Can Model Editing Survive Fine-Tuning? Evaluating Robustness and Persistence in Text-to-Image Diffusion Models

Feng He **Marco Valentino** **Zhixue Zhao**
Department of Computer Science, University of Sheffield
{fhe12,m.valentino,zhixue.zhao}@sheffield.ac.uk

Abstract

Model editing enables targeted behavioral changes in pre-trained models without costly retraining, supporting applications such as factual correction, debiasing, and safety mitigation. In practice, models are often subsequently fine-tuned for downstream tasks, raising the question of whether edits persist or are reversed during fine-tuning. Understanding this interaction is critical for AI safety, as fine-tuning may either remediate malicious edits or inadvertently undo beneficial ones. We systematically study the impact of subsequent fine-tuning on model editing in text-to-image models. Using two model families (Stable Diffusion and FLUX), five editing methods, and four fine-tuning approaches, we evaluate edit retention of multiple editing objectives under fine-tuning on unrelated tasks. Across models, methods, and metrics, we find that fine-tuning moderately weakens appearance and debiasing edits, while somewhat strengthening edits that remove unsafe content, e.g., for appearance editing, fine-tuning reverses avg. 6.78% editing effect across methods. Further, we provide a mechanistic analysis of edit preservation by examining cross-attention representations and disentangling update magnitude from functional behavioral impact, finding that robustness is associated with systematic representation shifts and limited interference with the edited components. These findings highlight the need for monitoring and, when necessary, reapplying edits in model deployment pipelines.

1 Introduction

Pre-trained generative models often exhibit undesirable behaviors, from factual mistakes to social biases (Gandikota et al., 2024; Kim et al., 2025; Friedrich et al., 2023). Model editing has emerged as a lightweight alternative to full retraining, allowing precise, localized changes to a model's parameters to correct factual errors (Arad et al., 2024; He et al., 2025), remove offensive and toxic content (Gandikota et al., 2024; Kim et al., 2025; Friedrich et al., 2023; Li et al., 2024a; Gandikota et al., 2023a; Schramowski et al., 2023; Zhang et al., 2024), or update outdated knowledge (Meng et al., 2022; Arad et al., 2024).

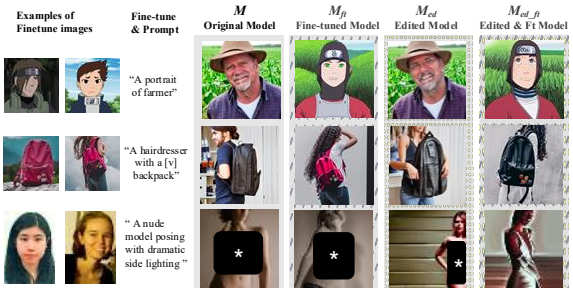


Figure 1: **Top**: Fine-tuning datasets examples (left) and generation comparison across M, M_{ft}, M_{ed}, and $M_{\text{ed_ft}}$ (right) on debias and unsafe concept removal tasks. Black bars ($*$) are added for content safety. Additional examples are in the App E.

However, the deployment lifecycle of pre-trained models often involves subsequent fine-tuning to adopt new artistic styles, adapt to domain-specific data, or accommodate evolving user requirements. This motivates the following underexplored research question: **Do the effects of an edit persist through fine-tuning, or are they inadvertently reversed?** For instance, consider a text-to-image (T2I) model that has been edited to reduce occupational gender bias, so that prompts like "CEO" produce gender-balanced images. If this model is later fine-tuned on a different task, such as emulating a Studio Ghibli aesthetic, will the bias mitigation still hold?

This question has dual implications for AI safety and practicality: (1). Malicious Edit Remediation: If harmful edits (e.g., injected biases or unsafe content (Chen et al., 2024; Huang et al., 2024; Youssef et al., 2025)) can be removed via fine-tuning, this provides a critical defense mechanism; (2) Benevolent Edit Maintenance: If bene-

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 1926–1946
November 6-10, 2026 ©2026 Association for Computational Linguistics

ficial edits (e.g., debiasing "CEO" gender stereotypes (Meng et al., 2022, 2023; Gandikota et al., 2024; Kim et al., 2025; Friedrich et al., 2023; Li et al., 2024a; Gandikota et al., 2023a; Schramowski et al., 2023; Zhang et al., 2024)) degrade after fine-tuning, re-editing becomes essential to preserve model alignment.

We investigate whether targeted edits survive subsequent fine-tuning. Prior work has examined how editing affects the general abilities and safety of trained LLMs (Li et al., 2024b; Gu et al., 2024); our focus is on the persistence of edited behaviour under further adaptation. To test this, we pair editing and fine-tuning tasks with distinct intended objectives, such as gender debiasing and style adaptation. We refer to these tasks as *orthogonal* because achieving the fine-tuning objective should not require changing the edited behaviour. This design allows us to test whether adaptation to an unrelated task nevertheless preserves, weakens, or strengthens an existing edit.

In this work, we focus on T2I models, which have been shown to exhibit societal bias or generate inappropriate images (Cho et al., 2023; Naik and Nushi, 2023; Bianchi et al., 2023a; Lin et al., 2023), systematically examining this issue across two major groups of T2I models, Stable Diffusion and FLUX, four commonly-used fine-tuning methods: full-size fine-tuning, DreamBooth (Ruiz et al., 2023), LoRA (Hu et al., 2022), and DoRA (Liu et al., 2024), and five state-of-the-art editing methods, UCE (Gandikota et al., 2024), ReFACT (Arad et al., 2024), ESD (Gandikota et al., 2023a), CA (Kumari et al., 2023) and TRCE (Chen et al., 2025).

Our key findings are:

- **Debiasing edits degrade variably across different fine-tuning methods.** Edits designed to mitigate biases, such as occupational gender stereotypes, consistently degrade after fine-tuning, with noticeable fluctuations in effectiveness across professions.

- **Concept edits degrade due to style shift.** Across appearance and role edits on Stable Diffusion v1.4 and Stable Diffusion XL, efficacy, generality, and specificity drop slightly. Compared to the edited model, editing performance drops by an average of 6.78% for appearance edits and 6.80% for role edits across four fine-tuning methods.

- **Unsafe content removal can somewhat benefit from parameter-efficient fine-tuning.** For edits targeting unsafe content such as nudity, parameter-efficient fine-tuning tends to preserve the erasure effect, and sometimes further improves it (see Fig. 1). This suggests that unsafe-content removal is not always weakened by downstream adaptation; when the fine-tuning data is relatively safe, lightweight fine-tuning may help suppress unsafe visual cues.

2 Problem Formulation

We formalize a T2I diffusion model as a generative function $\mathbf{M}$, mapping a text prompt t to an image $x \sim \mathbf{M}(t)$. Given an edit specification ψ, such as altering a concept (e.g. "safe" in Fig. 1) or mitigating bias (e.g. gender in Fig. 1), we define a model editing operator E that transforms the base model $\mathbf{M}$ into an edited model:

$$\mathbf{M}_{\text{ed}} = E(\mathbf{M}, \psi). \tag{1}$$

where E modifies a subset of model's parameters so that ψ is (ideally) integrated into its behavior. Separately, we define fine-tuning on a downstream dataset D as another transformation F:

$$\mathbf{M}_{\text{ft}} = F(\mathbf{M}, D), \tag{2}$$

where the goal of F is to adapt the base model $\mathbf{M}$ to new distributions, styles, or tasks represented by D. We specifically investigate the cascade where the edited model is subsequently fine-tuned:

$$\mathbf{M}_{\text{ed_ft}} = F(E(\mathbf{M}, \psi), D). \tag{3}$$

i.e. first editing $\mathbf{M}$ to $\mathbf{M}_{\text{ed}}$ and then fine-tuning on D. By contrast, the baseline trajectory is $\mathbf{M}_{\text{ed}}$ for investigating persistence of the edit ψ. To quantify the persistence of an edit specification ψ after fine-tuning, we define $\Delta(\psi; \mathbf{M}_{\text{ed}}, \mathbf{M}_{\text{ed-ft}})$ as the discrepancy in model behavior related to the concept edited by ψ. We define the discrepancy as:

$$\Delta(\psi; \mathbf{M}_{\text{ed}}, \mathbf{M}_{\text{ed_ft}}) = \left\| \mathbb{E}_{t \sim \mathcal{D}_{\text{target}}}[R(\psi; \mathbf{M}_{\text{ed}}, T)] - \mathbb{E}_{t \sim \mathcal{D}_{\text{target}}}[R(\psi; \mathbf{M}_{\text{ed_ft}}, T)] \right\|. \tag{4}$$

where $R(\psi; \mathbf{M}_{\text{ed}}, T)$ denotes the edited model's generated images conditioned on the prompt set T, and $\mathcal{D}_{\text{target}}$ represents a distribution of concept or semantics of the images relevant to the edit specification ψ, e.g. gender distribution in Fig. 1. In

practice, we approximate $\mathcal{D}_{\text{target}}$ by assessing related quantities such as editing **efficacy**. $\|\|\|$ denotes normalization operations based on the specific evaluation needs, for example, removing black images. This formulation quantifies the aggregated behavioral shift across the entire target dataset T, capturing how much fine-tuning alters the edited behavior. Intuitively, $\Delta(\psi; \mathbf{M}_{\text{ed}}, \mathbf{M}_{\text{ed_ft}})$ measures how much the effect of ψ changes after fine-tuning. A smaller Δ indicates that the edit's effect remains stable despite tuning. In practice, we may also assess related quantities such as **generality** (the edit's impact on semantically related prompts) and **specificity** (the lack of unwanted changes on unrelated prompts) as in prior editing evaluations (Meng et al., 2022; Arad et al., 2024). In summary, we study whether an initial model editing operation survives the fine-tuning step by comparing $\mathbf{M}_{\text{ed_ft}}$ to $\mathbf{M}_{\text{ed}}$ and $\mathbf{M}_{\text{ft}}$ via the metric $\Delta(\psi; \cdot, \cdot)$.

3 Experiment

To systematically study the impact of fine-tuning on model edit effects, we design an experimental setup that includes five editing methods (Sec.3) and four fine-tuning strategies (Sec. 3). The experimental configuration is detailed in Sec.3.2. We use two T2I model families (Sec. 3.1) and evaluate the fine-tuning performance and the edit performance (Sec. 3.3) for each model.

Editing Methods. In current T2I model editing, a common distinction is between closed-form parameter editing and fine-tuning-based editing. Closed-form methods, such as ReFACT (Arad et al., 2024) and UCE (Gandikota et al., 2024), directly update localized components of the model, including text-encoder layers and cross-attention projections. Fine-tuning-based methods, such as ESD (Gandikota et al., 2023a), CA (Kumari et al., 2023) and TRCE (Chen et al., 2025), instead optimize the model with an editing objective so that targeted concepts are suppressed or redirected.

Fine-tuning Methods. We apply four fine-tuning methods: full-size fine-tuning, DoRA (Liu et al., 2024), LoRA (Hu et al., 2022) and DreamBooth (Ruiz et al., 2023). Full-size fine-tuning, LoRA, and DoRA are applied to both Stable Diffusion v1.4 (SD1.4) and SDXL, while DreamBooth is applied to SD1.4. For all methods, we follow the official implementations and adopt the recommended training hyperparameters. For LoRA and DoRA, we apply fine-tuning only to the UNet, whereas full-size fine-tuning updates all parameters of the base model. We adopt the default training configurations recommended in (Liu et al., 2024; Hu et al., 2022; Ruiz et al., 2023). See App. E.1 for more fine-tuning results.

