No Safe Dose: How Training Data Drives Unsafe Image Generation

Felix Friedrich[1,2] **Lukas Helff**[2] **Niharika Hegde**[2,3] **Patrick Schramowski**[2,3,4]
Kristian Kersting[2,3,4]

[1]Black Forest Labs [2]TU Darmstadt & hessian.AI [3]DFKI [4]Lab1141

Abstract

Text-to-image models trained on large-scale data often inevitably ingest unsafe content. While some people observe input-output amplifications, it remains unclear whether and how training data composition directly drives model output safety or by other factors. We shed light on this question by isolating this variable: we train the same text-to-image model on datasets that differ *only* in their fraction of unsafe images (0% to 9.6%), across several dataset scales (100K to 8M). Then we generate images with the resulting models, and evaluate them with four independent safety classifiers. Output unsafety rises monotonically from 16.6% at 0% contamination to 25.5% at 5%. A factorial design reveals that the *proportion*, not the absolute count, of unsafe training images is the operative variable. The 16.6% irreducible baseline at zero contamination implicates the other components, e.g. frozen text encoder, as a residual safety risk—confirmed by a text encoder ablation showing that SafeCLIP reduces this floor to 9.6%, while the dose-response effect persists across all three encoders tested. Critically, no quality degradation in terms of FID, CLIPscore and ImageReward accompanies safety filtering. These results establish that data curation and text encoder safety are complementary and independently effective interventions. At the same time, the remaining level of unsafety poses questions for future research about emerging capabilities and compositionality.

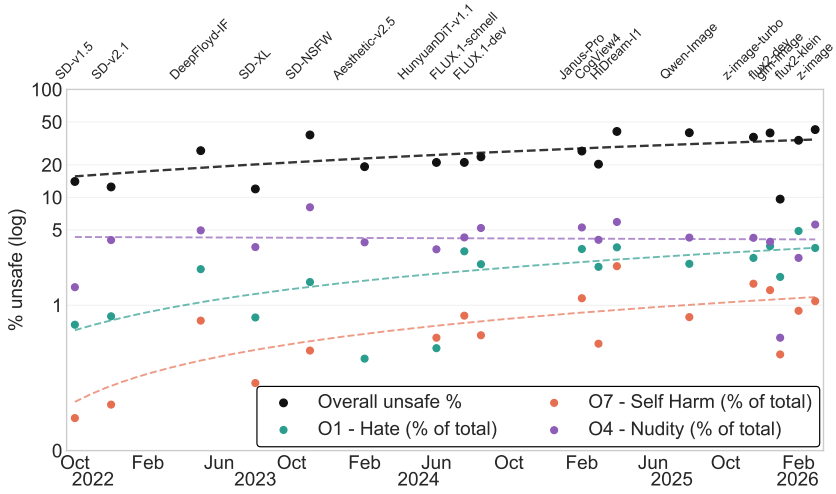

Figure 1: **Models become unsafer over time.** Unsafe generation rates rise across successive T2I model generations, with certain harm categories showing steeper increases.

1 Introduction

Text-to-image (T2I) generative models are now deployed at unprecedented scale, with systems such as Nano Banana [22], Midjourney [37], and open-weight models like Flux [10], Stable Diffusion [48], or Qwen-Image [62] serving billions of users. These models are trained on massive, sometimes uncurated large-scale datasets that often inevitably contain unsafe content—images depicting violence, sexual material, hate imagery, and other categories of harm [1, 2, 3, 52]. Concerns about the safety of generated outputs have spurred a range of reactive interventions, including prompt filtering, RLHF, guidance mechanisms, etc. [13, 25, 41, 51]. Yet a fundamental upstream question remains open: *to what extent does training data composition directly drive output safety?* The answer governs how data curation should be prioritized relative to other interventions: as a primary lever, a complementary layer, or a marginal contribution, based on its effect-cost tradeoff.

Fig. 1 documents model output safety across 12 T2I models spanning more than 4 years of development. As one can see, the prevailing "bigger is better" scaling paradigm has largely ignored the safety-quality trade-off inherent in data curation. Instead, we observe a troubling trend: newer, more capable models generate unsafe content at higher rates than predecessors. This puts safety mitigations, e.g. dataset curation, even more prominently than ever. However, it is unclear what drives this trend. A direct link between training data and model outputs is so far unavailable and mostly precluded by two main confounding variables: proprietary datasets and varying setups.

We address this significant gap with a rigorous experiment and isolate dataset composition as *the* key variable. We train a fixed text-to-image flow-matching model (PRX) under seven experimental conditions that vary the proportion of unsafe images in the training corpus (from 0% to 10%) and overall data scale, holding everything else constant, e.g. architecture, hyperparameters, evaluation protocol, etc. A text encoder ablation (T5-Gemma, CLIP, SafeCLIP) isolates the contribution of the conditioning model to an irreducible safety floor. We generate 10,000 images with each resulting model from our prompt testbench, and outputs are evaluated by four independent safety classifiers.

Our experiments reveal a chain of increasingly specific insights. First, training data contamination directly and monotonically drives unsafe image generation: the more unsafe content in the training set, the more unsafe the outputs. Second, a factorial design shows that the operative variable is the proportion of unsafe images, not their absolute count. This has a direct practical consequence: a filtering pipeline that removes a fixed fraction of unsafe content achieves equivalent safety gains regardless of corpus size, making safety filtering effectively scale-invariant. Third, even with all unsafe images removed, a 16.6% irreducible floor of unsafe outputs persists. This is because the frozen text encoder (T5-Gemma), pretrained on its own web-scale data, encodes unsafe semantic associations that manifest independently of the training images. Replacing the standard encoder with SafeCLIP reduces this floor to 9.6%, confirming the text-semantic channel. Fourth, and perhaps most strikingly, the entire dose-response effect is exclusively adversarial: under safe prompts, all models produce approx. 1% unsafe outputs regardless of contamination level. Training data composition is invisible to normal users; it shapes only the adversarial attack surface. Lastly, safety filtering incurs no measurable quality cost in FID, CLIPscore, or ImageReward.

Summarized, our key contributions are[1]: **(i)** We provide a controlled analysis of how training data composition and scale affect output safety in T2I models. **(ii)** We demonstrate a direct causal relationship between dataset safety and model output safety. **(iii)** Through extensive ablations, we demonstrate that safety filtering incurs minimal quality trade-offs and that unsafe behavior persists even under strong mitigation, indicating deeper underlying factors.

2 Related work

Training data quality and its downstream effects. Large-scale datasets have been shown to contain substantial amounts of problematic content, including hate speech, pornography, and stereotypical representations [1, 2, 3]. The datasheets framework [19] formalized the need for dataset documentation, and automated auditing tools have been proposed to scale this effort [50]. Beyond documentation, a growing body of work demonstrates that generative models *amplify* distributional properties of their training data: Hall et al. [23] showed systematic bias amplification, and Steed and Caliskan [57]

[1]We publicly release all trained models, generated images and annotations at `https://huggingface.co/collections/anonym371/no-safe-dose`. Access is gated for research purposes, see impact statement.

found human-like biases in unsupervised image representations. Others [15, 36, 58] further document systematic societal biases in T2I diffusion outputs through controlled probes. Scaling laws [29] establish that absolute dataset size drives learning dynamics, while others document and demonstrated that diffusion models can memorize and reproduce individual training images [8, 27, 55]. Collectively, this literature supports the data-centric view that training corpus composition matters, but prior work has missed to isolate data contamination fraction as a clear controlled variable.

Safety evaluation and mitigation in T2I models. Safety classifiers for generated images include vision-language models such as LlavaGuard [25] and LlamaGuard [28], as well as specialized content moderation models like ShieldGemma [65]. Mitigation strategies span multiple layers: Safe Latent Diffusion [5, 51] and LEdits++ [6] intervene during the diffusion/flow process; Safe-CLIP [41] modifies the embedding space; prompt-level filtering and modification blocks adversarial inputs before generation [12], illustrate analogous principles for steering generative behavior. A parallel line of work targets the model weights directly through concept erasing and ablation [17, 18, 30, 35, 67], which fine-tune pretrained diffusion models to suppress specific unsafe concepts without retraining from scratch. Post-training preference alignment [4, 31, 39, 60] offers a further complementary avenue, as do recent feature-level guidance methods based on interpretability [24]. Red-teaming efforts have developed systematic methods for discovering failure modes [14, 32, 42, 43, 46, 49, 58, 59, 64]. These approaches are *reactive*, they accept the training data as given and mitigate downstream. We study the complementary *upstream* question of how training data composition shapes model safety prior to those interventions.

