 5, 7, 8, 9, 10, 11, 12, 13, 14
2. Bedapudi, P.: Nudenet: Neural nets for nudity classification, detection and selective censoring (2019) 11
3. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollar, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server (2015), https://arxiv.org/abs/1504.00325 11

4. Chin, Z.Y., Jiang, C.M., Huang, C.C., Chen, P.Y., Chiu, W.C.: Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts (2026), https://arxiv.org/abs/2309.06135 11
5. Fan, C., Liu, J., Zhang, Y., Wong, E., Wei, D., Liu, S.: Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation (2024), https://arxiv.org/abs/2310.12508 5
6. Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., Bau, D.: Erasing concepts from diffusion models (2023), https://arxiv.org/abs/2303.07345 5
7. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment (2023), https://arxiv.org/abs/2310.11513 4, 11, 13
8. Gong, C., Chen, K., Wei, Z., Chen, J., Jiang, Y.G.: Reliable and efficient concept erasure of text-to-image diffusion models (2024), https://arxiv.org/abs/2407.12383 5, 11, 13
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium (2018), https://arxiv.org/abs/1706.08500 2
10. Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., Smith, N.A.: Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering (2023), https://arxiv.org/abs/2303.11897 2, 4, 5, 11, 13
11. Kim, C., Min, K., Yang, Y.: R.a.c.e.: Robust adversarial concept erasure for secure text-to-image diffusion model (2024), https://arxiv.org/abs/2405.16341 5
12. Li, X., Yang, Y., Deng, J., Yan, C., Chen, Y., Ji, X., Xu, W.: Safegen: Mitigating sexually explicit content generation in text-to-image models. In: Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security. pp. 4807–4821 (2024) 2
13. Liu, R., Chen, I.C., Gu, J., Zhang, J., Pi, R., Chen, Q., Torr, P., Khakzar, A., Pizzati, F.: Alignguard: Scalable safety alignment for text-to-image generation (2025), https://arxiv.org/abs/2412.10493 5
14. Liu, R., Khakzar, A., Gu, J., Chen, Q., Torr, P., Pizzati, F.: Latent guard: a safety framework for text-to-image generation (2024), https://arxiv.org/abs/2404.08031 11
15. Lu, S., Wang, Z., Li, L., Liu, Y., Kong, A.W.K.: Mace: Mass concept erasure in diffusion models (2024), https://arxiv.org/abs/2403.06135 5, 11, 13
16. OpenAI: Improving image generation with better captions. OpenAI Technical Report (2023), https://cdn.openai.com/papers/dall-e-3.pdf 1
17. Poppi, S., Poppi, T., Cocchi, F., Cornia, M., Baraldi, L., Cucchiara, R.: Safe-clip: Removing nsfw concepts from vision-and-language models (2024), https://arxiv.org/abs/2311.16254 2, 5, 7, 8, 11, 12
18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021) 11
19. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2022), https://arxiv.org/abs/2112.10752 1, 11
20. Schramowski, P., Brack, M., Deiseroth, B., Kersting, K.: Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22522–22531 (June 2023) 1, 2, 11

21. Schramowski, P., Brack, M., Deiseroth, B., Kersting, K.: Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models (2023), https://arxiv.org/abs/2211.05105 5
22. Srivatsan, K., Shamshad, F., Naseer, M., Patel, V.M., Nandakumar, K.: Stereo: A two-stage framework for adversarially robust concept erasing from text-to-image diffusion models (2025), https://arxiv.org/abs/2408.16807 2, 5, 11, 12, 13
23. Tsai, Y.L., Hsu, C.Y., Xie, C., Lin, C.H., Chen, J.Y., Li, B., Chen, P.Y., Yu, C.M., Huang, C.Y.: Ring-a-bell! how reliable are concept removal methods for diffusion models? (2024), https://arxiv.org/abs/2310.10012 11
24. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z.: Qwen3 technical report (2025), https://arxiv.org/abs/2505.09388 11
25. Yang, Y., Gao, R., Wang, X., Ho, T.Y., Xu, N., Xu, Q.: Mma-diffusion: Multimodal attack on diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7737–7746 (2024) 11
26. Yang, Y., Hui, B., Yuan, H., Gong, N., Cao, Y.: Sneakyprompt: Jailbreaking text-to-image generative models (2023), https://arxiv.org/abs/2305.12082 11
27. Yousaf, A., Fioresi, J., Beetham, J., Bedi, A.S., Shah, M.: Safer-clip: Mitigating nsfw content in vision-language models while preserving pre-trained knowledge (2025), https://arxiv.org/abs/2511.16743 2, 5, 7, 8, 11, 12
28. Zhang, Y., Chen, X., Jia, J., Zhang, Y., Fan, C., Liu, J., Hong, M., Ding, K., Liu, S.: Defensive unlearning with adversarial training for robust concept erasure in diffusion models. Advances in neural information processing systems **37**, 36748–36776 (2024) 2, 5, 7, 8, 11, 13
29. Zhang, Y., Jia, J., Chen, X., Chen, A., Zhang, Y., Liu, J., Ding, K., Liu, S.: To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In: European Conference on Computer Vision. pp. 385–403. Springer (2024) 1, 2