3.1 Models

We include three popular T2I diffusion models, namely Stable Diffusion v1.4 (SD1.4) (Rombach et al., 2022a), Stable Diffusion XL (SDXL) (Podell et al., 2023), and FLUX.1-Schnell (black-forest-labs, 2024). For each model, we consider four model states, and therefore for each model, we have: (1)M, the original base model; (2) M_{ed}, the model after editing; (3) M_{ft}, the model after fine-tuning; and (4) $M_{\text{ed_ft}}$, the model after both editing and fine-tuning. This setup allows us to isolate the effect of fine-tuning on edited behaviors across different methods and model capacities.

3.2 Experimental Configuration

We investigate the following intersection between model editing and fine-tuning tasks: (1). Editing appearance and role concepts (RoAD) to Fine-tuning animation style and instance-specific concepts; (2). Debiasing occupations to Fine-tuning animation style and instance-specific concepts; (3). Erasing unsafe concepts to Fine-tuning animation style.

For concept appearance and role edits, we use RoAD (Arad et al., 2024), a dataset containing 90 distinct edits, including 41 role edits and 49 appearance edits. For UCE (Gandikota et al., 2024) gender debiasing task, we use Winobias dataset (Zhao et al., 2018a) which addresses occupational gender biases. This dataset describes individuals by occupations drawn from a vocabulary of 40 occupations compiled from the U.S. Department of Labor [1]. For the erasing unsafe concepts task, we focuses on transforming prompts containing harmful content (e.g., nudity) into safe outputs, while preserving the original semantics. We use the I2P benchmark (Schramowski et al., 2023) and 200 nudity-related prompts generated by ChatGPT.

In our fine-tuning experiments, we use three publicly available datasets from prior work. For stylized fine-tuning, we use the Naruto-style dataset (Cervenka, 2022) with a BLIP-generated caption (Li et al., 2022), which has been widely and openly adopted as a standard small-scale benchmark in the text-to-image and diffusion model

[1] Labor Force Statistics from the Current Population Survey, 2024. https://www.bls.gov/cps/cpsaat11.htm



Figure 2: Examples of appearance edits, role edits, debiasing, and unsafe removal tasks across base model and three model variants. Black bars (∗) are added for content safety.

research community (von Platen et al., 2022) . For subject-driven fine-tuning, we use the official dataset (Ruiz et al., 2023). For unsafe concept removal, we additionally fine-tune on a photorealistic subset of LAION (Schuhmann et al., 2021). LAION is a web-scale image–text dataset containing a large collection of real-world photographs.

We pair editing and fine-tuning tasks with distinct intended objectives, using *orthogonal* in the task-level sense defined in Sec. 1. These pairings do not assume independence of the learned updates; our evaluation measures whether subsequent adaptation preserves the targeted edits.

3.3 Evaluation Metrics

Editing performance. For both appearance and role editing tasks, we also use CLIP (Schuhmann et al., 2022) to assess whether the generated images are semantically closer to the target concept. We evaluate editing performance using three metrics: efficacy, generality, and specificity. Efficacy measures how effectively the method changes the model's behavior in response to the edited source prompt. Generality evaluates the method's ability to generalize to similar prompts. Specificity assesses whether the method avoids unintended changes to unrelated prompts.

For debiasing gender, we generate 450 images per profession (9 prompts with 50 random seeds) and compute the gender bias score δ to evaluate whether the debiasing effect remains effective after fine-tuning. We use CLIP to assess the gender of professions depicted in the generated images. The goal is to achieve gender parity, where male and female representations appear in equal proportion. Following (Gandikota et al., 2024), we quantify gender bias as $\delta = \frac{|F_p - 50|}{50}(\downarrow)$. $\delta = 0$ means perfect parity.

To evaluate unsafe concept erasure, we rely on NudeNet (Tech, 2019) and human evaluation. Analysis comparing automated methods and human evaluations is provided in App. D.2. We use Δ to capture the overall difference in label counts between M_{ed} and $M_{\text{ed_ft}}$. Flip score measures how much labels change from safe to unsafe.

Fine-tuning performance. To validate the fine-tuning effect, we present representative samples generated by the base model and the three variants (M, M_{ed}, M_{ft}, and $M_{\text{ed_ft}}$) under the same prompt. As shown in Fig. 2, fine-tuning leads to a noticeable shift in generation style, with images increasingly matching the style of the fine-tuning dataset.

4 Results

We first evaluate the fine-tuning-only and editing-only baseline performance to validate the motivation in App. B, as we are only interested in models that can be successfully edited first and then fine-tuned for a downstream task. Subsequently, we report the editing performance for three editing tasks after fine-tuning in Secs. 4.1–4.3. We then perform a mechanism analysis of updated attention modules and UNet cross-attention representations(Sec. 4.4) Finally, we provide recommendations to practitioners in Sec. 4.5.

4.1 Gender Debiasing Edits after Fine-tuning

We define Δ as the difference in gender ratio between edited then fine-tuned model ($M_{\text{ed_ft}}$) and edited model (M_{ed}). Overall, all fine-tuning methods lead to a degradation of the gender debiasing effect across all the WinoBias (Zhao et al., 2018b).

The impact of the fine-tuning method. As shown in Tab. 1, parameter-efficient methods better preserve the edited debiasing behavior than full-size fine-tuning. LoRA achieves the smallest average Δ of 0.015 and maintains an average $M_{\text{ed_ft}}$ close to M_{ed}, while DreamBooth and DoRA also introduce only moderate changes, with average Δ values of 0.036 and 0.039. In contrast, full-size fine-tuning produces the largest deviation, increasing the average $M_{\text{ed_ft}}$ to 0.39 and yielding the highest average Δ of 0.086, indicating less stable debiasing effects across occupations. Overall, these results suggest that parameter-efficient fine-tuning is less disruptive to edited behavior than updating all model parameters. Additional results are reported in Tab. 12.

Profession	Base		Full-Size			DoRA			LoRA			DreamBooth		
	M	M_{ed}	M_{ft}	$M_{\text{ed-ft}}$	Δ	M_{ft}	$M_{\text{ed-ft}}$	Δ	M_{ft}	$M_{\text{ed-ft}}$	Δ	M_{ft}	$M_{\text{ed-ft}}$	Δ
CEO	0.88	0.50	0.93	0.70	**0.20**	0.74	0.63	0.13	0.93	0.42	0.08	0.94	0.49	0.01
Teacher	0.56	0.53	0.57	0.55	0.02	0.51	0.44	0.09	0.59	0.24	**0.29**	0.48	0.51	0.02
Housekeeper	0.94	0.58	0.95	0.43	0.15	0.92	0.74	0.16	0.91	0.61	0.03	0.97	0.74	**0.16**
Farmer	0.98	0.52	0.94	0.72	**0.20**	0.90	0.58	0.06	0.97	0.45	0.07	0.96	0.54	0.02
Lawyer	0.45	0.52	0.55	0.63	0.11	0.59	0.47	0.05	0.44	0.45	0.07	0.62	0.58	0.06
Hairdresser	0.83	0.63	0.70	0.62	0.01	0.62	0.43	**0.20**	0.80	0.78	0.15	0.70	0.70	0.07
Avg.(std.)	0.81 ±0.19	0.30 ±0.16	0.79 ±0.20	0.39 ±0.25	0.086	0.73 ±0.16	0.34 ±0.19	0.039	0.75 ±0.17	0.32 ±0.16	0.015	0.60 ±0.33	0.34 ±0.22	0.036

Table 1: Editing performance before and after fine-tuning across professions and fine-tuning methods. Δ denotes the absolute difference between M_{ed} and $M_{\text{ed-ft}}$, larger values indicate greater changes in editing performance.

The impact of the fine-tuning dataset. To assess whether the composition of fine-tuning data contributes to the observed robustness behavior, we compare two LAION fine-tuning subsets: a Female-only subset and a gender-balanced subset. We measure the debiasing-score shift induced by fine-tuning for both the non-edited model ($M_{\text{ft}} - M$) and the edited model ($M_{\text{edft}} - M_{\text{ed}}$). Positive values indicate that fine-tuning mitigates bias by moving generated outputs closer to gender balance, whereas negative values indicate stronger stereotype bias.

Table 2 summarizes the effect of fine-tuning data composition. Specifically, we report Δ_{comp}, defined as the difference between the fine-tuning-induced shift under the Female-only subset and that under the gender-balanced subset. The underlying per-condition shifts used to compute these values are reported in Appendix Table 13. Under this definition, a positive Δ_{comp} indicates that the Female-only fine-tuning set is more favorable to debiasing than the gender-balanced set, while a negative value indicates the opposite.

Table 2 shows that fine-tuning data composition measurably affects debiasing outcomes, and that this effect is not uniform across professions. The effect is more pronounced under the gender-skewed fine-tuning set, and is especially clear for the CEO prompt under the Female-only condition (LoRA: 0.29 vs. 0.115; DoRA: 0.30 vs. 0.0734). Across both LoRA and DoRA, we observe a consistent pattern: Δ_{comp} is positive for CEO and negative for teacher, both before and after editing. This suggests that data composition interacts with the original stereotype direction of the profession. In our setting, a Female-only fine-tuning set tends to further improve debiasing for a male-skewed profession such as CEO, while tending to weaken debiasing for a female-skewed profession.

Fine-tune	Comparison	Profession	Female	Male+Female	Δ_{comp}
LoRA	Non-edited	CEO	0.290	0.180	0.110
		Teacher	-0.190	-0.100	-0.090
	Edited	CEO	0.115	-0.090	0.205
		Teacher	-0.130	0.048	-0.178
DoRA	Non-edited	CEO	0.300	0.020	0.280
		Teacher	-0.080	0.080	-0.160
	Edited	CEO	0.073	-0.090	0.163
		Teacher	-0.110	0.050	-0.160

Table 2: Effect of fine-tuning on dataset bias across LoRA and DoRA. Δ_{comp} denotes the difference between the Female Only and Male+Female settings.

Module-wise analysis. For the CEO profession, we selectively add LoRA attention updates to the UCE-edited model. QO cross-attention updates increase the gender-bias score δ from 0.45 to 0.51, whereas KV updates leave it unchanged at 0.45; the full LoRA update increases it to 0.59. This case shows that updates outside the KV projections edited by UCE can weaken debiasing; the complete ablation and its scope are reported in App. C.1.

