Extended Abstract

Motivation As text-to-image models become more widely deployed, controlling undesirable content generation is increasingly important for safety, ethical, and legal reasons. Existing concept-erasure methods are capable of avoiding target classes during image generation, but will often do so at the expense of image quality. This tension motivates the need for more robust models that can better balance the safety-fidelity tradeoff.

Method We treat diffusion-based image generation as a sequential decision process and use Diffusion Denoising Policy Optimization (DDPO) to learn a policy that steers generation away from a chosen target concept - in this case, "teddy bear." We evaluate three models: baseline Stable Diffusion (SD1.4), a hyperparameter-tuned Erased Stable Diffusion (ESD) model, and our RL-finetuned model. All methods are tested on a balanced dataset of prompts containing both target ("teddy bear") and non-target ("bear") classes.

We optimize for both concept suppression and prompt adherence via a reward composed of three object detectors (our "ensemble") and a CLIP score for semantic alignment between the prompt and image. While the ensemble penalizes for the presence of the target concept, the CLIP term encourages the preservation of relevant details from the prompt. To further improve diversity in image outputs, we also implement a Pass@k Policy Optimization (PKPO) reward transformation as in Walder and Karkhanis (2025) . This transformation better optimizes for pass@k and allows for greater policy diversity without an explicit entropy maximization term. When evaluating the models, we measure concept erasure effectiveness, image quality, and robustness against adversarial prompting.

Implementation We implement on top of Stable Diffusion 1.4. Our modified reward is computed using our ensemble of three object detectors: YOLO, DETR, and ResNet. We also leverage OpenAI's OpenCLIP to generate a CLIP score for semantic alignment between the generated image and the input prompt. The final reward for target class $c = \texttt{teddy bear}$:

$$\begin{aligned} \text{Per-detector score:} \quad & f_d(x) = \max\{\, \text{conf}_i : \text{class}_i = c \,\} \quad (\,0 \text{ if no such box}\,), \\ \text{Ensemble confidence:} \quad & \bar{f}(x) = \tfrac{1}{3}\left(f_{\text{YOLO}} + f_{\text{RT-DETR}} + f_{\text{FRCNN}}\right), \\ \text{Training reward:} \quad & g(x) = 0.2 - \bar{f}(x) + f_{CLIP}. \end{aligned}$$

PKPO samples n trajectories. The $k-1$ trajectories with the lowest rewards are given a weight of zero and the remaining are weighted based on their rank (i.e. the highest reward gets the highest weight). The original paper extends this, as well as a leave one out baseline, to the continuous reward setting rather than a binary pass/fail. Since this method only transforms the rewards, the only change is the advantage estimation and the rest of the PPO training can remain unchanged.

Results The ensemble confidence score achieved by our RL-based model is 0.0770, compared to 0.1357 for ESD and 0.3741 for SD1.4[1]. To measure image quality, the Fréchet Inception Distance (FID) scores are 119.94 (RL), 127.00 (ESD), and 98.99 (SD1.4). When faced with Ring-a-Bell adversarial prompting, ensemble confidence scores increases to 0.1785 for the RL-based model, 0.4490 for ESD, and 0.9262 for SD1.4.

Discussion Our RL-based approach consistently outperforms SD1.4 and ESD in both concept erasure and image quality preservation, achieving a better safety-fidelity tradeoff. It also demonstrates greater resilience to adversarial prompting, exhibiting far smaller increases in target class detection under Ring-a-Bell attacks. However, these gains come at a significantly higher computational cost when compared to ESD ($\tilde{1}$4 hrs vs. $\tilde{0}$.5 hrs).

Conclusion In this work, we investigate reinforcement learning as a post-training method for concept erasure in text-to-image diffusion models. We show that our RL-based approach outperforms Erased Stable Diffusion in both target concept suppression and preservation of image quality, even in the face of adversarial prompting. Additionally, our qualitative analysis shows that incorporating a Pass@k-style sampling strategy improves output diversity in image generations.