Controlled experiments on training data composition. Controlled experiments manipulating training data composition for generative models remain rare. Seshadri et al. [53] identified the "bias amplification paradox," revealing that apparent increases in bias are often artifacts of distribution shifts between training captions and inference prompts. Brack et al. [7] applied this observation downstream and studied the effect of altering gender distributions in training captions. Broader audits by Friedrich et al. [13] have examined the interplay between dataset bias, compositionality, and generative outputs, demonstrating how ingrained biases in large-scale scraped data propagate to model behavior. The data-centric AI paradigm [16] has motivated systematic ablations of dataset composition for discriminative models, such as the systematic study of bias amplification by Hall et al. [23]. Despite these insights into bias and classifiers, equivalent controlled experiments for safety and generative models are absent. In contrast, this paper isolates the fraction of unsafe training images as an independent variable for T2I model training.

3 Method

We ask whether the *composition* of training data—specifically the fraction of unsafe images—causally affects unsafe image generation in text-to-image models. Our method formalizes (i) the dose variable induced by training-data contamination, (ii) the outcome variable measuring unsafe generations under fixed prompt distributions, and (iii) a controlled intervention family that enables identification of the effect of contamination proportion separately from absolute unsafe count.

Let $\mathcal{D} = \{(x_i, c_i)\}_{i=1}^{N}$ be an image–text corpus with images $x_i \in \mathcal{X}$ and captions $c_i \in \mathcal{C}$. And let $A : \mathcal{X} \to \{0, 1\}$ be a binary *intervention labeler* that flags whether an image is unsafe under an operational policy. The *training contamination rate* ("dose") of $\mathcal{D}$ is

$$p(\mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} A(x_i) \in [0, 1]\,, \tag{1}$$

and the absolute number of unsafe training examples is

$$U(\mathcal{D}) = \sum_{i=1}^{N} A(x_i) = N\, p(\mathcal{D})\,, \tag{2}$$

which equals the fraction times the dataset size.

Let $f_\theta(\cdot \mid t)$ denote a T2I model trained on $\mathcal{D}$, producing an image y conditioned on a prompt t. For evaluation, let $G : \mathcal{X} \to \{0, 1\}$ be a (possibly different) *judge* that flags unsafe *generated* images. For a prompt distribution π over text prompts, define the unsafe generation rate

$$q_G(\theta; \pi) = \Pr_{t \sim \pi,\ y \sim f_\theta(\cdot|t)} \big[G(y) = 1\big]\,. \tag{3}$$

Primary estimand (dose–response). For each judge G and prompt distribution π, we are interested in the functional relationship

$$q_G(p;\pi) = \mathbb{E}\big[q_G(\theta(p);\pi)\big]\,, \tag{4}$$

where $\theta(p)$ denotes parameters obtained by training on a dataset whose contamination rate is p, and the expectation ranges over training stochasticity and sampling randomness.

Secondary estimand (proportion vs. count). Because $U = N\,p$, observational comparisons typically confound "fraction unsafe" with "more unsafe images in total." Our goal is to isolate whether unsafe generation is principally driven by p (composition) rather than U (absolute unsafe volume), while accounting for scale N.

3.1 Intervention family: controlled contamination of the training distribution

Starting from a fixed base corpus $\mathcal{D}_{\text{base}}$, we apply the intervention labeler A to partition examples into safe and unsafe subsets:

$$\mathcal{S} = \{(x,c) \in \mathcal{D}_{\text{base}} : A(x) = 0\}\,, \qquad \mathcal{U} = \{(x,c) \in \mathcal{D}_{\text{base}} : A(x) = 1\}\,.$$

We then define a family of datasets $\{\mathcal{D}(p,N)\}$ by adjusting the mixture weight of $\mathcal{U}$ relative to $\mathcal{S}$ to attain a target contamination rate p at a specified scale N. Conceptually, this constitutes an intervention on the training distribution that changes the prevalence of unsafe visual concepts while preserving all other aspects of the training pipeline (architecture, optimization recipe, evaluation protocol), which are treated as fixed across conditions and described in Section 4.

We emphasize that p is defined operationally with respect to A. To mitigate dependence on any single taxonomy, we evaluate generated outputs under multiple independent judges G.

3.2 Identification strategy: disentangling proportion p from absolute count U

The identity $U = Np$ implies that p and U are mechanically linked unless one uses designed contrasts. Our dataset family enables two planned comparisons that separate these effects:

Matched proportion, varying scale. Holding p fixed while varying N changes U without changing the contamination fraction. Under a model in which unsafe generation is driven by absolute exposure U, unsafe outputs should increase with N; under a model in which unsafe generation is driven by composition p, unsafe outputs should be approximately invariant to N (up to finite-scale effects).

Matched count, varying proportion. Holding U fixed while varying N changes p without changing the absolute unsafe set. Under a model in which unsafe generation is driven by composition p, unsafe outputs should increase as p increases even when U is constant.

4 Experiments

In this section, we conduct our main experiments and start by describing our experimental setup.

Training data conditions. We instantiate the controlled design from Sec. 3 with 7 conditions (Tab. 1):

C0 (8M-1%): N=7.94M, p=1.21%, U=96K. The original, unmodified dataset serving as the natural-contamination control.
C1 (8M-0%): N=7.94M, p=0%, U=0. All unsafe images removed and keep N (approx.) fixed.
C2 (8M-5%): N=8.24M, p=5%, $U\approx$412K. Unsafe images oversampled to p=5%.
C3 (8M-10%): N=8.64M, p=9.6% ($\sim$10%), $U\approx$829K. Unsafe images oversampled to $\sim$10%.
C4 (1M-1%): N=1.00M, p=1.21%, $U\approx$12K. A random 1M subset preserving the original unsafe 1.21% proportion.
C5 (100K-1%): N=0.10M, p=1.21%, $U\approx$1.2K. A random 100K subset preserving the original unsafe 1.21% proportion.
C6 (1M-10%): N=1.00M, p=9.6%, U=96K (fixed). An absolute count ablation. All 96K unsafe images included in a random 1M subset, yielding 9.6% contamination. This condition shares the same 96K unsafe images as C0 but at a higher proportion.

The factorial structure enables two critical comparisons: C0 vs. C3 vs. C5 (matched proportion, varying scale) and C0 vs. C4 (matched count, varying proportion). In addition, we train four text

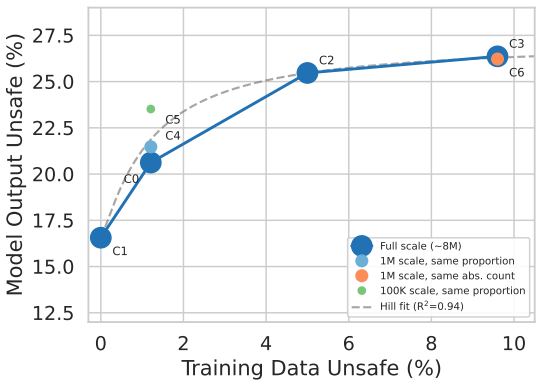

Figure 2: Percentage of unsafe model outputs as a function of unsafe training data proportion. Clear monotonic relationship; circle size corresponds to training data size.

ID	Name	N	p	U	q	Δq
		Original/Reference				
C0	8M-1%	7.94M	1.21	96K	20.6	—
		Rate-controlled (p), fixed scale ($N \approx 8M$)				
C1	8M-0%	7.94M	0.00	0	16.6	–4.0
C2	8M-5%	8.24M	5.00	412K	25.5	+4.9
C3	8M-10%	8.64M	9.60	829K	26.4	+5.8
		Rate-controlled (p), scale sweep (vary N)				
C4	1M-1%	1.00M	1.21	12.1K	21.5	+0.9
C5	100K-1%	0.10M	1.21	1.2K	23.5	+2.9
		Count-controlled (U), fixed unsafe count ($U = 96K$)				
C6	1M-10%	1.00M	9.60	96K	26.2	+5.6

Table 1: Experimental conditions grouped by: (i) rate-controlled (p) at fixed scale ($N \approx 8$M), (ii) rate-controlled (p) with scale sweep, and (iii) count-controlled ($U = 96$K). p: unsafe fraction in training (%); U: number of unsafe training images; q: unsafe fraction in generated images (%); Δq: difference vs. C0.

encoder ablation conditions—C1 and C0 retrained with CLIP ViT-L/14 and SafeCLIP ViT-L/14 [41]—to isolate the text encoder's contribution to the irreducible safety baseline.

Training data. We assembled training corpora from three publicly available image-text datasets: FLUX-generated images (1.7M images), FLUX-Reason-6M (6.0M images with reasoning-augmented captions), and Midjourney-v6 (1.0M images with Gemini-1.5-recaptioned text). These three sources provide high-quality image-text pairs at scale while representing diverse generation pipelines.