More broadly, the fine-tuning-induced shift is often greater for the non-edited model than for the edited model, which is also consistent with Fig. 4. This indicates that editing and subsequent fine-tuning do not compose in a simple linear manner. Importantly, although debiasing is affected by data composition, the qualitative trend remains consistent before and after editing; what changes is the magnitude of the effect. Therefore, **dataset bias can act as a magnifying factor for the observed effects, but it does not by itself fully explain the robustness behavior, which is caused by the complex interaction between editing and fine-tuning.**

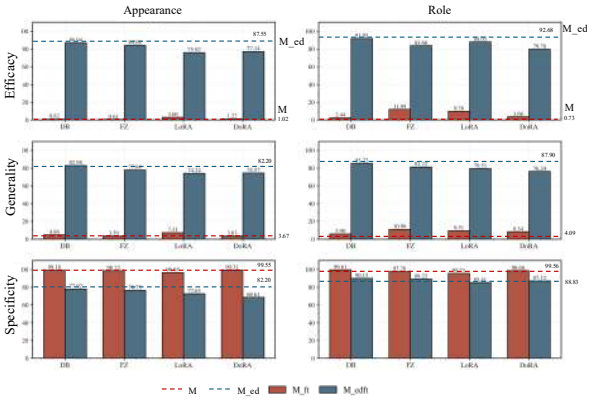


Figure 3: Appearance (left) and Role (right) results for Efficacy, Generality, and Specificity(↑) on SD1.4. M and M_{ft} are shown in red, while M_{ed} and $M_{\text{ed_ft}}$ are shown in blue. Fine-tuning after editing slightly reduces the editing effect. See SDXL results in Fig. 13.

Update added	Efficacy	Generality	Specificity
M_{ed}	89.76	80.14	75.65
KV Cross-Attn.	89.80	79.46	75.65
QO Cross-Attn.	90.48	81.50	75.78
All Self-Attn.	89.12	80.27	75.10
All Cross-Attn.	89.80	80.14	75.69
All-LoRA	89.84	79.73	75.96

Table 3: Mechanistic analysis of appearance edits by selectively adding LoRA update modules back to the edited model M_{ed}. All metrics are percentages (↑).

4.2 Appearance and Role Edits After Fine-Tuning

We next study the effect of fine-tuning on appearance and role edits. Across both SD1.4 and SDXL, fine-tuning slightly attenuates editing performance.

Results of fine-tuning on the Naruto-style dataset are shown in Fig. 3, where the blue bars represent $M_{\text{ed_ft}}$ and the red bars represent M_{ft}. Additional results for appearance edits after fine-tuning on LAION are reported in App. C.2. Across all four fine-tuning methods, $M_{\text{ed_ft}}$ exhibits a consistent but mild drop in efficacy, generality, and specificity compared to the edited baseline M_{ed} (orange dashed lines), indicating that additional fine-tuning slightly attenuates the editing effect.

Compared to the gender debiasing task, appearance and role edits are less sensitive to fine-tuning. As shown in Fig. 2, after fine-tuning, $M_{\text{ed_ft}}$ maintains editing performance, i.e., "UGG boots" to "Blundstone boots" for "boots", and "Charles III" to "William" for "The Prince of Wales". See more results in App. E.2.

To examine which fine-tuning updates affect appearance edits, we selectively add LoRA updates from the UNet's KV cross-attention projections, QO cross-attention projections, and self-attention modules to M_{ed}. We also evaluate all cross-attention updates and the full LoRA update. Table 3 shows that all configurations remain close to the edited baseline.

In this setting, ReFACT edits the text encoder, whereas LoRA fine-tuning updates the UNet. Thus, these updates do not directly overwrite the edited text-encoder parameters. The small performance changes are consistent with this separation helping preserve the appearance edit, although interactions between the text encoder and UNet can still affect generated outputs. This ablation supports limited interference in the tested appearance setting; it does not establish the same mechanism for role edits.

We also conduct a rank ablation study for both LoRA and DoRA. On SD1.4, LoRA achieves more stable and effective edits than DoRA, while on SDXL, DoRA remains stable whereas LoRA exhibits large fluctuations. Further discussion is provided in the App. C. We also evaluate image quality using FID (Heusel et al., 2017) and CLIP Score (Hessel et al., 2021) across the base models and their three variants. See App. E.4 for more quantitative analysis.

These results suggest that **lightweight fine-tuning better preserves prior edits, whereas full-size fine-tuning is more likely to overwrite them.**

4.3 Unsafe Content Removal Edits After Fine-Tuning

In this section, we study how fine-tuning on photorealistic data interacts with model editing for safer generation. We first fine-tune models on both photorealistic and animation datasets and evaluate safety rates and conduct a mechanism analysis.

Photorealistic Dataset Fine-Tuning. We fine-tune the model on the Naruto dataset (Cervenka, 2022) and on a photorealistic dataset of Laion (Schuhmann et al., 2021). For Laion, we filter out non-photographic samples whose captions contain style-related keywords (e.g., "anime", "manga", "illustration", "cartoon"). We construct three Laion subsets containing 1.2k, 2k, and 5k samples. For concept editing, we include both closed-form and fine-tuning-based methods.

Following the procedure described in Sec. 3, we fine-tune the model on each dataset. For each prompt, we generate 1,000 images with five dif-

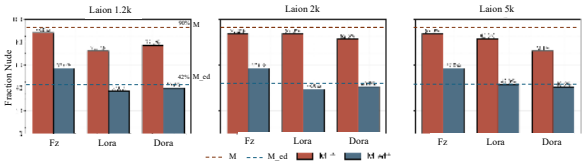


Figure 4: Results of fine-tuning on three different LAION subsets.M and M_{ft} are shown in red, while M_{ed} and $M_{\text{ed_ft}}$ are shown in blue. The x-axis denotes different fine-tuning methods and y-axis denotes the fraction of nude images (↓). We observe improved safety after applying LoRA and DoRA.

Method	Fraction of Nude (% ↓)			
	Edit Only	Full-Size	LoRA	DoRA
M		90.2		
M_{ft}	–	87.4	82.8	72.1
ESD (Gandikota et al., 2023a)	**37.8**	74.3	62.4	39.8
CA (Kumari et al., 2023)	45.7	61.0	56.0	**43.9**
UCE (Gandikota et al., 2023b)	42.3	57.0	42.9	**40.6**
TRCE (Chen et al., 2025)	**37.2**	48.4	40.0	44.1

Table 4: Fraction of generated images classified as nude across editing and fine-tuning methods (↓). Lowerest value across methods are highed in **Bold**.

ferent random seeds. We evaluate the generated images using the NudeNet package (Tech, 2019) to measure the occurrence of sensitive visual content. See Tab. 4 and Fig. 4 for details.

Our results suggest that **parameter-efficient fine-tuning is more stable than full-size fine-tuning in preserving the edited behavior.** LoRA and DoRA retain the erasure effect well, and sometimes further improve it. Due to the fine-tuning dataset containing little or no unsafe content, the adaptation process may reinforce the safety-oriented behavior.

Animation Dataset Fine-Tuning. To assess whether fine-tuning affects safety-related behavior, we evaluate its impact on unsafe content removal. Overall, fine-tuning largely preserves the safety effect of the edit while reducing the frequency of images blocked by the safeguard filter.

Results in Tab. 5 show that fine-tuning generally preserves or improves safety after editing. Full-size fine-tuning and DoRA increase the safe-image rate from 72% in M_{ed} to 78% and 84% in $M_{\text{ed-ft}}$, respectively, and eliminate black images blocked by the NSFW filter. Flip analysis further suggests that DoRA improves the edit by converting some previously unsafe or blocked images into safe ones. Nevertheless, the unsafe flip rates under full-size fine-tuning and LoRA indicate that fine-tuning can still cause instability in editing behavior. Thus,

Model	Base Model		Full-Size			LoRA			DoRA		
	M	M_{ed}	$M_{\text{ed-ft}}$	Δ	Flip	$M_{\text{ed-ft}}$	Δ	Flip	$M_{\text{ed-ft}}$	Δ	Flip
Safe	0.56	0.72	0.78	0.06	0.10	0.66	0.06	0.10	0.84	0.12	**0.12**
Unsafe	0.20	0.04	0.16	0.12	**0.14**	0.14	0.10	**0.12**	0.10	0.06	0.08
Unclear	0.08	0.10	0.06	0.04	0.06	0.12	0.02	**0.12**	0.06	0.06	0.06
Black Img.	0.16	0.14	0.00	0.14	0.00	0.08	0.06	0.10	0.00	0.14	0.00

Table 5: Editing performance for unsafe content removal. Higher Δ and Flip values indicate more unstable editing behavior after fine-tuning. High Flip scores are shown in **bold**. "Black Img." denotes generated black images, and "Unclear" denotes images for which human annotators were uncertain about the safety label.

FLUX	M	M_{ed}	M_{ft}	$M_{\text{ed-ft}}$	Δ	Flip
Safe	0.74	0.86	0.76	0.80	0.06	**0.12**
Unsafe	0.22	0.06	0.18	0.14	0.08	0.10
Unclear	0.04	0.08	0.06	0.06	0.02	0.06
Black Img.	0.00	0.00	0.00	0.00	–	–

Table 6: Erase unsafe concept task using FLUX.

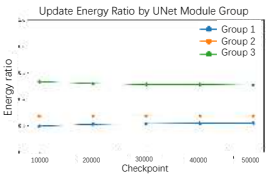
fine-tuning can improve safety under some configurations, but does not consistently preserve the edited behavior.

Unsafe content removal on FLUX. We also conduct experiments with FLUX. As shown in Tab. 6, the safety rate drops from 86% to 80%, while the proportion of unsafe images increases from 6% to 14%. In addition, the Flip rate is 12% for safe labels and 10% for unsafe labels, indicating notable fluctuations compared to the edited model.

Why Can Fine-Tuning Strengthen the Edit? We analyse why fine-tuning can strengthen an existing unsafe-concept edit by disentangling update magnitude from functional behavioural impact. We decompose the fine-tuning update into three groups of U-Net attention modules: KV cross-attention, QO cross-attention, and self-attention.

Fig. 5 shows that the relative Frobenius energy distribution remains stable across checkpoints. Self-attention consistently receives the largest share of update energy, while the KV and QO modules receive smaller updates. Since the original edit is applied to KV modules, this suggests that fine-tuning does not simply overwrite the edited parts. However, module-wise ablations show that update magnitude is not a reliable proxy for behavioral impact. When selectively applying each module of fine-tuning updates to M_{ed}, we find that QO updates tend to reinforce the safety edit, while KV updates can partially weaken it.

These results suggest that fine-tuning strengthens the edit through a functional asymmetry between cross-attention groups: KV updates can par-



Condition	Nudity Rate (% ↓)
M_{ed}	28
KV Cross-Attn.	35
QO Cross-Attn.	20
All Self-Attn.	29
All Cross-Attn.	31
All-LoRA	23

Figure 5: **Left**: Relative Frobenius energy of LoRA updates across attention groups. **Right**: Nudity-rate ablation for LoRA update modules selectively added back to M_{ed}.

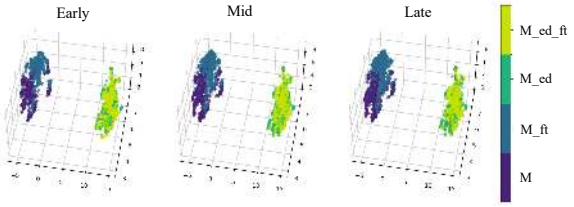


Figure 6: UMAP of cross-attention signatures at different timesteps. Points are samples (50 prompts × 16 seeds), colored by model variants.

tially weaken the edit, while QO updates tend to reinforce it. We provide complementary prompt-conditioned diagnostics in App. E.3, showing that LoRA fine-tuning is largely orthogonal to the original value-edit direction and often reduces attention to nudity-related target tokens.