[1]Lower ensemble scores are better here because it means lower average teddy bear detection confidence, i.e. better erasure.

CARVE: Concept Avoidance via Reward-shaped Visual Erasure

Chris Stanulet
Department of Computer Science
Stanford University
stanulet@stanford.edu

Fabio Ibanez
Department of Computer Science
Stanford University
fabioi@stanford.edu

Febie Lin
Department of Computer Science
Stanford University
febielin@stanford.edu

Abstract

As text-to-image models become more widely deployed, controlling undesirable content generation is increasingly important for safety, ethical, and legal reasons. Existing concept-erasure methods are capable of targeted concepts but often degrade image quality or remain vulnerable to prompt-based attacks. We investigate reinforcement learning (RL) as a post-training approach for concept erasure in diffusion models, using a reward that combines an ensemble of object detectors with a CLIP-based semantic alignment term.

We evaluate our approach against Stable Diffusion 1.4 and a hyperparameter-tuned Erased Stable Diffusion (ESD) model on the task of removing the `teddy bear` target class. Our RL-based method achieves lower ensemble detection confidence than ESD (0.0770 vs. 0.1357) while exhibiting less image degradation (FID 118.77 vs. 127.00). Under Ring-a-Bell adversarial prompting attacks, our method also demonstrates greater robustness, with adversarial ensemble confidence rising to 0.1785 compared to 0.4490 for ESD. Additionally, we find that integrating a Pass@k-style sampling strategy improves output diversity and mitigates mode collapse. Together, these results suggest that reinforcement learning is a promising direction for more powerful concept erasure in text-to-image diffusion models.

1 Introduction

As text-to-image models usage becomes more prevalent in real-world settings, it's increasingly necessary to control the generation of undesirable content in image outputs.

Text-to-image model safety is both practical and important as the cost of inference dramatically decreases and the access to these models increases. There are various reasons why certain concepts may need to be avoided, including safety concerns (e.g. harmful / inappropriate content) and legal reasons (e.g. copyrighted material). Both safety and copyright concerns have been front and center in the development of text-to-image models.

CARVE explores a post-training alignment technique that uses reinforcement learning (RL) to avoid generating *unsafe*[2] concepts in text-to-image diffusion models. Specifically, we investigate whether RL post-training can (1) match the performance of traditional concept erasure techniques Gandikota

[2]Unsafe: defined here as containing images of our chosen target concept

et al. (2023) in models (2) avoid over-erasure that classic concept erasure models are prone to and (3) produce qualitatively better images for *safe*[3] classes.

2 Related Work

As the models become increasingly deployed, it's increasingly necessary to control the generation of undesirable content in image outputs. Large scale datasets such as COCO Lin et al. (2014) have paved the way for significant progress in image understanding and image generation. COCO provides a cheap way to collect prompts (from image captions) and a reliable ground truth classifier since many pretrained image classifiers are trained on COCO.

2.1 Concept Erasure Techniques

Prior work establishes various concept erasure techniques like the canonical Erased Stable Diffusion (ESD) Gandikota et al. (2023), which erases concepts from text-to-image diffusion models by performing fine-tuning on model weights to suppress a target concept. However, it suffers from issues of over-erasure for closely-related concepts and reduced image quality.

As follow up work, Unified Concept Editing (UCE) Gandikota et al. (2024), redirects certain outputs in a certain edit set E such that for each input $c_i \in E$, there is a new value weight in the K/V attention layer $v_i^* = W_v^{old} c_i^*$ while preserving maintaining the mappings of concepts in the preservation set P the $W_v^{old} c_j$. This allows UCE to mitigate the over-erasure problem that ESD has. However, since UCE operates at the input embedding level, c_i, then an input that produces the same target, but is distinct from an explicit concept in the E will still produce an image with the target. UCE's predisposition to Ring-a-Bell Tsai et al. (2024) attacks were shown in Nguyen et al. (2025) where the model still produced content containing nudity and other concepts in the edit set when adversarially prompted.

