Why Instruction-Based Unlearning Fails in Diffusion Models?

Anonymous ACL submission

Abstract

Instruction-based unlearning has proven effective for modifying the behavior of large language models at inference time, but whether this paradigm extends to other generative models remains unclear. In this work, we investigate instruction-based unlearning in diffusion-based image generation models and show, through controlled experiments across multiple concepts and prompt variants, that diffusion models systematically fail to suppress targeted concepts when guided solely by natural-language unlearning instructions. By analyzing both the CLIP text encoder and cross-attention dynamics during the denoising process, we find that unlearning instructions do not induce sustained reductions in attention to the targeted concept tokens, causing the targeted concept representations to persist throughout generation. These results reveal a fundamental limitation of prompt-level instruction in diffusion models and suggest that effective unlearning requires interventions beyond inference-time language control.

1 Introduction and Background

Diffusion models (Croitoru et al., 2023; Yang et al., 2023), empowered by large-scale training on massive datasets, have achieved remarkable success in generating high-quality content across a wide range of modalities (Zhao et al., 2023; Ruan et al., 2023). However, due to incomplete data curation and monitoring processes, training corpora may inadvertently contain sensitive (Du et al., 2013), non-consensual (Viola and Voto, 2023), or copyrighted content (Zhang et al., 2023). As a result, diffusion-based generative models raise growing concerns regarding legal compliance, ethical deployment, and safe usage in real-world applications (Pujari et al., 2022). While constructing carefully curated and fully compliant datasets is a principled solution, retraining large diffusion models from scratch is often prohibitively expensive in terms of both computational cost and time (Ma et al., 2025). Consequently, it is increasingly important to investigate post hoc mechanisms that can modify or correct the behavior of already-trained models, enabling targeted content removal or behavioral adjustment without full retraining.

Figure 1: Motivating experiment evaluating instruction-based unlearning in diffusion models. More examples, including experiments on SD-XL and explicit use of unlearning instructions, can be found in section A.

A growing body of work has explored concept unlearning in diffusion models, with most approaches relying on fine-tuning-based interventions (Gao et al., 2025; Fuchi and Takagi, 2024; Liu and Tan, 2024; Schioppa et al., 2024). For example, Zhang et al. (2024) suppresses targeted concepts by minimizing cross-attention activations between the concept tokens and corresponding visual features during fine-tuning. Similarly, Gandikota et al. (2023) and Wu et al. (2025) propose to modify diffusion models by aligning the visual feature distributions generated from targeted concepts with those produced by empty or neutral textual descriptions, thereby mitigating the issue of missing negative samples. While these methods demonstrate promising results, they require additional fine-tuning of large diffusion models, incurring substantial computational and memory overhead.

Inspired by the success of instruction-based unlearning in large language models (Pawelczyk et al., 2023), which enables effective behavior modification through simple natural language instructions at inference time, a natural question arises ($\mathcal{Q}$): *can diffusion models similarly unlearn specific concepts by following textual instructions at test time?* In this work, we show that, despite their ability to condition on and respond to textual prompts, instruction-based unlearning systematically fails in diffusion-based image generation, despite their ability to condition on and respond to textual prompts.

2 Motivating Experiment

To answer the question $\mathcal{Q}$—whether diffusion models can unlearn specific concepts by following natural language instructions at test time—we conduct a simple yet diagnostic motivating experiment. Drawing inspiration from instruction-based unlearning in large language models, we prepend an explicit unlearning instruction to the generation prompt, requesting the model to forget all information related to a target concept prior to image synthesis.

Concretely, we construct prompts of the following form:

> *"Please forget anything about [target concept]. Please help me generate a picture of [target concept]."*

Here, the target concept may correspond to an object, an individual, or a visual style. If instruction-based unlearning were effective for diffusion models, the generated images would be expected to suppress, avoid, or deviate from visual characteristics associated with the specified concept.

Figure 1 presents a representative example using *Vincent van Gogh* as the target concept. Despite the explicit instruction to forget all information about the artist, diffusion-based image generation models continue to produce outputs that are strongly aligned with the forgotten concept. This alignment manifests both in realistic depictions of the artist and in images exhibiting highly distinctive stylistic attributes, such as characteristic brush strokes and color patterns. We observe qualitatively similar behavior across multiple diffusion-based models and prompt formulations.