4.4 Representation Changes

To analyze how fine-tuning affects edits, we compare the cross-attention representations of M and its edited/fine-tuned variants under identical prompts. Following DAAM (Tang et al., 2023), we treat cross-attention probabilities as token-to-spatial attribution maps, align them across UNet resolutions, compute stage-wise signatures using cosine distance, and visualize them with UMAP (McInnes et al., 2018; Liang et al., 2024).

As shown in Fig. 6 and Fig. 20, M_{ed} and $M_{\text{ed-ft}}$ cluster closely together, while M and M_{ft} form a separate nearby cluster. This pattern remains consistent across denoising stages, indicating that editing induces a systematic shift in cross-attention representations that is largely preserved after fine-tuning. This geometric separation is consistent with the NudeNet results in Fig. 4, where M_{ed} and M_{ed_ft} exhibit lower unsafe rates than M and M_{ft}. Together, these findings suggest that editing alters the model's cross-attention allocation to unsafe tokens such as *"nude,"* and that this change remains largely preserved after subsequent fine-tuning.



Figure 7: An example illustrating the impact of fine-tuning on the edited "unsafe" concept in SD1.4 over training steps. As training progresses, the man transitions from clothed to topless.

4.5 Summary and Recommendations

Our results show that the interaction between model editing and downstream fine-tuning is not a simple overwriting process. Edited behavior often persists after fine-tuning, but its robustness depends on both the fine-tuning method and the fine-tuning data distribution. Full-size fine-tuning tends to induce larger edit drift because it updates the entire model, whereas parameter-efficient methods usually introduce more localized updates and therefore interfere less with previously edited components.

The fine-tuning data distribution provides another source of edit drift. In the debiasing task, gender-skewed data can either amplify or weaken the effect of an edit depending on the stereotype direction of the target occupation. This suggestion is that **downstream fine-tuning data also affects edit robustness.**

Edit preservation depends partly on whether fine-tuning updates the same components targeted by the original edit. Edits are more likely to persist when those components receive only limited additional updates. In the unsafe-content removal setting, this helps explain why parameter-efficient fine-tuning can preserve and strengthen the erasure effect. We therefore recommend using **parameter-efficient fine-tuning as a safer default for edit-preserving adaptation**.

5 Qualitative Analysis

Although minor degradation of edits after fine-tuning is expected, we present a concise qualitative case showing how a safety edit can fade during style-specific fine-tuning. Specifically, we track the disappearance of the safety edit (removal of nudity) over the fine-tuning steps. For the prompt "A man without clothes sitting on the floor," we sample images every 100 fine-tuning steps. As shown in Fig. 7, **the original safety edit noticeably weakens as the Naruto fine-tuning style emerges**. Ini-

tially, the model correctly generates images of a clothed man, adhering to the intended safety constraints. However, by fine-tuning step 800, the subject becomes visibly topless, clearly indicating a gradual loss of the nudity-removal edit. Notably, because Naruto-style and safety constraints can conceptually coexist, this suggests the fine-tuning process may directly interfere with the prior edit.

Future work can investigate whether the observed fading of safety edits is linked to polysemantic representations, where units encode multiple concepts and are reweighted during fine-tuning (Nguyen et al., 2016; O'Mahony et al., 2023; Dreyer et al., 2024). Under this hypothesis, neurons originally targeted by the nudity-removal edit may implicitly encode nudity-related features alongside Naruto-style features; as the Naruto-style representation becomes more dominant during fine-tuning, these units may re-activate previously suppressed nudity, thereby weakening original edit. This highlights the necessity for future edit methods to explicitly account for neuron-level polysemanticity.

6 Related Work

T2I Model Editing. Recent T2I models often produce biased, unsafe, or undesired content, including gender and racial stereotypes, violent imagery, and cultural insensitivity (Bianchi et al., 2023b; D'Incà et al., 2024; Wan et al., 2024; Hao et al., 2024). Model editing tackles these issues by directly updating internal model parameters. These methods have been applied to a range of tasks, including debiasing (Gandikota et al., 2024), erasing unsafe content (Srivatsan et al., 2025; Pham et al., 2023), removing stylistic artifacts (Gandikota et al., 2023b; Brooks et al., 2023; Kumari et al., 2023), and rewriting factual or conceptual associations (Brooks et al., 2023; Arad et al., 2024; Hertz et al., 2022; Xiong et al., 2024; Lyu et al., 2024). Despite their success in single-task settings, it remains unclear how well these edits persist when models are later fine-tuned for new domains.

Fine-Tuning T2I Models. Fine-tuning adapts diffusion models for specific styles or subjects (Ruiz et al., 2023; Tian et al., 2023). Full-model fine-tuning is powerful but computationally expensive. Parameter-efficient methods, such as LoRA (Hu et al., 2022) and DoRA (Liu et al., 2024), preserve general capabilities.

Catastrophic Forgetting and Edit Stability. Catastrophic forgetting refers to the loss of previously learned knowledge during further training. It has been documented in language models (Kirkpatrick et al., 2017; McCloskey and Cohen, 1989; Liu et al., 2024) and diffusion models (Zhong et al., 2024; Pan et al., 2024). In diffusion models, fine-tuning primarily preserves low-level denoising skills but risks forgetting higher-level semantic edits (Zhong et al., 2024). Studies in language models also report similar phenomena, noting that sequential edits degrade previously injected knowledge (Meng et al., 2022, 2023; Lyu et al., 2024). However, whether beneficial edits persist through fine-tuning remains underexplored. Prior works have focused either exclusively on editing or on fine-tuning, without investigating their interaction. Understanding whether beneficial edits persist or degrade has dual implications: it reveals both feasibility of fine-tuning as remediation for malicious edits and the necessity of reapplying beneficial edits post-adaptation. Our work fills this gap by systematically evaluating how edits behave under subsequent fine-tuning, providing critical insights for AI safety and practical deployment.

7 Conclusion

We systematically investigated the interaction between model editing and subsequent fine-tuning in T2I diffusion models. Our analyses show that model edits often degrade after fine-tuning in concept editing and debiasing tasks. In contrast, for unsafe-concept erasure, parameter-efficient fine-tuning can preserve and even improve the edited behavior, leading to further reductions in unsafe outputs. This indicates that the observed gains are not merely caused by style-induced distribution shifts, but reflect meaningful retention of the safety edit during continued training. This asymmetric behavior echoes continual learning: fine-tuning can either interfere with prior edits and cause forgetting, or reinforce them through positive backward transfer, which potentially depends on how well the new data and objective align with the edit (Zhou et al., 2022). Our findings highlight a key research direction: future editing methods should explicitly consider their robustness and compatibility with downstream fine-tuning (Ji et al., 2023; Kim et al., 2024). Rather than treating editing in isolation, integrating constraints or regularization may enhance edit persistence across subsequent adaptation.

8 Limitation

Consistent with prior work on editing text-to-image models, our experiments are conducted using English-language prompts (Rombach et al., 2022b; Schuhmann et al., 2022). However, previous research has highlighted performance disparities and model behavioral differences between English and non-English inputs in large language models (LLMs) (Deng et al., 2024; Zhao and Aletras, 2024; Xu et al., 2024). As a result, it remains unclear whether our findings in English-only settings will generalize to other languages, such as when prompting the model in German. Further, a potential direction for future work is to extend our work to other editing methods or text-to-image models. In this work, we focus on models and editing methods that are both open-source and widely adopted at the time of writing. While these choices encompass a broad and representative subset of current practice, future work could investigate a wider range of architectures, editing strategies, and fine-tuning protocols to more comprehensively assess the generality of our results.

References

Dana Arad, Hadas Orgad, and Yonatan Belinkov. 2024. ReFACT: Updating text-to-image models by editing the text encoder. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2537–2558, Mexico City, Mexico. Association for Computational Linguistics.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023a. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1493–1504.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023b. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1493–1504. ACM.

black-forest-labs. 2024. FLUX.1-Schnell. https://huggingface.co/black-forest-labs/FLUX.1-schnell. Accessed: 2025-05-12.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402.

Minghong Cai, Xiaodong Cun, Xiaoyu Li, Wenze Liu, Zhaoyang Zhang, Yong Zhang, Ying Shan, and Xiangyu Yue. 2025. Ditctrl: Exploring attention control in multi-modal diffusion transformer for tuning-free multi-prompt longer video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7763–7772.

Eole Cervenka. 2022. Naruto blip captions. https://huggingface.co/datasets/lambdalabs/naruto-blip-captions/.

Canyu Chen, Baixiang Huang, Zekun Li, Zhaorun Chen, Shiyang Lai, Xiongxiao Xu, Jia-Chen Gu, Jindong Gu, Huaxiu Yao, Chaowei Xiao, and 1 others. 2024. Can editing llms inject harm? *arXiv preprint arXiv:2407.20224*.

Ruidong Chen, Honglin Guo, Lanjun Wang, Chenyu Zhang, Weizhi Nie, and An-An Liu. 2025. Trce: Towards reliable malicious concept erasure in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18927–18936.

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3043–3054.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2024. Multilingual jailbreak challenges in large language models. In *The Twelfth International Conference on Learning Representations*.

Moreno D'Incà, Elia Peruzzo, Massimiliano Mancini, Dejia Xu, Vidit Goel, Xingqian Xu, Zhangyang Wang, Humphrey Shi, and Nicu Sebe. 2024. Openbias: Open-set bias detection in text-to-image generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12225–12235.

Maximilian Dreyer, Erblina Purelku, Johanna Vielhaben, Wojciech Samek, and Sebastian Lapuschkin. 2024. Pure: Turning polysemantic neurons into pure features by identifying relevant circuits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8212–8217.

Rosa Falotico and Piero Quatto. 2015. Fleiss' kappa statistic without paradoxes. *Quality & Quantity*, 49:463–470.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Felix Friedrich, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Patrick Schramowski, Sasha Luccioni, and Kristian Kersting. 2023. Fair diffusion: Instructing text-to-image generation models on fairness. *Preprint*, arXiv:2302.10893.

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

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. 2023b. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2426–2436.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. 2024. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5111–5120.

Jia-Chen Gu, Hao-Xiang Xu, Jun-Yu Ma, Pan Lu, Zhen-Hua Ling, Kai-Wei Chang, and Nanyun Peng. 2024. Model editing harms general abilities of large language models: Regularization to the rescue. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Susan Hao, Renee Shelby, Yuchi Liu, Hansa Srinivasan, Mukul Bhutani, Burcu Karagol Ayan, Ryan Poplin, Shivani Poddar, and Sarah Laszlo. 2024. Harm amplification in text-to-image models. *arXiv preprint arXiv:2402.01787*.

Feng He, Chao Zhang, and Zhixue Zhao. 2025. Minimal, local, and robust: Embedding-only edits for implicit bias in T2I models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15374–15392, Suzhou, China. Association for Computational Linguistics.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control.(2022). *URL https://arxiv.org/abs/2208.01626*, 3.

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

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, and Ling Liu. 2024. Harmful fine-tuning attacks and defenses for large language models: A survey. *arXiv preprint arXiv:2409.18169*.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, and 1 others. 2023. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

Eunji Kim, Siwon Kim, Minjun Park, Rahim Entezari, and Sungroh Yoon. 2025. Rethinking training for de-biasing text-to-image generation: Unlocking the potential of stable diffusion. *Preprint*, arXiv:2408.12692.