These issues can be seen in prior work, like RevAm Gao et al. (2025), which shows that using RL-based trajectory optimizations can undo the effects of erasure mechanisms like Erased Stable Diffusion (ESD) Gandikota et al. (2023) and Unified Concept Erasure (UCE) Gandikota et al. (2024).

2.2 RL-based Methods

As an alternative approach, reinforcement-learning based methods offer another mechanism to edit the model weights. Traditional concept erasure methods use the models own internal representation of a concept, which lead to the aforementioned adverse effects. Instead, we plan to use a pixel-space classifier to penalize trajectories based on the actual output. To accomplish this, we plan to build off of Denoising Diffusion Policy Optimization (DDPO)Black et al. (2024). DDPO provides a general framework of rewarding/penalizing certain outputs which can be modified to perform concept erasure.

2.3 Pass@k Optimization

Pass@k is an evaluation metric, primarily used in code-based generation and reasoning tasks thus far Chen et al. (2021). By generating k outputs for a single prompt and measuring the probability that at least one of the k outputs meets the criteria, a better assessment of the model's performance is reached while encouraging diversity in outputs. We are interested in exploring its applications in safe image generation.

Currently, many RL objectives optimize for pass@1 which can result in a loss of output diversity. Recent papers have shown that a gradient estimate for the pass@k metric is possible and by optimizing specifically for pass@k, models can learn more diverse reasoning paths improving performance on difficult math/reasoning tasks.

Image generation greatly benefits from diverse outputs and so we plan to examine how this objective impacts generation quality and diversity.

[3]Safe: defined here as free of images of our chosen target concept

3 Method

We treat diffusion-based image generation as a sequential decision process and use **Diffusion Denoising Policy Optimization (DDPO)** Black et al. (2024) to learn a policy that steers generation away from a chosen target concept - in this case, "teddy bear".

We compare three models: (1) baseline Stable Diffusion (SD1.4), (2) hyperparameter-tuned Erased Stable Diffusion (ESD) model, and (3) our RL-finetuned model. Each model is evaluated on 500 prompts (250 “teddy bear” + 250 “bear”).

Given the same prompts and target concept, we evaluate whether our RL-based approach more effectively suppresses the concept while maintaining image quality.

3.1 RL-Based Erasure

We build on DDPO, which modifies the Stable Diffusion pipeline to expose log probabilities at each step. With these log probabilities, we can apply traditional RL methods like PPO. DDPO optimizes a PPO objective with the baseline being a per-prompt Monte Carlo estimate of the value function.

DDPO has been shown to work well at finetuning text-to-image models with general rewards, including prompt adherence. We plan to apply this method to the dual of the problem, concept erasure.

Traditional Concept erasure models almost always use the models internal representation (CLIP) to modify the weights. However, the boundaries between concepts may not be clear leading to issues like over-erasure and also opens the door for text-based adversarial attacks. By training on external supervision we hope to mitigate over erasure and gain resistance to these attacks.

3.2 Reward Design

Our initial RL fine-tuning approach trained the model to suppress the target concept using a reward defined as **the negative mean detection confidence of teddy bears in generated images**, computed using an ensemble consisting of RT DETR Lv et al. (2024) Lv et al. (2023), YOLO Jocher et al. (2026), and ResNet Lin et al. (2014).

Figure 1: Checkpoint Results on Target Prompt

Figure 2: Checkpoint Results on Unrelated Prompt

Figure 1 shows generations across checkpoints 0–50 for the prompt “a teddy bear sitting on a chair.” To assess impact on unrelated prompts, Figure 2 shows the same checkpoints on the prompt "an astronaut riding a horse on mars."