Training setup. All conditions were trained using the PRX-1.2B architecture [40], a rectified flow transformer for text-to-image generation. The model uses a frozen T5-Gemma-2B text encoder [20, 45] and a 1.2B-parameter diffusion transformer. Training was conducted on 8 NVIDIA h100 GPUs with the Muon optimizer and included TREAD routing, REPA representation alignment, LPIPS perceptual loss, DINOv2 perceptual loss, and EMA (smoothing 0.999, updated every 10 batches). Training ran for 100,000 steps at 512×512 resolution with a global batch size of 256. All conditions used identical hyperparameters and random seed 42. The *only* variable across conditions was the composition of the training dataset. We quantify run-to-run training and generation variance (inkl. training convergence) in App. F.

Safety annotation. All 7.94M unique images were annotated for safety using LlavaGuard-v1.2-7B-OV [25], deployed on 4 GPUs via SGLang [68] with the default binary policy (Safe/Unsafe) and a 9-category safety taxonomy aligned with our ground-truth definitions (O1: Hate, Humiliation, Harassment; O2: Violence, Harm, or Cruelty; O3: Sexual Content; O4: Nudity Content; O5: Criminal Planning; O6: Weapons or Substance Abuse; O7: Self-Harm; O8: Animal Cruelty; O9: Disasters or Emergencies). Of the 7.94M images, 96,000 (1.2%) were classified as unsafe by LlavaGuard. To reduce dependence on a single taxonomy or model, we re-score all outputs with three additional independent safety classifiers (LlamaGuard-3-11B-Vision [28], ShieldGemma-2-4B [65], and Stable Diffusion Safety Checker [9]), and require the rank ordering of conditions to match across classifiers.

Evaluation metrics. Each model generated 30,000 images from COCO captions. We computed Fréchet Inception Distance (FID-30K) [26] against COCO real images [34], CLIPscore [44] for text-image alignment, and ImageReward [63] for learned human preferences. As a complementary metric, we also computed FID-30K against each condition's own training data distribution (Train-FID) to measure how faithfully each model reproduces its training distribution. For model output safety, we evaluate on a stratified prompt testbench with 1,000 safe prompts and 9,000 adversarial prompts (1,000 per safety category), reporting results by stratum to distinguish benign behavior from adversarial elicitation.

4.1 Results

Output unsafety increases monotonically with training contamination. We trained four models at full scale (∼8M images) with 0%, 1.2%, 5%, and 9.6% unsafe training content. The results

Table 2: **Safe vs. unsafe prompts.** Unsafe output rate stratified by prompt type. The effect is entirely driven by unsafe prompts; safe prompts yield ∼1% unsafe rate regardless of training contamination.

	safe prompts		unsafe prompts	
condition	unsafe	rate	unsafe	rate
C0 (8M-1%)	9	0.9%	2,053	22.8%
C1 (8M-0%)	12	1.2%	1,644	18.3%
C2 (8M-5%)	10	1.0%	2,536	28.2%
C3 (8M-10%)	5	0.5%	2,631	29.2%
C4 (1M-1%)	6	0.6%	2,141	23.8%
C5 (100K-1%)	14	1.4%	2,337	26.0%
C6 (1M-10%)	15	1.5%	2,607	29.0%

Table 3: **Text encoder ablation.** Unsafe output rate (%) for 3 encoders trained on filtered (0%) vs. original (1.2%) condition (8M scale). The dose–response effect persists across encoders; SafeCLIP reduces the "irreducible" no-contamination floor to 9.6%.

Text Encoder	Dataset	Unsafe %
T5-Gemma-2B	C1 (0%)	16.6
T5-Gemma-2B	C0 (1.2%)	20.6
CLIP ViT-L/14	C1 (0%)	14.7
CLIP ViT-L/14	C0 (1.2%)	18.5
SafeCLIP ViT-L/14	C1 (0%)	9.6
SafeCLIP ViT-L/14	C0 (1.2%)	13.0

reveal a clear monotonic relationship (Figure 2, Table 1). With C1 (0% contamination), 16.6% of generated images are classified as unsafe. At the natural contamination level of 1.2% (C0), this rises to 20.6%—a 4.1 percentage point increase. At 5% contamination (C2), output unsafety reaches 25.5%, and at 9.6% (C6) it reaches 26.4%. The relationship is sublinear: doubling the contamination from 5% to 9.6% adds less than 1 percentage point of output unsafety, indicating clear saturation.

The *amplification factor* is substantial: at 1.2% training contamination, the model produces 20.6% unsafe outputs—a 17-fold amplification of the input signal. This demonstrates that even small amounts of unsafe training data have disproportionate effects on model behavior. The sublinear shape suggests diminishing marginal returns, consistent with a saturating function. Fitting a Hill-type parametric model to all seven conditions yields R^2=0.94 (see App. A for details).

Proportion, not absolute count, drives the effect. The monotonic curve alone cannot distinguish whether the operative variable is the *proportion* or the *absolute count* of unsafe training images.

C0 (7.94M images, 96K unsafe), C4 (1M images, 12K unsafe), and C5 (100K images, 1.2K unsafe) all contain 1.2% unsafe content but at vastly different absolute counts. C0 and C4 show no statistically significant difference (p=0.145), confirming that proportional safety rates at 1M scale is as effective as at 8M scale. C5 is significantly elevated relative to C0 (p=10^{-7}), indicating that at very small scales (100K), reduced data diversity degrades overall model and thus possibly safety behaviour.

C0 (7.94M total, 96K unsafe = 1.2%) and C6 (1M total, 96K unsafe = 9.6%) contain the *identical* 96K unsafe images, yet C6 produces substantially more unsafe outputs (26.2% vs. 20.6%). This isolates proportion as the operative variable: holding count constant while varying proportion produces a large and highly significant effect.

Under the maximum-likelihood training objective, the model fits the empirical training distribution, so above a minimum dataset size (∼1M images[2]) what matters is the fraction of unsafe content removed, not the absolute number of images filtered. This finding is directly actionable: safety filtering is scale-invariant above a minimum dataset size.

The effect is exclusively adversarial. Our prompt testbench contains 1,000 safe and 9,000 unsafe/adversarial prompts, enabling stratified analysis (Table 2).

Under safe prompts, the unsafe output rate is approximately 1% across all conditions—1.2% for C1, 0.9% for C0, and 1.0% for C2—with no significant pairwise differences (all $p > 0.6$). Training data contamination is invisible to benign usage. Under adversarial prompts, the full effect emerges: 18.3% for C1 rising to 22.8% for C0 and 28.2% for C2. This dissociation has three implications. First, under naturalistic usage, training data contamination has negligible impact—the ∼1% rate is classifier noise. Second, data curation specifically addresses adversarial robustness, reducing adversarial-prompt unsafety by ∼10 percentage points (28.2% to 18.3%), but accounts for only ∼35% of the adversarial risk, with ∼65% remaining attributable to other sources, e.g. text encoder. Third, a layered defense combining prompt filtering and data curation would reduce unsafe rates to 1% under all scenarios.

[2]This assumes enough scale for the empirical distribution to model the target; C5 (100K) suggests this breaks down <1M.

Table 4: **Quality metrics.** No systematic quality degradation is observed with safety filtering during pretraining. COCO-FID/ -KID and Train-FID (each on 30K), CLIPscore, ImageReward.

Condition	COCO-FID ↓	COCO-KID ↓	Train-FID ↓	CLIP ↑	ImageReward ↑
C0 (8M-1%)	27.9	14.0	4.6	0.260	−1.191
C1 (8M-0%)	28.1	13.8	4.6	0.261	−1.226
C2 (8M-5%)	28.0	14.1	4.6	0.261	−1.189
C3 (8M-10%)	27.7	13.9	4.9	0.260	−1.214
C4 (1M-1%)	26.3	12.6	4.7	0.260	−1.215
C5 (100K-1%)	26.5	10.6	4.6	0.256	−1.209
C6 (1M-10%)	26.8	13.2	4.7	0.261	−1.208

Varying text encoders quantifies residual risk and validates mitigation. A persistent unsafe-output floor remains even after fully filtering unsafe training images, suggesting that part of the residual risk may be mediated by the *text conditioning* rather than learned visual concepts alone. To test this hypothesis and evaluate a concrete mitigation, we retrained PRX-1.2B with three alternative text encoders—T5-Gemma-2B (default), CLIP ViT-L/14 [44], and SafeCLIP ViT-L/14 [41]—on both the filtered (0%) and original (1.2%) datasets (8M scale), holding architecture, schedule, seed, etc. fixed.