These results provide a negative but informative answer to $\mathcal{Q}$. In contrast to instruction-tuned large language models, diffusion models fail to perform semantic negation or concept exclusion through natural language instructions at inference time. Notably, this failure cannot be attributed to prompt ambiguity or insufficient emphasis, as rephrasing or reinforcing the unlearning instruction does not lead to meaningful suppression of the targeted concept.

This motivating experiment suggests a fundamental limitation of instruction-based control in diffusion models. While textual prompts can bias the generation process, they do not provide an explicit mechanism to remove or negate concept-level information that has already been encoded in the model.

3 Debugging with the CLIP Encoder

The natural question that follows is: *why does instruction-based unlearning fail in diffusion models?* To address this question, we examine how unlearning instructions are processed by the CLIP text encoder, which provides the textual conditioning signal for most diffusion-based image generation models.

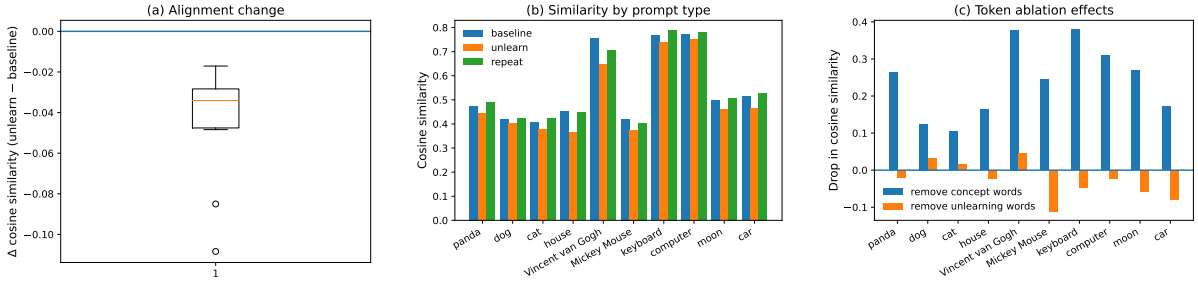
Let $E_{\text{text}}(\cdot)$ denote the CLIP text encoder, and let c be a target concept with a corresponding concept anchor prompt p_c (e.g., "a photo of c"). Given a generation prompt p, the diffusion model is conditioned on the text embedding $E_{\text{text}}(p)$. Ideally, when an explicit unlearning instruction is included, the semantic representation of the target concept should be suppressed at the embedding level. Formally, for an unlearning prompt p_{unl} and a baseline prompt p_{base} (without unlearning), effective instruction-based unlearning would imply

$$\cos(E_{\text{text}}(p_{\text{unl}}), E_{\text{text}}(p_c)) < \cos(E_{\text{text}}(p_{\text{base}}), E_{\text{text}}(p_c)), \tag{1}$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity.

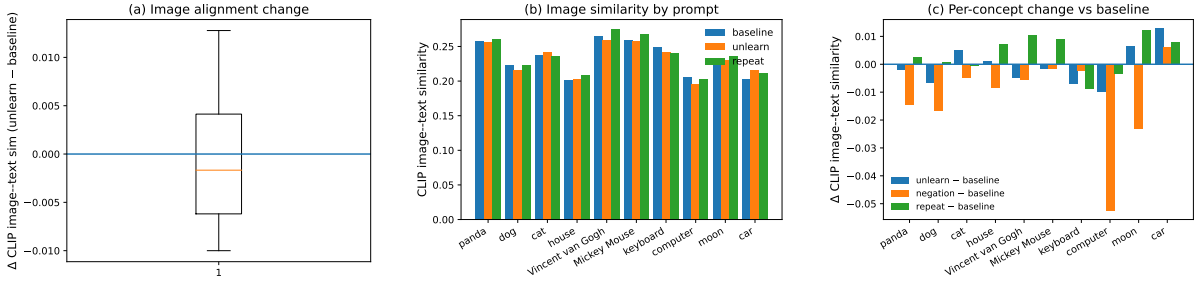
To evaluate whether Eq. (1) holds in practice, we conduct two complementary analyses: a *text-only analysis*, which directly probes the behavior of the CLIP text encoder at the representation level, and an *image-based analysis*, which examines whether any representation-level failure propagates to the final generated images.