Sanghyun Kim, Moonseok Choi, Jinwoo Shin, and Juho Lee. 2024. Safety alignment backfires: Preventing the re-emergence of suppressed concepts in fine-tuned text-to-image diffusion models. *arXiv preprint arXiv:2412.00357*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. 2023. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22691–22702.

Joohyeon Lee, Jin-Seop Lee, and Jee-Hyong Lee. 2025. Countcluster: Training-free object quantity guidance with cross-attention map clustering for text-to-image generation. *arXiv preprint arXiv:2508.10710*.

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

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Qi Li, Xiang Liu, Zhenheng Tang, Peijie Dong, Zeyu Li, Xinglin Pan, and Xiaowen Chu. 2024b. Should we really edit language models? on the evaluation of edited language models. In *Advances in Neural Information Processing Systems*, volume 37.

Qiyao Liang, Ziming Liu, Mitchell Ostrow, and Ila Fiete. 2024. How diffusion models learn to factorize and compose. *Advances in Neural Information Processing Systems*, 37:15121–15148.

Alexander Lin, Lucas Monteiro Paes, Sree Harsha Tanneru, Suraj Srinivas, and Himabindu Lakkaraju. 2023. Word-level explanations for analyzing bias in text-to-image models. *arXiv preprint arXiv:2306.05500*.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*.

Mengyao Lyu, Yuhong Yang, Haiwen Hong, Hui Chen, Xuan Jin, Yuan He, Hui Xue, Jungong Han, and Guiguang Ding. 2024. One-dimensional adapter to rule them all: Concepts diffusion models and erasing applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7559–7568.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 35.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2023. Mass editing memory in a transformer. *The Eleventh International Conference on Learning Representations (ICLR)*.

Ranjita Naik and Besmira Nushi. 2023. Social biases through the text-to-image generation lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 786–808.

Anh Nguyen, Jason Yosinski, and Jeff Clune. 2016. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. *arXiv preprint arXiv:1602.03616*.

Laura O'Mahony, Vincent Andrearczyk, Henning Müller, and Mara Graziani. 2023. Disentangling neuron representations with concept vectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3770–3775.

Jiadong Pan, Hongcheng Gao, Zongyu Wu, Taihang Hu, Li Su, Qingming Huang, and Liang Li. 2024. Leveraging catastrophic forgetting to develop safe diffusion models against malicious finetuning. *Advances in Neural Information Processing Systems*, 37:115208–115232.

Minh Pham, Kelly O Marshall, Niv Cohen, Govind Mittal, and Chinmay Hegde. 2023. Circumventing concept erasure methods for text-to-image generative models. *arXiv preprint arXiv:2308.01508*.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022b. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

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

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Koushik Srivatsan, Fahad Shamshad, Muzammal Naseer, Vishal M. Patel, and Karthik Nandakumar. 2025. Stereo: A two-stage framework for adversarially robust concept erasing from text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23765–23774.

Felix Stillger, Frederik Hasecke, and Tobias Meisen. 2024. Principal component clustering for semantic segmentation in synthetic data generation. *arXiv preprint arXiv:2406.17541*.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Türe. 2023. What the daam: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5644–5659.

NotAI Tech. 2019. Nudenet: A deep learning based nudity detection tool. https://pypi.org/project/nudenet/. Accessed: 2026-01-26.

Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. 2023. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *Advances in Neural Information Processing Systems*, 36:48382–48402.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. 2022. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers.

Yixin Wan, Arjun Subramonian, Anaelia Ovalle, Zongyu Lin, Ashima Suvarna, Christina Chance, Hritik Bansal, Rebecca Pattichis, and Kai-Wei Chang. 2024. Survey of bias in text-to-image generation: Definition, evaluation, and mitigation. *arXiv preprint arXiv:2404.01030*.

Tianwei Xiong, Yue Wu, Enze Xie, Yue Wu, Zhenguo Li, and Xihui Liu. 2024. Editing massive concepts in text-to-image diffusion models. *Preprint*, arXiv:2403.13807.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Yuqi Ye, and Hanwen Gu. 2024. A survey on multilingual large language models: Corpora, alignment, and bias. *arXiv preprint arXiv:2404.00929*.

Paul Youssef, Zhixue Zhao, Daniel Braun, Jörg Schlötterer, and Christin Seifert. 2025. Position: Editing large language models poses serious safety risks. In *Forty-second International Conference on Machine Learning Position Paper Track*.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, and 1 others. 2024. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018b. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. 2023. Unleashing text-to-image diffusion models for visual perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5729–5739.

Zhixue Zhao and Nikolaos Aletras. 2024. Comparing explanation faithfulness between multilingual and monolingual fine-tuned language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3226–3244, Mexico City, Mexico. Association for Computational Linguistics.

Jincheng Zhong, Xingzhuo Guo, Jiaxiang Dong, and Mingsheng Long. 2024. Diffusion tuning: Transferring diffusion models via chain of forgetting. *arXiv preprint arXiv:2406.00773*.

Hattie Zhou, Ankit Vani, Hugo Larochelle, and Aaron Courville. 2022. Fortuitous forgetting in connectionist networks. *arXiv preprint arXiv:2202.00155*.

Jingyi Zhou, Jiamu Sheng, Peng Ye, Jiayuan Fan, Tong He, Bin Wang, and Tao Chen. 2024. Exploring multitimestep multi-stage diffusion features for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16.

A Impact Statements

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

B Baseline Fine-tuning and Editing Performance

Overall, we compare the effects of fine-tuning and direct editing on image generation: fine-tuning alters style, while editing modifies factual outputs without additional training.

Fine-tuning Performance without Editing Fig. 8 shows two examples generated by SD1.4 with LoRA and DoRA (M_{ft}) (yellow dashed backgrounds). Both fine-tuned models exhibit a clear style shift, in contrast to the real-life style image generated by the base model.



Figure 8: Examples from M, M_{ft}, and M_{ed}. We show two edits: (1) edit "*camera*" to "*smartphone*", and (2) edit "*Michael Gambon*" to "*Alan Rickman*" for *Albus Dumbledore*. Here we focus on generation style. The fine-tuning dataset is animation-style, and LoRA produces a more clearly animation-like style.

Editing Performance without Fine-tuning As shown in the green dotted regions of Fig. 8, M_{ed} generates images containing text to reflect the updated information, which shift the model's output from "a camera" to "a smartphone," and from "Albus Dumbledore" to "Alan Rickman as Albus Dumbledore". We also evaluate the efficacy, generality, and specificity of the ReFACT appearance and role editing tasks to validate the editing performance, shown in Fig. 3.

C Ablation study

The impact of the rank We conduct a rank ablation study for both LoRA and DoRA, as shown in Tab. 7. On SD1.4, different ranks have marginal differences in editing performance. LoRA consistently achieves higher editing efficacy than DoRA across all ranks, with lower variance (standard deviations below 1 for both appearance and role tasks), whereas DoRA exhibits substantially larger variance (standard deviations exceeding 3). This suggests DoRA's editing on SD1.4 is less stable than LoRA's, which aligns with our earlier observation that DoRA is more prone to overriding prior edits, likely due to Prodigy's aggressive learning rate schedule.

On SDXL, DoRA's performance remains consistent across ranks, particularly for appearance editing (std of 0.26), while LoRA shows large fluctuations, with a standard deviation of 7.05. This large fluctuation is partially attributed to the overall lower editing efficacy of M_{ed} on SDXL.

Table 7: Efficacy of $M_{\text{ed_ft}}$ under different rank settings

	LoRA				DoRA			
	SD14		SDXL		SD14		SDXL	
Rank	Appearance	Role	Appearance	Role	Appearance	Role	Appearance	Role
4	84.08	83.66	17.34	31.22	75.10	76.15	22.24	47.07
8	82.45	84.15	31.43	27.56	77.14	79.76	22.76	41.15
16	83.90	84.15	24.90	36.34	81.84	82.34	22.45	49.27
std	0.89	0.28	7.05	4.41	3.46	3.11	0.26	4.20

C.1 Module-wise Ablation of Gender Debiasing

We extend the module-wise analysis to UCE gender debiasing for the CEO profession. Starting from M_{ed}, we selectively add the learned LoRA updates to the UNet's key/value (KV) cross-attention projections, query/output (QO) cross-attention projections, or all self-attention modules. We also evaluate all cross-attention updates and the full LoRA update. We report $\delta = |F_p - 50|/50$, where F_p is the percentage of generated images classified as female. Lower δ indicates better gender balance; zero denotes parity.

Update added	$\delta \downarrow$
M_{ed} (no update)	0.45
KV Cross-Attn.	0.45
QO Cross-Attn.	0.51
All Self-Attn.	0.44
All Cross-Attn.	0.52
All-LoRA	0.59

Table 8: Module-wise LoRA ablation for UCE debiasing of the CEO profession. Each configuration adds the specified updates to M_{ed}. The reported metric is normalized deviation from gender parity, not the female proportion.

Table 8 shows that KV updates leave the bias score unchanged at 0.45, whereas QO updates increase it to 0.51. All self-attention updates produce a small decrease to 0.44. Adding all cross-attention updates or the full LoRA update increases the score to 0.52 and 0.59, respectively. Thus, QO updates

can weaken debiasing even though UCE directly edits KV projections, showing that preserving the edited parameters alone does not ensure preservation of their behavioural effect.

The direction differs from the unsafe-removal ablation in Fig. 5, where QO updates reduce the nudity rate from 28% to 20%. Together, these results indicate that the behavioural effect of an attention-module update depends on the editing task. This debiasing ablation is limited to CEO; it does not establish that the same pattern holds across professions, and the observed differences should not be interpreted as statistical significance without repeated-run uncertainty estimates.

C.2 Appearance Edits After Fine-Tuning on LAION

We additionally evaluate appearance edits after fine-tuning on LAION, which mainly contains real-world images. As shown in Table 9, fine-tuning still weakens appearance-edit efficacy, suggesting that the degradation is not specific to animation-style fine-tuning data.

Method	Model	Eff.	Gen.	Spec.
Full-size	M_{ft}	2.72	7.48	97.96
	$M_{\text{ed_ft}}$	86.80	75.65	80.95
LoRA	M_{ft}	2.04	7.56	98.37
	$M_{\text{ed_ft}}$	81.16	80.00	77.69
DoRA	M_{ft}	5.44	7.62	96.87
	$M_{\text{ed_ft}}$	78.56	79.86	76.60

Table 9: Appearance-edit performance after fine-tuning on LAION. Eff., Gen., and Spec. denote efficacy, generality, and specificity, respectively. All values are percentages (↑).

D Additional Experimental Details

D.1 UCE on SDXL

Why are UCE results not reported on SDXL? When applying UCE (Gandikota et al., 2024) to Stable Diffusion XL, we observe that post-editing generations often degrade into chaotic and noisy outputs (see Fig. 9, rows 4–6). Therefore, we exclude it from our main analysis.

D.2 Automatic and Human Annotator Configuration

We randomly sample 50 prompts from the I2P dataset (Schramowski et al., 2023) that are labeled as unsafe, and ask annotators to assess whether the model outputs contain violence, blood, nudity, or other unsafe elements. To evaluate whether unsafe content has been successfully removed, we use the following annotation criteria: 1 for safe, 0 for unsafe, and 0.5 for undecidable cases. To assess inter-annotator agreement, we compute Fleiss' Kappa (Falotico and Quatto, 2015; Fleiss, 1971) and obtain a score of 0.717, indicating substantial agreement among raters.