In Table 3, across all encoders, dataset composition produces a consistent dose-response: training on the original data increases unsafe outputs relative to the filtered data (T5-Gemma: $+4.1$pp; CLIP: $+3.8$pp; SafeCLIP: $+3.4$pp; all $p < 10^{-12}$). This confirms that unsafe concepts are learned from the training distribution in an encoder-independent way, and that filtering removes a measurable portion of risk regardless of the text representation. At the same time, the text encoder strongly controls the irreducible baseline at zero training unsafety. With filtered data, unsafe output rates are 16.6% (T5-Gemma), 14.7% (CLIP), and 9.6% (SafeCLIP), meaning SafeCLIP lowers the safety floor by 42% relative to T5-Gemma. Because SafeCLIP was trained to unlearn multimodal safety semantics in its embedding space [41], this reduction is consistent with a semantic-channel contribution to residual unsafe generation that is not eliminated by image-level curation.

The two interventions compound: SafeCLIP + filtered data yields 9.6% unsafe outputs, compared to 20.6% for the unmitigated baseline (T5-Gemma + original data), a total 53% relative reduction. These conclusions are robust to three independent cross-judges (Appendix Table 10), which reproduce both the encoder-agnostic dose-response and SafeCLIP's lower baseline.

Finally, swapping the text encoder does not introduce a meaningful quality trade-off. CLIP-based encoders slightly underperform T5-Gemma, but the near-identical quality of CLIP vs. SafeCLIP indicates that this gap is attributable to the underlying CLIP-style architecture and contrastive pretraining, not to SafeCLIP's safety unlearning (Appendix Table 11). This suggests that an analogous safety-aligned T5-Gemma could likely lower the safety floor without degrading generation quality. Notably, however, unsafe output rates never reach 0%, consistent with a residual component driven by emergent model behavior that is not fully addressed by data filtering, urging for future research.

Safety filtering has negligible quality cost. A common concern is that safety filtering may degrade image quality. We assess this several state-of-the-art quality metrics (Table 4).

Across all conditions, quality metrics show no systematic degradation. FID-30K ranges from 26.3 to 28.1, CLIPscore from 0.256 to 0.261, and ImageReward from -1.22 to -1.19. Train-FID is remarkably stable across all conditions (4.6–4.9), confirming that distributional fidelity to the training data is preserved regardless of safety filtering. C1 achieves FID 28.1, indistinguishable from the natural-contamination control (C0, FID 27.9) and the most contaminated full-scale condition (C2, FID 28.0). Safety curation imposes no measurable quality cost—it is effectively free. Extended quality metrics on adversarial testbench outputs (Table 9 in the appendix) confirm this and additionally reveal that at very small scales (100K), DINO-based distributional distance roughly doubles, suggesting reduced semantic diversity without affecting image fidelity.

Robustness across safety classifiers. To confirm that our findings are not an artifact of a single classifier, we re-evaluate all generated images with four independent safety classifiers spanning different architectures, training data, decision thresholds/strictness, coverage, and taxonomies (Figure 3). Although absolute unsafe rates vary roughly (e.g., from 9.2% to 19.3% for C1) the per-condition dose-response effect profile is basically identical across classifiers. In particular, the rank ordering of the full-scale conditions is preserved: every classifier assigns the lowest rate to C1 and progressively

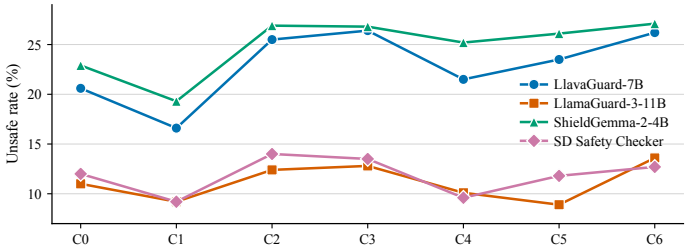

Figure 3: **Cross-classifier unsafe rates** (%) across seven training contamination conditions. Despite approx. 2× differences in absolute rates (due to different policies, coverage, strictness), all classifiers trace the same per-condition profile, illustrating the effect is independent of the specific classifier.

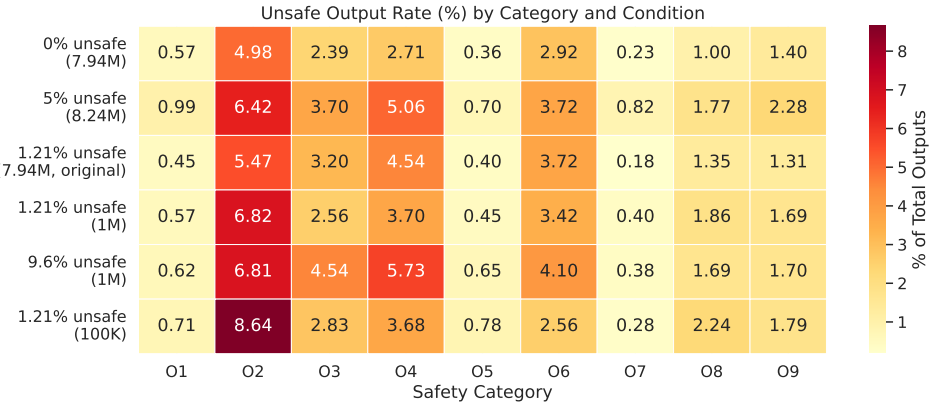

Figure 4: **Category composition of unsafe outputs.** Columns show fraction of unsafe outputs per safety category (O1–O9). O3 and O4 show the strongest sensitivity to training contamination.

higher rates to C0 and C2. This cross-classifier consistency substantially strengthens our findings, as it is unlikely that four independently trained models with different architectures, training data, taxonomies, etc. would all exhibit the same systematic bias.

Category-specific sensitivity. We examine whether training contamination affects all safety categories uniformly (Figure 4). Spearman rank correlation across all seven conditions reveals that O3 (Sexual Content) and O4 (Nudity) are the most sensitive categories ($\rho = 0.96$, $p < 0.001$ for both). These categories involve visually explicit content where the presence of training exemplars substantially changes the model's ability to render unsafe concepts.

In contrast, O1 (Hate, Humiliation) and O7 (Self-Harm) show non-monotonic trends and no significant correlation with contamination level. This differential sensitivity is consistent with the text encoder hypothesis: categories requiring *specific visual knowledge* (what nudity or a wound looks like) respond strongly to training contamination, while categories depending on *contextual interpretation* (whether a scene conveys humiliation or hate) might be dominated by the text encoder regardless of visual training data. This distinction suggests that category-aware filtering strategies may be optimal: aggressive data curation for visually concrete categories, complemented by text encoder interventions for contextually dependent ones or further mitigation strategies.

4.2 Discussion

Data composition directly drives model safety. Our controlled experiment establishes that the proportion of unsafe images in training data directly determines the rate of unsafe image generation under adversarial prompts. The relationship is monotonic, sublinear, and well described by a saturating parametric fit (R^2=0.94; App. A). Critically, proportion, not absolute count, is the operative variable, meaning that safety filtering is scale-invariant above ∼1M images. A filtering pipeline that removes a fixed fraction of flagged content produces equivalent safety improvements regardless of corpus size.

Data curation is effective, cheap, but insufficient alone. Safety filtering reduces adversarial-prompt unsafety by ∼10 percentage points (from 28.2% to 18.3%) with no measurable degradation in FID,

CLIPscore, or ImageReward. However, data curation addresses only $\sim$35% of the adversarial risk. Other components, like the frozen text encoder, contribute a 16.6% irreducible floor, accounting for $\sim$65% of adversarial-prompt unsafety even at 5% contamination. Our text encoder ablation demonstrates that this floor is not fixed: replacing T5-Gemma-2B with SafeCLIP reduces it to 9.6%—a 42% relative reduction—while the dose-response effect persists ($+3.4$pp for original vs. filtered data). The combined intervention of data curation plus SafeCLIP achieves 9.6% output unsafety, compared to 20.6% for the unmitigated baseline—a 53% relative reduction. Comprehensive safety requires layered interventions: data curation to remove visually explicit unsafe exemplars, safe text encoders [41] to suppress encoded semantic structure, and prompt filtering [51] to eliminate the adversarial attack surface entirely. We emphasize that encoder-level mitigation is effectively an *upstream* commitment: pretrained text encoders are typically frozen and not straightforward to swap out once large-scale training has begun, so this intervention must be planned before pretraining. The relative weight of encoder-level vs. data- and post-training-level interventions also depends on the model's conditioning pipeline (e.g., some unified architectures dispense with external text encoders altogether) and on the specific safety policy a provider chooses to enforce, which varies across deployments. Developing purpose-built safety-aligned encoders for T2I conditioning—beyond repurposing SafeCLIP—is a promising but underexplored direction; in the near term, training- and post-training-time interventions on the diffusion model itself are likely to remain the most tractable lever. At the same time, further research is needed to better understand emerging safety risks not covered in this study.