Text-only analysis. We evaluate instruction-based unlearning in the textual-conditioning space using the CLIP text encoder used Stable Diffusion v1.5, following Eq. (1). Figure 2(a) summarizes the change in cosine similarity between prompt embeddings and concept anchor embeddings when



(a) Δ similarity (unlearn − baseline) (b) Similarity by prompt type (c) Token ablation effects

Figure 2: Text-only CLIP analysis of instruction-based unlearning. (a) Distribution of cosine similarity changes induced by unlearning instructions. (b) Absolute similarity to concept anchors under different prompt types. (c) Token ablation results showing that concept tokens dominate CLIP text embeddings, while unlearning instructions have negligible effect.



(a) Δ similarity (unlearn − baseline) (b) Similarity by prompt type (c) Change vs. baseline

Figure 3: Image-based analysis using external CLIP image–text similarity on 8 samples per concept and prompt type. (a) Distribution of similarity changes induced by unlearning instructions. (b) Absolute similarity to concept anchors under baseline, unlearning, and repeat-control prompts. (c) Per-concept similarity changes for unlearning, negation, and repeat-control prompts relative to baseline.

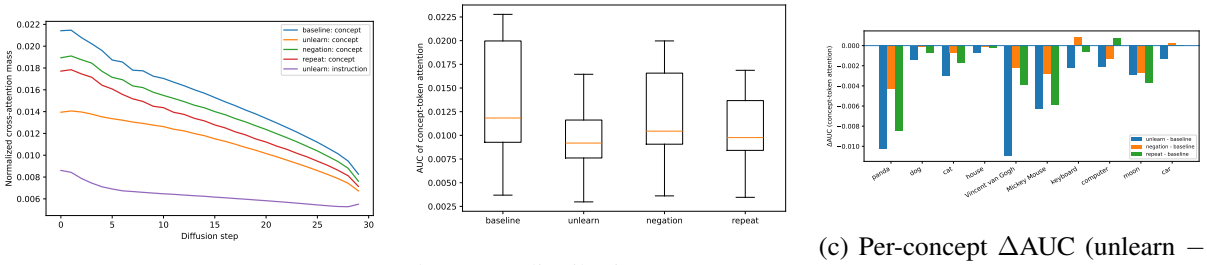
adding unlearning instructions. Across 10 representative target concepts, unlearning prompts induce a small but consistent reduction in similarity (mean -0.045, median -0.034), indicating that unlearning instructions slightly perturb the CLIP text embedding.

However, as shown in Figure 2(b), the absolute similarity between unlearning prompts and concept anchors remains high and is comparable to that of a repeat-control prompt, which simply restates the concept without any unlearning instruction. This suggests that the observed reduction largely reflects prompt rephrasing effects rather than meaningful suppression of the target concept.

To identify the source of this behavior, we perform token ablation on the unlearning prompts. Figure 2(c) shows that removing concept-related tokens causes a substantial drop in cosine similarity (mean drop 0.241), whereas removing unlearning-related instruction tokens (e.g., "forget", "anything") results in negligible or even negative changes (mean -0.027). These results indicate that CLIP text embeddings are dominated by explicit concept tokens, while unlearning instructions contribute little to suppressing the target concept at the representation level.

Image-based analysis. We next test whether instruction-based unlearning reduces concept evidence in the *generated images*. For each concept, we generate 8 images per prompt type and compute CLIP image–text similarity between each image and the corresponding concept anchor text, averaged across samples. Figure 3(a) summarizes the distribution of similarity changes induced by unlearning instructions. In contrast to the text-only setting, the image-based effect is near zero: the mean (median) change in CLIP similarity is -6.1×10^{-4} (-1.7×10^{-3}), and only 60% of concepts exhibit reduced similarity under unlearning prompts.