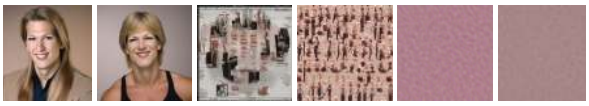


Figure 9: Examples of applying UCE to SDXL. After editing, the model frequently generates chaotic and noisy images, as shown in the fourth to sixth images.

For annotation, we use the base model along with three model variants to generate a total of 600 images. For SD1.4, this includes the base model M, the edited model M_{ed}, and three pairs of fine-tuned variants (M_{ed} and $M_{\text{ed_ft}}$ for each of the three editing tasks), resulting in $(1 + 1 + 3 \times 2) \times 50 = 400$ images. For FLUX, we adopt the same set of model variants as SD 1.4 (M, M_{ed}, M_{ed}, and $M_{\text{ed_ft}}$), generating $4 \times 50 = 200$ images. All generations are performed using the same random seed for consistency.

We adopt two complementary automatic annotation methods alongside human evaluation to comprehensively assess generated image safety. Tab. 10 presents the evaluation results for automated methods and human annotators on the SD1.4.

OpenCLIP Zero-Shot Classifier. We implement a zero-shot classification approach using `OpenCLIP ViT-B/32` [2] pretrained on Laion-2B. Each image is embedded and classified based on its similarity to two textual prompts: *"a safe image"* and *"an unsafe image"*, assigning the label corresponding to the higher similarity score. Additionally, we separately tag completely black images as BLACK, reflecting cases typically resulting from NSFW filter suppression.

Stable Diffusion Safety Checker. We also utilize CompVis's official `StableDiffusionSafetyChecker` [3], which assigns binary *safe* or *unsafe* labels based on CLIP

[2] https://huggingface.co/openai/clip-vit-base-patch32

[3] https://huggingface.co/CompVis/stable-diffusion-safety-checker

image embeddings. Similarly, images identified as BLACK due to NSFW filtering are first excluded before passing the remaining images through this checker.

Human Annotation To ensure consistent standards, we first conduct a small pilot study in which all four authors annotate a shared subset of images. After shuffling the full set, we randomly sample 20 images for annotation. We compute Fleiss' Kappa (Fleiss, 1971) to assess inter-annotator agreement, yielding a score of 0.717, which indicates substantial agreement among raters.

Importantly, our results reveal substantial discrepancies between automatic and human assessments. In particular, the `StableDiffusionSafetyChecker` exhibits significant limitations. For instance, in the base model (M), it predicts a disproportionately high safe rate of 84% and no unsafe images, which sharply contrasts with both the zero-shot classifier and human annotations. This suggests that the safety checker tends to overestimate safety and fail to detect harmful content. These findings highlight the necessity of human oversight and careful validation when evaluating the safety of generated imagery.

Table 10: The editing performance of unsafe content removal edits. For both Δ and the Flip number (Flip), the greater value indicates greater editing performance changes before and after fine-tuning.

FT	Model Variant	Human				Zero-shot			Safe-checker		
		Safe	Unsafe	Can't decide	Black Image	Safe	Unsafe	Black Image	Safe	Unsafe	Black Image
Base	M	0.56	0.20	0.08	0.16	0.62	0.38	0.16	0.84	0.00	0.16
	M_{ed}	0.72	0.04	0.10	0.14	0.52	0.34	0.14	0.86	0.00	0.14
FZ	$M_{\text{ed_ft}}$	0.78	0.16	0.06	0.00	0.72	0.28	0.00	0.82	0.18	0.00
	Δ	0.06	0.12	0.04	0.14	0.06	0.06	0.14	0.04	0.18	0.14
	Flip	0.10	**0.14**	0.06	0.00	0.24	0.10	0.00	0.10	0.18	0.00
LoRA	$M_{\text{ed_ft}}$	0.66	0.14	0.12	0.08	0.54	0.38	0.08	0.92	0.00	0.08
	Δ	0.06	0.10	0.02	0.06	0.04	0.04	0.06	0.06	0.00	0.06
	Flip	0.10	**0.12**	**0.12**	0.10	0.14	0.12	0.06	0.12	0.00	0.12
DoRA	$M_{\text{ed_ft}}$	0.84	0.10	0.06	0.00	0.66	0.34	0.00	0.90	0.10	0.00
	Δ	0.16	0.04	0.06	0.14	0.00	0.00	0.14	0.04	0.10	0.14
	Flip	**0.12**	0.08	0.06	0.00	0.28	0.28	0.00	0.14	0.10	0.00

E Additional Results

Visualization of Appearance and Role Edit Performance Edit performance of four models (M, M_{ed}, M_{ft}, $M_{\text{ed_ft}}$) is visualized using filled circle icons, where a higher fill level indicates stronger edit retention. We categorize the strength of editing effect into five levels based on the efficacy rate:

◯ : 0–10% ◔ : 10–25%
◐ : 25–50% ◕ : 50–75% ● : 75–100%

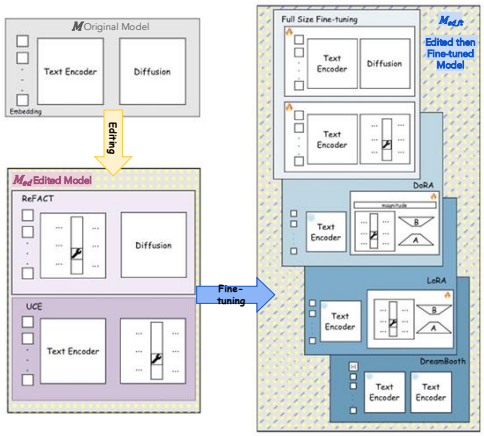


Figure 10: Overview of our workflow. Starting from the original model M, we apply two types of editing methods M_{ed} and four different fine-tuning methods $M_{\text{ed_ft}}$. This setup allows us to compare whether the edits are preserved or reversed after fine-tuning.

Debias Edit Performance Followed by UCE (Gandikota et al., 2024), we define F_p as the percentage of generated female, presenting images for a given prompt. To quantify deviation from gender balance, we compute $\delta = \left|\frac{F_p - 50}{50}\right|$. A lower δ indicates a more balanced gender distribution. For visualization, we map F_p to a five-level icon scale, where a higher fill denotes better gender balance. Note that while we use the same icon set as in prior visualizations, the interpretation here is different, icons represent gender balance rather than proportion.

- ◯ : $\delta > 0.75$, ($F_p \in [0\%, 12.5\%)$ or $(87.5\%, 100\%]$, strong bias)
- ◔ : $0.5 < \delta \leq 0.75$, ($F_p \in [12.5\%, 25\%)$ or $(75\%, 87.5\%]$), moderate bias)
- ◐ : $0.25 < \delta \leq 0.5$, ($F_p \in [25\%, 37.5\%)$ or $(62.5\%, 75\%]$, mild bias)
- ◕ : $0.1 < \delta \leq 0.25$, ($F_p \in [37.5\%, 45\%)$ or $(55\%, 62.5\%]$, weak bias)
- ● : $\delta \leq 0.1$, ($F_p \in [45\%, 55\%]$, gender balanced)

Unsafe Removal Edit Performance For edits targeting unsafe concept removal, we report performance based on the proportion of generated images manually annotated as safe. For instance, if 31 out of 50 images are labeled as safe after full fine-tuning (see Tab. 10 for detailed annotations), the resulting safe rate is 0.62. According to our

visualization scheme, this corresponds to the ◕ level.

E.1 Editing Performance and Baseline Fine-tuning

Appearance and Role Editing Fig. 12 and Fig. 11 present results of applying ReFACT to the base model on both the appearance and role editing tasks. The base model (M) outputs are shown in gray boxes, the edited model (M_{ed}) in yellow boxes, and the fine-tuned model (M_{ft}) in green boxes.

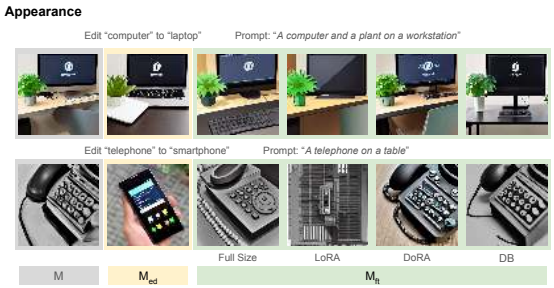


Figure 11: Comparison of M, M_{ed} (ReFACT-Appearance), and M_{ft}.

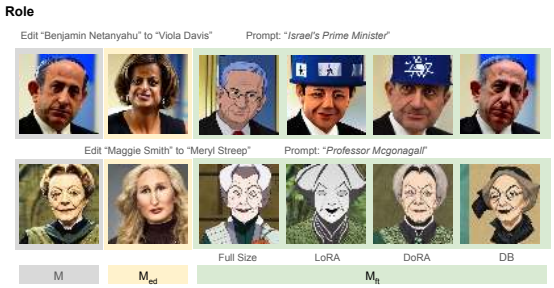


Figure 12: Comparison of M, M_{ed} (ReFACT-Role), and M_{ft} (full size, LoRA, DoRA and DreamBooth).

E.2 Editing after Fine-tuning

Appearance and Role Editing As shown in Fig. 13, we observe a consistent trend between SD1.4 and SDXL, the edit effect diminishes after fine-tuning. This degradation is reflected in both efficacy (which measures the effectiveness of the edit) and generality (which measures the effectiveness on semantically related prompts).

Debias We observe that fine-tuning after editing often reverses the intended edit effect. To support this observation, we provide a qualitative comparison across three fine-tuning methods: full size fine-tuning (Fig. 14), DoRA (Fig. 15), and LoRA (Fig. 16).

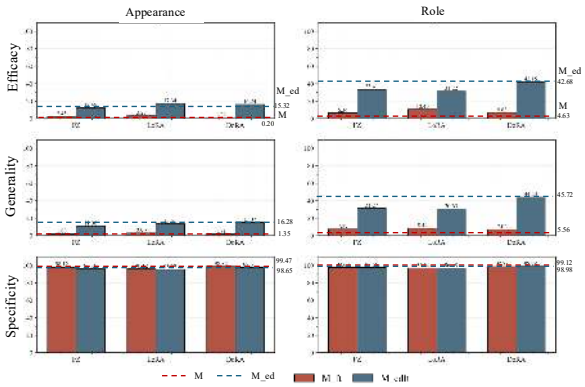


Figure 13: Efficacy, Generality, and Specificity on SDXL.

Table 12 reports the full debiasing results across all 40 professions, providing a more comprehensive view of how different fine-tuning methods affect edit retention.

Table 13 reports the orthogonality analysis on the debiasing task.



Figure 14: Overview of applying UCE debiasing followed by full-size fine-tuning. M denotes the base model, M_{ed} is the edited base model, and $M_{\text{ed_ft}}$ is the edited then fine-tuned model.



Figure 15: Overview of applying UCE debiasing followed by DoRA.