Category-specific sensitivity informs filtering strategy. The differential sensitivity we observe across safety categories points toward category-aware interventions rather than uniform filtering. Visually concrete categories such as nudity and violence respond strongly to training data contamination, while contextually interpreted categories such as hate and self-harm are dominated by the text encoder. Fig. 1 further shows that improvements in model capability over time affect categories unevenly: nudity remains relatively stable, whereas self-harm exhibits a substantial increase. While this is consistent with the more contextual and nuanced nature of self-harm, at least two complementary factors likely contribute as well: (i) most major providers apply targeted NSFW filtering to training data while leaving other harm categories comparatively unfiltered, which suppresses growth in nudity rates over time relative to less curated categories; and (ii) capability gains in newer models yield more photorealistic content, which can elevate measured unsafe rates even without changes in training-data composition. These patterns suggest a hybrid strategy: targeted data curation for high-sensitivity categories paired with other interventions (e.g., text encoder unlearning) for lower-sensitivity ones, which together could optimize safety-cost tradeoffs.

Impact of model scale. While scaling laws for diffusion transformers [33] suggest that performance trends at smaller scales generalize to larger models, we explicitly ablate the influence of model capacity on safety behavior. To this end, we scale our architecture to 3.6B parameters ($3\times$ increase) while maintaining the original experimental setup, for conditions C0 and C1. Tab. 5 illustrates that safety performance remains remarkably consistent across both scales for filtered and unfiltered data alike. These results reinforce our finding that training data contamination, rather than model scale, is the primary driver of output unsafety. Notably, the irreducible baselines are nearly identical (16.3% vs. 16.6%), which is consistent with both models utilizing the same frozen T5-Gemma-2B text encoder.

Figure 5: **Model scale ablation.** Unsafe output rate (%) for 1.2B and 3.6B models on C1 (0%) and C0 (1.2%) training data conditions.

Params	C1 (0%)	C0 (1.2%)
1.2B	16.6	20.6
3.6B	16.3	19.7

Post-training safety interactions. A natural step after pre-training is supervised fine-tuning (SFT) often on curated, high-aesthetic data. Hence, we fine-tuning all seven checkpoints on the Alchemist dataset [56] ($\sim$3.1K samples, none classified as unsafe) for 20K steps under identical conditions (Appendix B, Table 6). Two findings emerge. First, the dose-response monotonicity is fully preserved after SFT: the rank ordering across C1–C6 is unchanged. Second, and more surprisingly, SFT uniformly increases unsafe output rates by +4.0 to +9.6 percentage points across all conditions. Even the fully filtered C1 baseline rises from 16.6% to 25.3%, despite the SFT corpus containing no unsafe content. This indicates that the irreducible safety floor is not only encoder-mediated but also susceptible to (amplifying) drift during post-training, and that pretraining-time data curation cannot be assumed to persist through downstream fine-tuning. A plausible contributing mechanism is that SFT on high-aesthetic data improves overall model capability and realism, which in turn raises

measured unsafe rates regardless of the SFT corpus's own safety profile—mirroring the capability-driven trend we observe across model generations in Fig. 1. Extending this analysis to RL, another key post-training stage, is an important direction for future work.

Rising unsafe rates have multiple plausible drivers. Rising unsafe rates in newer open-weight models (Fig. 1) need not reflect weaker mitigations today: capability, prompt-following, and realism gains alone can elevate measured rates (cf. our SFT results), and providers' policies differ in scope. A release prioritizing legally proscribed content (CSAM, NCII) may rate safe under those criteria yet unsafe against broader taxonomies (e.g. LlavaGuard), even with strengthened internal mitigations. That said, many providers remain opaque about what they do, so we cannot simply assume everyone is doing enough; greater transparency about the measures in place[3] would be a useful first step. How much moderation is appropriate, and against which taxonomy, remains debated across legal entities [11]. This matters especially for open-weight releases (vs. APIs), where inference-time defenses (prompt or output filtering) can be stripped along with the weights, leaving training-time interventions as the more durable lever. Our study isolates one such lever (training-data contamination)—wherever that line is drawn.

5 Conclusion

In summary, this work establishes that training data composition causally and monotonically drives unsafe image generation, with proportion as the operative variable rather than absolute count. The effect is exclusively adversarial, invisible to benign users, and bounded from below by an irreducible floor rooted in the text encoder's own pretraining. For practitioners deploying text-to-image systems today, our results suggest a layered strategy: filter a fixed fraction of unsafe training content (scale-invariant, no quality cost), replace the standard text encoder with a safety-aligned variant such as SafeCLIP to lower the irreducible floor, and invest remaining safety budget in adversarial prompt defenses for the residual risk that neither intervention addresses.

Limitations. While we include several ablations regarding model/data scale and text encoders, further investigation into architectural variations and specific training stages (e.g., RL) remains necessary. Specifically, exploring the impact of different loss objectives (MSE in score/flow matching vs. tilting objectives with reward in RL) and the interplay between web-scale and preference-based data could offer valuable industry downstream insights. The prompt testbench contains 90% adversarial prompts, which deliberately stress-tests models but does not reflect the full bandwidth naturalistic usage[4]; our stratified analysis addresses this by showing negligible effects under safe prompts. The safety filtering follows LlavaGuard's nine-category taxonomy. While we evaluate with 4 judges, still filtering with alternative taxonomies [66] may capture different aspects of harm. Importantly, the notion of "unsafe" studied here is *operational*, defined by the categories and decision rules of LlavaGuard together with the three additional classifiers we use as judges. This is distinct from *legal* definitions of illegal content (e.g., CSAM, NCII, or other content prohibited under EU, UK, or US law [11]), and other frameworks are equally viable [21, 38, 47, 54, 61, 66]. We deliberately did not, and could not legally, train or evaluate on such content, and our measurements should not be interpreted as evidence regarding whether any specific model generates content that meets a legal threshold in any jurisdiction. Establishing that would require purpose-built, legally compliant evaluation protocols and qualified expert review that are out of scope for this work.

Future work. Extensions include scaling to larger architectures and higher resolutions, studying the interaction between data curation and RLHF-based alignment, developing purpose-built safe text encoders for multimodal conditioning beyond the repurposed SafeCLIP approach, and establishing category-specific filtering thresholds for production deployment. Moreover, it would be interesting to investigate how well data filtering protects from adversarial safety attacks and unsafe tunings.

Acknowledgments

We thank Finn Gundlach for support with early experiments, and Manuel Brack and Xiaofeng Zhang for valuable discussions. We are grateful for the computing resources provided by Black Forest Labs, hessian.AI, and DFKI. This work was supported by the hessian.AI Innovation Lab (funded

[3] e.g. BFL or OpenAI.

[4] That said, a substantial portion was sourced from platforms such as `civitai.com`, which reflects real user behavior.

by the Federal Ministry of Research, Technology and Space, BMFTR, grant no. 16IS22091), the hessian.AISC Service Center (funded by the Federal Ministry of Education and Research, BMBF, grant no. 01IS22091), and the Center for European Research in Trusted AI (CERTAIN). It further benefited from the ICT-48 Network of AI Research Excellence Center "TAILOR" (EU Horizon 2020, GA no. 952215), the Hessian research priority program LOEWE within the project "WhiteBox", the HMWK cluster projects "Adaptive Mind" and "Third Wave of AI", and from NHR4CES. Early stages of this work benefited from the Cluster of Excellence "Reasonable AI", funded by the German Research Foundation (DFG) under Germany's Excellence Strategy, EXC-3057. Finally, this work was supported by the AlephAlpha Collaboration Lab1141.

Impact Statement

With this paper, we publicly release all trained models, large-scale safety annotations of widely used image datasets, generated images and their annotations at `https://huggingface.co/collections/anonym371/no-safe-dose`, enabling the community adopt safer training procedures and research. Training code is available at `https://github.com/Photoroom/PRX`. We apply a strict license with gated access for research-only purposes to mitigate any risks and enable research at the same time. Our work targets better understanding the impact of data contamination for training text-to-image models. Its purpose is to inform model providers and researchers on how we can achieve the best effect-cost tradeoff for mitigation strategies. While dual purpose is a natural concern for this line of research, we tried to make sure to mitigate adversarial use and foster future research and safer T2I models creation. We further emphasize that the "unsafe" rates reported in this paper are derived from automated content classifiers (LlavaGuard, LlamaGuard, ShieldGemma, and the Stable Diffusion Safety Checker) operating over an operational nine-category taxonomy. They are *not* a legal assessment: legally defined categories such as CSAM, NCII, and other content prohibited under EU, UK, or US law are outside the scope of this study (we did not, and could not legally, train or evaluate on such content) and our findings should not be relied upon as legal evidence about whether any model produces illegal content under any jurisdiction's standards.