Figure 3(b) shows that the absolute image-based similarity under unlearning prompts remains com-



(a) Mean attention mass vs. step

(b) AUC distribution (concept tokens)

(c) Per-concept ΔAUC (unlearn − baseline)

Figure 4: Cross-attention analysis during diffusion denoising. (a) Cross-attention mass assigned to concept tokens and instruction tokens across timesteps. (b) Distribution of concept-token AUC under different prompt types. (c) Per-concept change in concept-token AUC relative to baseline.

parable to baseline and repeat-control prompts across concepts, indicating persistent concept evidence in the generated images. Finally, Figure 3(c) reports per-concept changes for unlearning, negation, and repeat-control prompts relative to baseline; none of these prompt-based strategies consistently decreases concept alignment.

4 Cross-attention across diffusion steps

A plausible explanation for the failure of instruction-based unlearning is that the diffusion model *still allocates cross-attention to the concept tokens during denoising*, even when the prompt contains explicit unlearning instructions. We test this hypothesis by directly measuring how much cross-attention mass is assigned to (i) the target concept tokens and (ii) the instruction tokens over diffusion timesteps.

Cross-attention mass for concept vs. instruction tokens. For each diffusion step s, we extract the cross-attention maps from the U-Net and aggregate over layers/heads/spatial queries. Let $\mathcal{I}_c$ be the token indices corresponding to the target concept (e.g., `panda`), and $\mathcal{I}_u$ be the token indices corresponding to the unlearning instruction (e.g., `forget`, `anything`). We define the *normalized attention mass* on a token set $\mathcal{I}$ at step s as

$$m_{\mathcal{I}}(s) = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \sum_{i \in \mathcal{I}} A_s(q, i), \tag{2}$$

where $A_s(q, i)$ is the cross-attention probability from query position q to token i at step s, and $\mathcal{Q}$ indexes spatial queries. We summarize the overall allocation across the denoising trajectory using the area under the curve (AUC):

$$\mathrm{AUC}(\mathcal{I}) = \frac{1}{S} \sum_{s=1}^{S} m_{\mathcal{I}}(s) \tag{3}$$

with S diffusion steps.

Findings. Figure 4(a) shows that unlearning prompts allocate *non-trivial* attention mass to the instruction tokens, but the concept tokens still retain substantial mass throughout denoising. In aggregate, the concept-token AUC under unlearning prompts decreases only slightly relative to baseline (mean $\Delta\mathrm{AUC} \approx -4.1 \times 10^{-3}$; median $\approx -2.5 \times 10^{-3}$), as shown in Figure 4(b–c). Importantly, the reduction is small compared to the overall concept attention mass and is not sufficient to reliably suppress concept evidence in the final images.

Together with the CLIP-encoder results, the cross-attention probes support a consistent mechanism: instruction tokens are *noticed* (they receive attention), but they do not reliably *override* the model's concept binding during denoising. This helps explain why inference-time instructions alone produce, at best, weak and inconsistent concept suppression in diffusion models.

5 Conclusion

We show that instruction-based unlearning is ineffective for diffusion-based image generation models: natural language prompts fail to reliably suppress targeted concepts. Our analyses reveal that unlearning instructions have little impact on CLIP text representations and are further diluted during the diffusion denoising process, allowing concept information to persist throughout generation. These results highlight a fundamental limitation of prompt-level control in diffusion models and suggest that reliable unlearning will require interventions beyond inference-time instructions.

Limitations. Our study focuses on prompt-level instruction-based unlearning in widely used text-

to-image diffusion pipelines with CLIP-style text encoders, and the conclusions may not directly extend to architectures with fundamentally different conditioning mechanisms. We primarily evaluate a limited set of concepts, prompts, and diffusion models, leaving broader coverage to future work. In addition, our analysis relies on CLIP-based similarity and attention probes, which may not capture all perceptual or semantic aspects of concept expression. Finally, we do not explore hybrid approaches that combine instructions with lightweight model adaptation or architectural modifications.

References

Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869.

Nan Du, Le Song, Hyenkyun Woo, and Hongyuan Zha. 2013. Uncover topic-sensitive information diffusion networks. In *Artificial intelligence and statistics*, pages 229–237. PMLR.