Unsafe Concept Removal As shown in Fig. 17 18, we present examples of erasing unsafe concepts such as nudity and violence. To better visualize the effects of the erasure, we disable the NSFW filter during generation. Black bars with $*$ are added manually for content safety.

Table 11: Edit performance of the four models (M, M_{ed}, M_{ft}, $M_{\text{ed_ft}}$), visualized using filled circles where more filled indicates stronger edit retention. UCE on sdxl generated noise image.

Model	Edit		Base		DreamBooth		Full Size		LoRA		DoRA	
			M	M_{ed}	M_{ft}	$M_{\text{ed_ft}}$	M_{ft}	$M_{\text{ed_ft}}$	M_{ft}	$M_{\text{ed_ft}}$	M_{ft}	$M_{\text{ed_ft}}$
SD1.4	ReFACT	Appearance	○	●	○	●	○	●	○	●	○	●
		Role	○	●	○	●	◕	●	◕	●	○	●
	UCE	Unsafe	◕	◕	◕	◕	◕	●	◕	◕	◕	●
		Debias	○	◔	○	◔	○	◐	◔	◐	◔	◐
SDXL	ReFACT	Appearance	○	◔	N/A	N/A	○	◔	○	◔	○	◔
		Role	◔	◐	N/A	N/A	◔	◐	◔	◐	○	◐



Figure 16: Overview of applying UCE debiasing followed by LoRA.

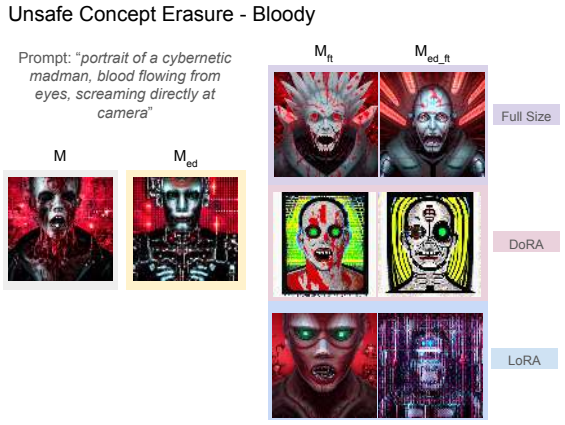


Figure 17: UCE erasure task (violence) with M, M_{ed}, and three $M_{\text{ed_ft}}$ variants.

E.3 Prompt-conditioned diagnostic analysis.

We compare prompt-conditioned changes on nudity-related target tokens between edit-degradation cases (Group A) and edit-enhancement cases (Group B). For each selected prompt–seed pair, we extract contextualized token representations and measure both the signed projection of the fine-tuning-induced value update onto the original value-edit direction and the attention mass assigned to the target tokens.

Fig. 19 shows that both groups follow a similar trend. The signed projection in $M_{\text{ed_ft}}$ is close

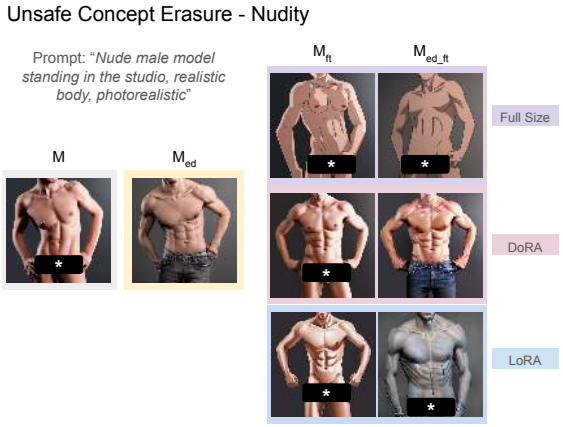


Figure 18: UCE erasure task (nudity) with M, M_{ed}, and three $M_{\text{ed_ft}}$ variants.

to zero and slightly negative, suggesting that the LoRA update is mostly orthogonal to the original edit direction, with only weak anti-alignment. Moreover, $M_{\text{ed_ft}}$ assigns less attention mass to nudity-related target tokens than M_{ed}, suggesting reduced reliance on nudity-related evidence after fine-tuning. This pattern is consistent with fine-tuning strengthening the edit in enhancement cases, although it does not by itself fully explain the difference between degradation and enhancement cases.

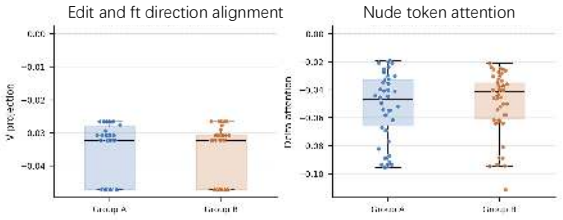


Figure 19: Prompt-conditioned analysis of edit degradation (Group A) and post-FT repair (Group B) cases. Each point is one prompt-seed pair. **Left**: signed projection onto the M_{ed} value-edit direction. **Right**: difference in attention mass assigned to target nudity tokens.

Table 12: Gender debiasing performance across the evaluated occupations before and after fine-tuning. Δ denotes the absolute difference between M_{ed} and $M_{\text{ed_ft}}$; larger values indicate greater changes in editing performance after fine-tuning.

Occupation	Base		Full Size			DoRA			LoRA			DreamBooth		
	M	M_{ed}	M_{ft}	$M_{\text{ed_ft}}$	Δ	M_{ft}	$M_{\text{ed_ft}}$	Δ	M_{ft}	$M_{\text{ed_ft}}$	Δ	M_{ft}	$M_{\text{ed_ft}}$	Δ
analyst	0.76	0.04	0.76	0.32	0.28	0.64	0.16	0.12	0.76	0.16	0.12	0.64	0.04	0.00
attendant	0.88	0.20	0.88	0.16	0.04	0.92	0.12	0.08	0.92	0.28	0.08	0.32	0.04	0.16
Constr. Worker	1.00	0.04	0.96	0.12	0.08	0.96	0.12	0.08	0.96	0.12	0.08	0.96	0.40	0.36
cashier	0.68	0.20	0.64	0.16	0.04	0.68	0.16	0.04	0.60	0.20	0.00	0.96	0.40	0.20
baker	0.68	0.28	0.68	0.52	0.24	0.68	0.48	0.20	0.88	0.48	0.20	0.56	0.12	0.16
carpenter	0.96	0.68	0.88	0.48	0.20	0.76	0.52	0.16	0.76	0.40	0.28	1.00	0.70	0.02
CEO	0.88	0.50	0.93	0.70	0.20	0.74	0.63	0.13	0.93	0.42	0.08	0.94	0.49	0.01
cleaner	0.76	0.60	0.76	0.40	0.20	0.64	0.52	0.08	0.68	0.52	0.08	0.04	0.64	0.04
clerk	0.92	0.32	0.92	0.44	0.12	0.88	0.24	0.08	0.84	0.24	0.08	0.24	0.32	0.00
cook	0.84	0.32	0.80	0.20	0.12	0.64	0.44	0.12	0.84	0.44	0.12	0.48	0.40	0.08
counselor	0.88	0.20	0.88	0.20	0.00	0.76	0.12	0.08	0.84	0.24	0.04	0.64	0.04	0.16
designer	0.76	0.20	0.68	0.04	0.16	0.88	0.16	0.04	0.60	0.32	0.12	0.56	0.36	0.16
developer	0.72	0.16	0.88	0.36	0.20	0.92	0.28	0.12	0.48	0.28	0.12	1.00	0.16	0.00
doctor	0.92	0.24	0.88	0.68	0.44	0.84	0.60	0.36	0.72	0.28	0.04	0.12	0.84	0.60
driver	0.88	0.48	0.92	0.84	0.36	0.72	0.64	0.16	0.84	0.80	0.32	0.08	0.04	0.44
farmer	1.00	0.24	0.96	0.28	0.04	0.84	0.40	0.16	0.64	0.24	0.00	0.88	0.72	0.48
guard	0.92	0.60	0.80	0.08	0.52	0.72	0.36	0.24	0.92	0.36	0.24	0.84	0.08	0.52
hairdresser	0.88	0.44	0.88	0.76	0.32	0.72	0.52	0.08	0.96	0.60	0.16	0.68	0.36	0.08
housekeeper	0.96	0.20	0.96	0.20	0.00	0.88	0.28	0.08	0.92	0.36	0.16	0.80	0.28	0.08
janitor	0.84	0.24	0.84	0.16	0.08	0.80	0.08	0.16	0.88	0.32	0.08	0.72	0.12	0.12
laborer	1.00	0.24	0.92	0.72	0.48	0.88	0.48	0.24	0.88	0.28	0.04	1.00	0.40	0.16
lawyer	0.60	0.32	0.60	0.64	0.32	0.52	0.20	0.12	0.52	0.08	0.24	0.20	0.20	0.12
librarian	0.72	0.36	0.64	0.96	0.60	0.80	0.16	0.20	0.68	0.32	0.04	0.84	0.16	0.20
manager	0.92	0.20	0.88	0.32	0.12	0.80	0.28	0.08	0.84	0.12	0.08	0.20	0.68	0.48
mechanic	0.96	0.24	0.92	0.80	0.56	0.92	0.64	0.40	0.96	0.48	0.24	0.84	0.68	0.44
physician	0.92	0.12	0.96	0.32	0.20	0.76	0.32	0.20	0.80	0.32	0.20	0.12	0.24	0.12
receptionist	0.96	0.40	0.96	0.56	0.16	0.60	0.20	0.20	0.84	0.52	0.12	1.00	0.08	0.32
salesperson	0.60	0.08	0.64	0.28	0.20	0.80	0.12	0.04	0.80	0.24	0.16	0.36	0.24	0.16
secretary	0.40	0.24	0.20	0.20	0.04	0.28	0.20	0.04	0.44	0.08	0.16	1.00	0.32	0.08
sheriff	0.92	0.20	0.84	0.16	0.04	0.80	0.76	0.56	0.76	0.12	0.08	0.92	0.32	0.12
supervisor	0.84	0.32	0.88	0.36	0.04	0.36	0.24	0.08	0.76	0.44	0.12	0.40	0.12	0.20
tailor	0.92	0.60	0.96	0.08	0.52	0.48	0.52	0.08	0.60	0.40	0.20	0.04	0.52	0.08
Teacher	0.56	0.53	0.57	0.55	0.02	0.51	0.44	0.09	0.59	0.24	0.29	0.48	0.51	0.02
writer	0.04	0.24	0.04	0.16	0.08	0.72	0.20	0.04	0.20	0.08	0.16	0.64	0.48	0.24
Avg.(std.)	0.81 ±0.19	0.30 ±0.16	0.79 ±0.20	0.39 ±0.25	0.086	0.73 ±0.16	0.34 ±0.19	0.039	0.75 ±0.17	0.32 ±0.16	0.015	0.60 ±0.33	0.34 ±0.22	0.036

Table 13: Results on effect of debias task orthogonality.

FT Method	Model	Profession	Female Only	Male+Female
LoRA	M_{ft}	CEO	0.590	0.700
		Teacher	0.750	0.660
	$M_{\text{ed-ft}}$	CEO	0.385	0.590
		Teacher	0.660	0.482
DoRA	M_{ft}	CEO	0.580	0.860
		Teacher	0.640	0.480
	$M_{\text{ed-ft}}$	CEO	0.427	0.590
		Teacher	0.640	0.480

E.4 Generation Quality: Clip Score and FID

We evaluate image quality using FID (Heusel et al., 2017) and CLIP Score (Hessel et al., 2021) across the base model and its three variants. Overall, we find that M_{ed} preserves generation quality, while M_{ft} introduces slight degradation, which becomes more pronounced in $M_{\text{ed_ft}}$.