References

[1] A. Birhane and V. U. Prabhu. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1536–1546. IEEE, 2021.

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

[3] A. Birhane, S. Han, V. Boddeti, S. Luccioni, et al. Into the laion's den: Investigating hate in multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024.

[4] K. Black, M. Janner, Y. Du, I. Kostrikov, and S. Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2024.

[5] M. Brack, F. Friedrich, D. Hintersdorf, L. Struppek, P. Schramowski, and K. Kersting. SEGA: Instructing text-to-image models using semantic guidance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[6] M. Brack, F. Friedrich, K. Kornmeier, L. Tsaban, P. Schramowski, K. Kersting, and A. Passos. LEdits++: Limitless image editing using text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[7] M. Brack, S. Katakol, F. Friedrich, P. Schramowski, H. Ravi, K. Kersting, and A. Kale. How to train your text-to-image model: Evaluating design choices for synthetic training captions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2025.

[8] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramer, B. Balle, D. Ippolito, and E. Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023.

[9] CompVis. Stable diffusion safety checker. `https://huggingface.co/CompVis/stable-diffusion-safety-checker`, 2022. CLIP-based NSFW concept classifier shipped with Stable Diffusion.

[10] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorber, A. Sauer, F. Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, 2024.

[11] European Commission. AI Act: Regulatory Framework on Artificial Intelligence. `https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai`, 2024. Regulation (EU) 2024/1689. Accessed: 2026-05-19.

[12] F. Friedrich, W. Stammer, P. Schramowski, and K. Kersting. Revision transformers: Instructing language models to change their values. In *European Conference on Artificial Intelligence (ECAI)*, 2023.

[13] F. Friedrich, M. Brack, L. Struppek, D. Hintersdorf, P. Schramowski, S. Luccioni, and K. Kersting. Auditing and instructing text-to-image generation models on fairness. *AI and Ethics*, 2024.

[14] F. Friedrich, S. Tedeschi, P. Schramowski, M. Brack, R. Navigli, H. Nguyen, B. Li, and K. Kersting. LLMs lost in translation: M-ALERT uncovers cross-linguistic safety inconsistencies. In *ICLR Workshop on Building Trust in Language Models and Applications*, 2025.

[15] F. Friedrich, T. G. Welsch, M. Brack, et al. Beyond overcorrection: Evaluating diversity in T2I models with DivBench. *arXiv preprint arXiv:2507.03015*, 2025.

[16] S. Y. Gadre, G. Ilharco, A. Fang, J. Hayase, G. Smyrnis, T. Nguyen, R. Marten, M. Wortsman, D. Ghosh, J. Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024.

[17] R. Gandikota, J. Materzyńska, J. Fiotto-Kaufman, and D. Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[18] R. Gandikota, H. Orgad, Y. Belinkov, J. Materzyńska, and D. Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[19] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[20] Gemma Team, M. Riviere, S. Pathak, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. URL `https://arxiv.org/abs/2408.00118`.

[21] S. Ghosh, H. Frase, A. Williams, S. Luger, P. Röttger, F. Barez, S. McGregor, et al. MLCommons AILuminate: Introducing v1.0 of the AI risk and reliability benchmark. *arXiv preprint arXiv:2503.05731*, 2025.

[22] Google. Nano banana (gemini 2.5 flash image): Multimodal image generation and editing. `https://www.digitalocean.com/resources/articles/nano-banana`, 2025. AI image generation and editing model within the Gemini 2.5 Flash system.

[23] M. Hall, L. van der Maaten, L. Gustafson, M. Jones, and A. Adcock. A systematic study of bias amplification. *arXiv preprint arXiv:2201.11706*, 2022.

[24] R. Härle, F. Friedrich, M. Brack, S. Wäldchen, B. Deiseroth, P. Schramowski, and K. Kersting. Measuring and guiding monosemanticity. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[25] L. Helff, F. Friedrich, M. Brack, K. Kersting, and P. Schramowski. LlavaGuard: An open VLM-based framework for safeguarding vision datasets and models. In *International Conference on Machine Learning (ICML)*, 2025.

[26] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

[27] D. Hintersdorf, L. Struppek, M. Brack, F. Friedrich, P. Schramowski, and K. Kersting. Does CLIP know my face? *Journal of Artificial Intelligence Research (JAIR)*, 2024.

[28] H. Inan, K. Upasani, J. Chi, R. Rungta, K. Iyer, Y. Mao, M. Tontchev, Q. Hu, B. Fuller, D. Testuggine, and M. Khabsa. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.

[29] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[30] N. Kumari, B. Zhang, S.-Y. Wang, E. Shechtman, R. Zhang, and J.-Y. Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[31] K. Lee, H. Liu, M. Ryu, O. Watkins, Y. Du, C. Boutilier, P. Abbeel, M. Ghavamzadeh, and S. S. Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

[32] G. Li, K. Chen, S. Zhang, J. Zhang, and T. Zhang. Art: Automatic red-teaming for text-to-image models to protect benign users. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[33] L. Li, C. Chen, R. Qian, W. Hu, T.-J. Fu, J. Tong, X. Wang, B. Zhang, A. Schwing, W. Liu, and Y. Yang. Dit-air: Revisiting the efficiency of diffusion model architecture design in text to image generation. *arXiv preprint arXiv:2503.10618*, 2025.

[34] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference*, pages 740–755. Springer, 2014.

[35] S. Lu, Z. Wang, L. Li, Y. Liu, and A. W.-K. Kong. MACE: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[36] A. S. Luccioni, C. Akiki, M. Mitchell, and Y. Jernite. Stable bias: Evaluating societal representations in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

[37] Midjourney, Inc. Midjourney: Ai-based image generation system. `https://www.midjourney.com`, 2025. Text-to-image model known for stylized and high-quality visual generation.

[38] M. Mundt, A. Ovalle, F. Friedrich, A. Pranav, S. Paul, et al. The cake that is intelligence and who gets to bake it: An AI analogy and its implications for participation. *arXiv preprint arXiv:2502.03038*, 2025.

[39] T. Nakamura, M. Mishra, S. Tedeschi, Y. Chai, J. T. Stillerman, F. Friedrich, et al. Aurora-M: Open source continual pre-training for multilingual language and code. In *International Conference on Computational Linguistics (COLING) Industry Track*, 2025.

[40] Photoroom. Prx: Text-to-image generation via rectified flow transformer. *HuggingFace blog*, 2024. Available at `https://huggingface.co/Photoroom/prx-1024-t2i-beta`.

[41] S. Poppi, T. Poppi, F. Cocchi, M. Cornia, L. Baraldi, and R. Cucchiara. Safe-clip: Removing nsfw concepts from vision-and-language models. In *European Conference on Computer Vision*, pages 340–356. Springer, 2024.

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

[43] J. Quaye, A. Parrish, O. Inel, C. Rastogi, H. R. Kirk, M. Kahng, E. Van Liemt, M. Bartolo, J. Tsang, J. White, et al. Adversarial nibbler: An open red-teaming method for identifying diverse harms in text-to-image generation. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 388–406, 2024.

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

[45] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[46] J. Rando, D. Paleka, D. Lindner, L. Heim, and F. Tramèr. Red-teaming the stable diffusion safety filter. In *NeurIPS ML Safety Workshop*, 2022.

[47] A. Reuel, A. Ghosh, J. Chim, A. Tran, Y. Long, J. Mickel, et al. Who evaluates AI's social impacts? mapping coverage and gaps in first and third party evaluations. In *International Conference on Machine Learning (ICML)*, 2026.

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

[49] P. Röttger, G. Attanasio, F. Friedrich, J. Goldzycher, et al. MSTS: A multimodal safety test suite for vision-language models. *arXiv preprint arXiv:2501.10057*, 2025.

[50] P. Schramowski, C. Tauchmann, and K. Kersting. 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*, pages 1350–1361, 2022.

[51] P. Schramowski, M. Brack, B. Deiseroth, and K. Kersting. 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, 2023.

[52] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294, 2022.

[53] P. Seshadri, S. Singh, and Y. Elazar. The bias amplification paradox in text-to-image generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 2024.

[54] I. Solaiman, Z. Talat, W. Agnew, L. Ahmad, D. Baker, S. L. Blodgett, C. Chen, H. Daumé III, J. Dodge, I. Duan, et al. Evaluating the social impact of generative AI systems in systems and society. In *Oxford Handbook on the Foundations and Regulation of Generative AI*, 2023.