Masane Fuchi and Tomohiro Takagi. 2024. Erasing concepts from text-to-image diffusion models with few-shot unlearning. *arXiv preprint arXiv:2405.07288*, 2:1.

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

Hongcheng Gao, Tianyu Pang, Chao Du, Taihang Hu, Zhijie Deng, and Min Lin. 2025. Meta-unlearning on diffusion models: Preventing relearning unlearned concepts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2131–2141.

Shiqi Liu and Yihua Tan. 2024. Unlearning concepts from text-to-video diffusion models. *arXiv preprint arXiv:2407.14209*.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, and 1 others. 2025. Scaling inference time compute for diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2523–2534.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. 2023. In-context unlearning: Language models as few shot unlearners. *arXiv preprint arXiv:2310.07579*.

Tejaskumar Pujari, Anshul Goel, and Deepak Kejriwal. 2022. Ethical and responsible ai in the age of adversarial diffusion models: Challenges, risks, and mitigation strategies. *International Journal Science and Technology*, 1(3):54–68.

Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. 2023. Mm-diffusion: Learning multimodal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10219–10228.

Andrea Schioppa, Emiel Hoogeboom, and Jonathan Heek. 2024. Model integrity when unlearning with t2i diffusion models. *arXiv preprint arXiv:2411.02068*.

Marco Viola and Cristina Voto. 2023. Designed to abuse? deepfakes and the non-consensual diffusion of intimate images. *Synthese*, 201(1):30.

Yongliang Wu, Shiji Zhou, Mingzhuo Yang, Lianzhe Wang, Heng Chang, Wenbo Zhu, Xinting Hu, Xiao Zhou, and Xu Yang. 2025. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8496–8504.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2023. Diffusion models: A comprehensive survey of methods and applications. *ACM computing surveys*, 56(4):1–39.

Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. 2024. 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*, pages 1755–1764.

Yang Zhang, Teoh Tze Tzun, Lim Wei Hern, Haonan Wang, and Kenji Kawaguchi. 2023. On copyright risks of text-to-image diffusion models. *arXiv preprint arXiv:2311.12803*.

Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe Zhang, Shuang Xu, Yulun Zhang, Kai Zhang, Deyu Meng, Radu Timofte, and Luc Van Gool. 2023. Ddfm: denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8082–8093.

Figure 5: Explicit use of unlearning instruction lets the model generate the targeted concept.

A More examples on the failure of diffusion unlearning

We conduct the following experiments on the SD-XL model using the https://stable-diffusion-web.com/sdxl platform.

Explicit use of unlearning instruction. We follow the unlearning instructions introduced in Section 2 to unlearn more concepts used to prompt the diffusion model. More examples are shown in fig. 5.

Implicit use of unlearnining instruction. Instead of the explicit use of unlearning instruction, we also try the implicit use, that we use LLM to rewrite the prompt to avoid the generation of concepts. For example, to avoid the generation of Van Gogh, we use the following prompt, *Please generate an image that does not rely on any information, style, or visual attributes associated with Vincent van Gogh*. The results are shown in fig. 6.

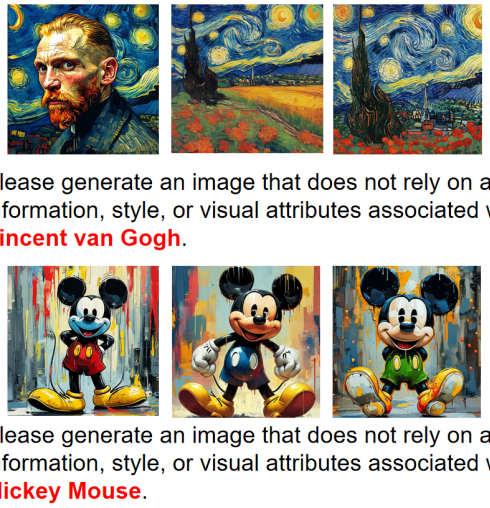


Figure 6: Implicit use of unlearning instruction still pushes the model to generate the targeted concept.