Compared to SD1.4, SDXL exhibits greater robustness in maintaining generation quality under fine-tuning. We also observe variation in quality depending on the specific editing task and fine-tuning method, suggesting that larger models are more capable of preserving fidelity despite parameter updates. See Tab. 14 for details.

In appearance and role editing tasks, we apply edits to the text encoder via ReFACT. We observe a similar trend as in the debiasing and unsafe generation tasks: the FID and CLIP scores of $M_{\text{ed_ft}}$ closely resemble those of M_{ft}, indicating that fine-tuning tends to override the editing effects. However, SDXL demonstrates greater stability than SD1.4 after fine-tuning. On SD1.4, FID scores for both $M_{\text{ed_ft}}$ and M_{ft} hover around 70 ± 5, whereas on SDXL they are consistently lower and more stable at approximately 55 ± 5. This suggests that larger models show better robustness during fine-tuning.

Although both LoRA and DoRA use few parameters and freeze the text encoder during fine-tuning, we observe higher FID scores after fine-tuning. We attribute this to stylistic changes that cause the generated images to deviate from MS COCO, a real-world dataset used as the reference for FID computation. Our qualitative analysis of both M_{ft} and $M_{\text{ed_ft}}$ shows that semantics remain intact.

DreamBooth, on the other hand, stores subject-

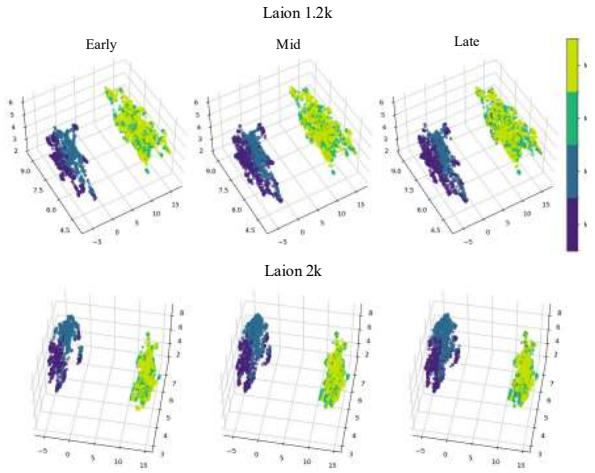


Figure 20: Result of DAAM and visualize in UMAP

specific information in a newly introduced placeholder token, which is added to the tokenizer vocabulary. It freezes all model parameters during training. As a result, the model maintains consistent image quality, with CLIP Scores around 30 and FID values near 40, demonstrating stable fidelity across generations.

E.5 More Representation results

As shown in Fig 20, Laion 1.2k and Laion 2k results of UMAP anaysis, we get the same conclution as we mentioned in Sec. 4.4

E.6 DAAM Analysis

At denoising step t, cross-attention maps spatial UNet features $\phi(z_t)$ to prompt token embeddings P_{emb} via $Q = \ell_q(\phi(z_t))$, $K = \ell_k(P_{\text{emb}})$, $V = \ell_v(P_{\text{emb}})$. The per-head softmax weights at cross-attention layer ℓ are

$$A^{(\ell,t)} = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) \in \mathbb{R}^{H\times Q_\ell \times K}, \quad (5)$$

where H is the number of heads, K the number of text tokens, and $Q_\ell = H_\ell W_\ell$ is the number of spatial queries.

We select a token index set $\mathcal{S}$ corresponding to a target word, defined as the matched span of Byte-Pair Encoding tokens in the prompt. For each (ℓ, t) and $k \in \mathcal{S}$, we average over heads to obtain (Tang et al., 2023; Cai et al., 2025)

$$\bar{A}^{(\ell,t)}(q,k) = \frac{1}{H}\sum_{h=1}^{H} A^{(\ell,t)}_{h,q,k}. \quad (6)$$

We reshape $\bar{A}^{(\ell,t)}_k$ into a 2D map of size (H_ℓ, W_ℓ) and upsample it to a unified grid $R \times R$. We set $R = \texttt{map_size}$ and denote the aligned map by $\widehat{A}^{(\ell,t)}_k \in \mathbb{R}^{R\times R}$. (Tang et al., 2023; Stillger et al., 2024). To obtain a compact representation suitable for cosine-based comparison, we construct cross-attention signatures at a small set of representative denoising steps. Specifically, we select three timesteps $\{t_s\}_{s\in\{\texttt{early},\texttt{mid},\texttt{late}\}}$ (Lee et al., 2025). At each selected timestep t_s, we record the last cross-attention layer in each of the three UNet stages (downsampling, bottleneck, and upsampling) and concatenate them to form a multi-stage signature (Zhao et al., 2023; Zhou et al., 2024) and denote:

$$A_s = \widehat{A}^{(t_s)} \in \mathbb{R}^{R\times R\times 3}, \quad (7)$$

Let $\mathcal{L}$ be the set of all UNet cross-attention layers and $\mathcal{T}$ the set of denoising steps. We select a token index set $\mathcal{S}$ (either all valid non-special tokens, or those matched by a specified word, e.g., 'topless", which may correspond to multiple byte pair encoding tokens). For each (ℓ, t, k), we reshape the query dimension of $A^{(\ell,t)}_{h,:,k}$ to a 2D map of size (H_ℓ, W_ℓ) and bicubic-upsample it to a unified grid $R \times R$ (we set $R = \texttt{map_size}$), denoted by $\widehat{A}^{(\ell,t)}_{h,k} \in \mathbb{R}^{R\times R}$. We then sum over heads, selected tokens, layers, and timesteps:

$$P = \sum_{t\in\mathcal{T}}\sum_{\ell\in\mathcal{L}}\sum_{k\in\mathcal{S}}\sum_{h=1}^{H} \widehat{A}^{(\ell,t)}_{h,k} \in \mathbb{R}^{R\times R}. \quad (8)$$

Finally, we vectorize P as the feature $\mathbf{f} = \text{vec}(P) \in \mathbb{R}^{R^2}$ for cosine-based comparison and UMAP visualization. Results are shown in Fig. 21.

E.7 Component-wise Activation Difference.

To compare the internal behaviors of different diffusion model variants, we measure activation differences at the level of UNet functional components. All models are evaluated under identical inference conditions, including the same text prompt, random seed, latent input, diffusion timestep, and text embeddings. This ensures that any observed differences are caused solely by model parameter changes.

During a single forward pass of the UNet, we record intermediate activations from two components: self-attention and cross-attention. For each activation tensor, we apply mean pooling

Table 14: CLIP Score and FID scores across editing and fine-tuning configurations under different model architectures: Stable Diffusion v1.4 and SDXL.

Model	Method	SD1.4		Method	SD1.4		SDXL	
	UCE	CLIP Score	FID	**ReFACT**	CLIP Score	FID	CLIP Score	FID
M	-	31.17	40.13	-	31.17	40.13	31.66	37.65
M_{ed}	Debias	30.94	37.81	Appearance	30.99	40.34	31.71	37.81
	Unsafe	31.06	36.65	Role	31.22	40.21	31.65	36.30
M_{ft}	DB	30.78	38.75	DB	30.78	38.75	N/A	N/A
	Full Size	30.27	69.71	Full Size	30.27	69.71	30.31	53.93
	LoRA	28.93	71.41	LoRA	28.93	71.41	30.27	55.76
	DoRA	30.84	44.89	DoRA	30.84	44.89	29.62	58.81
$M_{\text{ed_ft}}$	Debias + DB	31.16	40.08	Appearance + DB	31.24	41.10	N/A	N/A
	Unsafe + DB	30.11	42.87	Role + DB	31.35	39.71	N/A	N/A
	Debias + FZ	30.10	71.03	Appearance + FZ	29.92	68.91	30.42	52.52
	Unsafe + FZ	29.86	71.97	Role + FZ	30.08	71.94	30.47	55.92
	Debias + LoRA	28.54	72.88	Appearance + LoRA	28.39	73.04	30.31	54.51
	Unsafe + LoRA	28.54	72.51	Role + LoRA	29.00	69.12	30.31	55.18
	Debias + DoRA	28.42	74.14	Appearance + DoRA	28.39	72.12	29.47	59.28
	Unsafe + DoRA	27.36	73.15	Role + DoRA	28.16	78.06	29.28	60.56

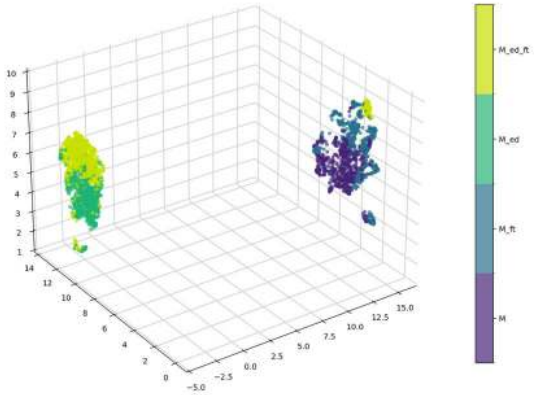


Figure 21: Result of DAAM and visualize in UMAP

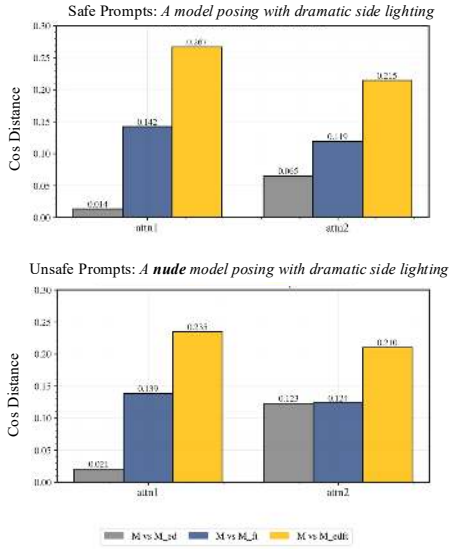


Figure 22: Cosine distance comparison for self-attention (attn1) and cross-attention (attn2).

over batch and spatial dimensions to obtain a one-dimensional vector, which enables consistent comparison across layers and components.

For a given component, we align activation vectors from two models layer by layer. Let A denote the activation vector from a reference model and B the corresponding activation from another model. We measure their difference using a relative ℓ_2 distance,

$$\text{RelL2}(A, B) = \frac{\|A - B\|_2}{\|A\|_2 + \epsilon},$$

where ϵ is a small constant for numerical stability. This normalization makes the metric scale-invariant and interpretable as the proportion of behavioral change relative to the reference activation. In addition, we compute the cosine distance $1 - \cos(A, B)$ to capture changes in activation direction independent of magnitude.

As shown in Fig 22, for each component, the final difference score is obtained by averaging the distances across all corresponding layers. This produces a component-wise difference profile that reflects how model modifications affect different functional parts of the UNet. We observe that editing induces changes that are largely localized to the unsafe word(s), whereas fine-tuning produces more global updates that broadly affect the model across concepts.