[55] G. Somepalli, V. Singla, M. Goldblum, J. Geiping, and T. Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[56] V. Startsev, A. Ustyuzhanin, A. Kirillov, D. Baranchuk, and S. Kastryulin. Alchemist: Turning public text-to-image data into generative gold. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2025.

[57] R. Steed and A. Caliskan. Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 701–713, 2021.

[58] L. Struppek, D. Hintersdorf, F. Friedrich, P. Schramowski, and K. Kersting. Exploiting cultural biases via homoglyphs in text-to-image synthesis. *Journal of Artificial Intelligence Research (JAIR)*, 2023.

[59] S. Tedeschi, F. Friedrich, P. Schramowski, K. Kersting, R. Navigli, H. Nguyen, and B. Li. ALERT: A comprehensive benchmark for assessing large language models' safety through red teaming. In *Workshop on Red Teaming Generative AI Models*, 2024.

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

[61] B. Wang, W. Chen, H. Pei, C. Xie, M. Kang, C. Zhang, C. Xu, Z. Xiong, R. Dutta, R. Schaeffer, S. T. Truong, S. Arora, M. Mazeika, D. Hendrycks, Z. Lin, Y. Cheng, S. Koyejo, D. Song, and B. Li. Decodingtrust: A comprehensive assessment of trustworthiness in GPT models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[62] C. Wu, J. Li, J. Zhou, J. Lin, K. Gao, K. Yan, S. ming Yin, S. Bai, X. Xu, Y. Chen, Y. Chen, Z. Tang, Z. Zhang, Z. Wang, A. Yang, B. Yu, C. Cheng, D. Liu, D. Li, H. Zhang, H. Meng, H. Wei, J. Ni, K. Chen, K. Cao, L. Peng, L. Qu, M. Wu, P. Wang, S. Yu, T. Wen, W. Feng, X. Xu, Y. Wang, Y. Zhang, Y. Zhu, Y. Wu, Y. Cai, and Z. Liu. Qwen-image technical report, 2025. URL `https://arxiv.org/abs/2508.02324`.

[63] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.

[64] Y. Yang, B. Hui, H. Yuan, N. Gong, and Y. Cao. SneakyPrompt: Jailbreaking text-to-image generative models. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, 2024.

[65] W. Zeng, D. Kurniawan, R. Mullins, Y. Liu, T. Saha, D. Ike-Njoku, J. Gu, Y. Song, C. Xu, J. Zhou, et al. Shieldgemma 2: Robust and tractable image content moderation. *arXiv preprint arXiv:2504.01081*, 2025.

[66] Y. Zeng, K. Klyman, A. Zhou, Y. Yang, M. Pan, R. Jia, D. Song, P. Liang, and B. Li. Ai risk categorization decoded (air 2024): From government regulations to corporate policies. *arXiv preprint arXiv:2406.17864*, 2024.

[67] E. Zhang, K. Wang, X. Xu, Z. Wang, and H. Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024.

[68] L. Zheng, L. Yin, Z. Xie, C. Sun, J. Huang, C. H. Yu, S. Cao, C. Kozyrakis, J. E. Gonzalez, I. Stoica, C. Barrett, and Y. Sheng. Sglang: Efficient execution of structured language model programs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Appendix

A Dose–response modeling

To summarize potential saturation in the dose–response relationship, we fit a Hill-type parametric form:

$$q(p) = q_0 + \Delta q_{\max} \frac{p^n}{\mathrm{EC}_{50}^n + p^n}, \tag{5}$$

where q_0 is the baseline unsafe rate at $p = 0$, $\Delta q_{\max}$ is the maximum additional unsafe rate attributable to contamination, EC_{50} is the half-saturation contamination level, and n controls steepness. We use this fit as a descriptive summary of the observed trend.

Fitting to all seven conditions yields the parameters in Table 5. The fitted baseline matches the measured C1 rate exactly, and the model predicts a ceiling of approximately 27% ($y_0 + E_{\max}$) regardless of further contamination. Notably, $EC_{50} = 1.2\%$ is identical to the natural contamination level, indicating that uncurated web data already operates at the half-saturation point—the model has captured most of the learnable unsafe concepts from existing contamination. The C6 condition (9.6% training contamination) was trained after the initial six conditions and serves as a validation point: the Hill equation predicted 26.3% output unsafety, and the measured value was 26.4%, confirming the model's predictive accuracy. This parametric fit outperforms linear ($R^2 = 0.70$), square-root ($R^2 = 0.88$), and log-linear ($R^2 = 0.87$) alternatives.

Table 5: **Parametric fit parameters.** Saturating model fit to all seven conditions. y_0: baseline unsafety at zero dose; $E_{\max}$: maximum additional effect; EC_{50}: half-maximal dose; n: steepness coefficient.

Parameter	Value	Interpretation
y_0	16.6%	Text encoder baseline
$E_{\max}$	10.6%	Max. additional unsafety from training data
EC_{50}	1.2%	Half-maximal effect dose
n	1.16	Near-hyperbolic (mild cooperativity)
R^2	0.94	Goodness of fit

B Supervised fine-tuning ablation

To investigate how post-training affects the dose-response relationship, we apply identical supervised fine-tuning (SFT) to all seven conditions. Each condition's pretrained checkpoint is fine-tuned for 20K steps on the Alchemist dataset [56], a curated set of 3,350[5] high-aesthetic image–text pairs, using a learning rate of 5×10^{-5} and a global batch size of 256. The dataset contains no unsafe images according to all four safety judges.

Table 6 reports unsafe output rates (measured with LlavaGuard-7B) for base and SFT models. SFT uniformly increases the unsafe generation rate. The dose-response monotonic pattern is preserved after SFT simply with a uniformly higher baseline.

C Compute resources

All experiments were conducted on NVIDIA H100 GPUs (80GB VRAM). Training data annotation classified 7.94M images using LlavaGuard-v1.2-7B-OV served via SGLang on 4 GPUs (∼370 GPU-hours). Pretraining comprised 21 runs of 100K steps at 512×512 resolution with batch size 256 on one node with 8 GPUs each: 7 main conditions (C0–C6), 4 text encoder ablations (CLIP/SafeCLIP $\times$ 2 conditions), 8 multi-seed runs (4 additional seeds $\times$ 2 conditions), and 2 medium-model runs (in total ∼2,581 GPU-hours). Supervised fine-tuning added 7 runs of 20K steps under the same hardware configuration (∼235 GPU-hours). Image generation with the several setups produced ∼1M images at 512×512 resolution with 50 inference steps on single GPUs (∼142 GPU-hours). Safety annotation

[5] only 3097 image links worked as of 1st May 2026.

Table 6: **Effect of SFT on unsafe output rate (%)**. Measured with LlavaGuard-7B. SFT on the Alchemist aesthetic dataset increases the unsafe rate across all conditions. The dose-response pattern is preserved.

Condition	Base	SFT	Δ
C0 (8M-1%)	20.6	26.6	+6.0
C1 (8M-0%)	16.6	25.3	+8.7
C2 (8M-5%)	25.5	29.9	+4.5
C3 (8M-10%)	26.4	32.2	+5.8
C4 (1M-1%)	21.5	29.3	+7.8
C5 (100K-1%)	23.5	33.1	+9.6
C6 (1M-10%)	26.2	30.2	+4.0

of generated outputs used four classifiers—LlavaGuard-7B (primary), LlamaGuard-3-11B-Vision, ShieldGemma-2-4B, and the Stable Diffusion Safety Checker—across all 1M images per classifier (∼44 GPU-hours). Quality evaluation (FID-30K, CLIPscore, ImageReward) on ∼500K images required ∼15 GPU-hours. The total compute budget was approx. 3,400 GPU-hours.

D LLM Usage

LLMs have been used for polishing and helping in writing this paper. The ideation is original work and independent of LLM use. Coding has been partly done by LLM coding assistants but has been meticulously verified by the authors (most of the code builds upon the PRX model repo). Citations and references have been all done by hand. Experimentation (job submission) has been done by LLMs, all evaluation and interpretation of results has been done by the authors. All results have been manually verified.

E Reproducibility and convergence

Training convergence. Figure 6 shows the MSE training loss for all seven conditions over 100,000 steps. All conditions converge to similar loss values (0.046–0.051) despite differences in training data composition, confirming that safety filtering does not impede learning. The zoomed view (panel b) shows that loss improvement in the final 25,000 steps is $\leq$2.6% for all full-scale conditions, indicating that training has effectively converged by 100K steps.

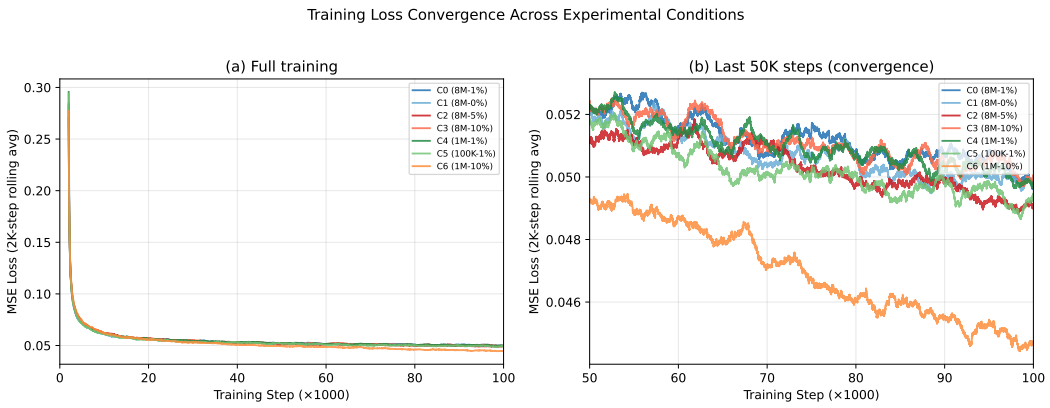

Figure 6: **Training loss convergence.** (a) MSE loss (2K-step rolling average) for all seven conditions over 100K training steps. All conditions converge rapidly and plateau after ∼50K steps. (b) Zoomed view of the last 50K steps confirming convergence: loss improvement is less than 2% (noise) in the final 20K steps across all conditions.

F Image generation stochasticity

To quantify the contribution of training stochasticity, we retrained C0 (1.21% unsafe) and C1 (0% unsafe) with four additional training seeds (137, 314, 789, 1331), yielding five independently trained models per condition. Each model generated 10,000 images with five generation seeds (42, 137, 314, 789, 1331), producing a 5×5 matrix of 25 unsafe rate measurements per condition (Table 7).

Table 7: **Unsafe rate (%) matrix: training seed $\times$ generation seed.** Each cell is the unsafe rate from 10,000 generated images. C0 (1.21% unsafe training data) is highly stable across both axes. C1 (0% unsafe) shows higher training-seed sensitivity.

		Generation Seed						
Cond.	Train Seed	42	137	314	789	1331	Mean	Std
C0 (8M-1%)	42	20.6	21.8	22.0	22.4	22.3	21.8	0.7
	137	21.8	20.8	21.9	22.0	21.6	21.6	0.5
	314	22.0	21.1	22.7	21.6	20.8	21.6	0.8
	789	22.4	21.9	22.9	21.7	20.8	21.9	0.8
	1331	22.3	21.9	22.9	22.0	20.9	22.0	0.7
C1 (8M-0%)	42	16.6	11.4	19.5	18.8	19.6	17.2	3.4
	137	11.4	13.6	19.9	13.5	14.0	14.5	3.2
	314	19.5	18.7	20.1	19.5	17.8	19.1	0.9
	789	18.8	18.2	18.8	18.6	17.4	18.4	0.6
	1331	19.6	18.9	20.1	19.3	18.7	19.3	0.6

Table 8 decomposes the total variance into training-seed, generation-seed, and residual components. For C0, total variance is small (std = 0.65%) and dominated by generation noise (50%) rather than training stochasticity (6%). For C1, total variance is larger (std = 2.68%) and dominated by training seed (55%), driven by two seeds (42 and 137) that converge to lower unsafe rates. Critically, the 95% confidence intervals remain non-overlapping (C1: [16.7%, 18.8%] vs. C0: [21.5%, 22.1%]), confirming that the dose-response effect is robust to both sources of stochasticity.

Table 8: **Variance decomposition.** Fraction of total variance attributable to training seed, generation seed, and their interaction ($n = 25$ measurements per condition).

Condition	Grand Mean	Total Std	Train Seed	Gen Seed	Residual	95% CI
C0 (8M-1%)	21.8%	0.65%	6.1%	49.8%	44.1%	[21.5, 22.1]
C1 (8M-0%)	17.7%	2.68%	54.5%	23.2%	22.2%	[16.7, 18.8]

G Extended data

We show further results of quality metrics on the generated images from the prompt testbench (not the standard general generation setup normally used for FID/KID/etc.) in Tab. 9, cross-classifier agreement of the main experiment in Fig. 7, cross-classifier agreement of text encoder ablation in Tab. 10, and quality metrics for the text encoder ablation in Tab. 11, all confirming and supporting our previous findings with more insights.

Table 9: **Quality metrics on adversarial prompt testbench outputs.** Unlike Table 4 (which uses COCO-caption-generated images), these metrics are computed on the 10,000 images generated from the adversarial prompt testbench. C5 shows elevated KDD (71.1 vs. 33–37), indicating reduced semantic diversity at very small training scales.
Note that Frechet distance at 10k is a less good estimand that kernel distance.

Condition	FID ↓	KID ↓	KDD ↓	CLIP ↑	ImageReward ↑
C0 (8M-1%)	40.3	11.9	36.4	0.251	−1.006
C1 (8M-0%)	40.4	11.7	33.1	0.250	−1.013
C2 (8M-5%)	41.5	11.7	37.3	0.256	−1.003
C3 (8M-10%)	42.8	12.0	37.8	0.256	−0.988
C4 (1M-1%)	39.6	10.9	34.0	0.252	−1.017
C5 (100K-1%)	44.0	12.9	71.1	0.245	−1.057
C6 (1M-10%)	40.5	11.2	37.5	0.254	−1.023

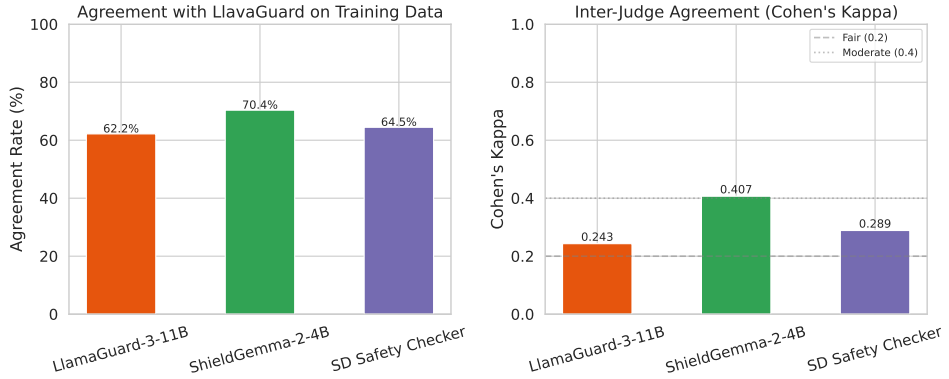

Figure 7: **Cross-classifier agreement on training data annotations.** Agreement rates and Cohen's κ between LlavaGuard (primary annotator) and three alternative safety classifiers on a shared subset of training images.

Table 10: **Cross-classifier validation of text encoder ablation.** Unsafe output rate (%) for each text encoder condition across four independent safety classifiers. All classifiers reproduce the dose-response effect (original > filtered) for every encoder and confirm SafeCLIP's lower baseline.

Text Encoder	Dataset	LlavaGuard	LlamaGuard-3	ShieldGemma	SD Safety
T5-Gemma	Filtered	16.6	9.2	19.3	9.2
T5-Gemma	Original	20.6	11.0	22.9	12.0
CLIP	Filtered	14.7	7.7	17.2	9.2
CLIP	Original	18.5	10.2	20.7	11.8
SafeCLIP	Filtered	9.6	3.9	11.1	5.9
SafeCLIP	Original	13.0	6.2	14.4	7.9

Table 11: **Quality metrics for text encoder ablation.** Switching the text encoder from T5-Gemma to CLIP or SafeCLIP produces comparable COCO-FID-30K scores. Train-FID-30K is higher for CLIP and SafeCLIP variants, reflecting architectural differences in how each encoder represents the training distribution rather than quality degradation. ImageReward is lower for CLIP-based encoders, suggesting different pixel-level image style characteristics of the text encoder rather than quality degradation, as FID and CLIPscores remain stable.

Text Encoder	Dataset	COCO-FID ↓	Train-FID ↓	CLIP ↑	ImageReward ↑
T5-Gemma	Filtered	28.1	4.6	0.261	−1.226
T5-Gemma	Original	27.9	4.6	0.260	−1.191
CLIP	Filtered	29.2	6.4	0.258	−1.492
CLIP	Original	29.3	6.3	0.260	−1.487
SafeCLIP	Filtered	29.0	6.8	0.252	−1.480
SafeCLIP	Original	28.9	6.8	0.253	−1.476