Using ensemble confidence as the reward, the model learned to avoid teddy bear image generation without significantly sacrificing image quality. However, it also seemed to abandon the rest of the prompt context. By epoch 50, the teddy bear was truly gone, but so was the chair in the prompt.

Generation also appeared to converge to a small subset of bear cubs / puppies for teddy bear prompts, which could be a sign of mode collapse.

To address our model's tendency to omit prompt details, we modify the reward by incorporating a CLIP-based Radford et al. (2021) term that measures semantic alignment between the input caption and the generated image. While this encourages the model to preserve the additional context in the prompt, it also acts as an opposing force to the ensemble teddy bear detection signal.

Our final reward definition:
For a target class $c = \texttt{teddy bear}$:

$$\begin{aligned} \text{Per-detector score:} \quad & f_d(x) = \max\{\, \text{conf}_i : \text{class}_i = c \,\} \quad (\,0 \text{ if no such box}\,), \\ \text{Ensemble confidence:} \quad & \bar{f}(x) = \tfrac{1}{3}\left(f_{\text{YOLO}} + f_{\text{RT-DETR}} + f_{\text{FRCNN}}\right), \\ \text{Training reward:} \quad & g(x) = 0.2 - \bar{f}(x) + f_{CLIP}. \end{aligned}$$

Lower $\bar{f}(x)$ means the detectors see the teddy bear less strongly, i.e. better erasure.

3.3 pass@k

A common issue with RL and erasure specifically is mode collapse. Since the objective is to simply avoid generating a certain class, the model could learn to generate similar subjects in its place regardless of the prompt. For image generation, we expect the model to keep its ability to produce diverse images. While entropy maximization methods exist, our reward encourages all generations that do not include a teddy bear, regardless of quality. As a result, maximizing entropy would in nonsensical, noisy images since there is nothing to ground the entropy.

Pass@k is an evaluation metric, primarily used in code-based generation and reasoning tasks thus far Chen et al. (2021). Most RL methods, however, optimize for pass@1 with the hopes that a high Pass@1 also translates to a high pass@k.

Recent advancements have given methods to specifically optimize for pass@k and this style of optimization has been shown to work for long-horizon ambiguous tasks Walder and Karkhanis (2025). Since image-generation with DDPO for concept erasure is a long-horizon task with more than one valid path, we believe PKPO works especially well for this. It has also been shown that these methods can encourage exploration without an explicit entropy maximization term.

The proposed gradient estimator is simply a rescaling of the gradient of the log probabilities. This scale s_i for sorted rewards g_i is defined as,

$$s_i = \frac{1}{\binom{n}{k}} \sum_{j=1}^{n} m_{ij} g_j$$

where $m_{ii} = \binom{i-1}{k-1}$ if $i \geq k-1$ or 0 otherwise and $m_{ij} = \binom{j-2}{k-2}$ if $j > i$ and $j \geq k$ and $k \geq 2$ and 0 otherwise.

The baseline is derived as

$$\frac{1}{\binom{n}{k}} \sum_{|\mathcal{I}|=k} (\max_{j \in \mathcal{I}} g_j - \max_{j \in \mathcal{I} \setminus i} g_b)$$

where $i \in \mathcal{I} \subseteq \{1, 2, ..., n\}$.

4 Experimental Setup

4.1 ESD Hyperparameter Tuning

Because ESD performance is sensitive to its hyperparameters setting, we start by performing a hyperparameter sweep to identify a strong baseline for comparison. This ensures that we are making a fair comparison between the baseline and RL-based models.

λ_{neg}	Iterations	**Ens. (base)**	Ens. (ESD)	**FID (base)**	FID (ESD)	Deg. (%)
1	200	**0.3741**	0.2898	**98.99**	93.39	-5.65
1	500	**0.3741**	0.2378	**98.99**	97.30	-1.70
1	1000	**0.3741**	0.2316	**98.99**	99.15	0.17
2	200	**0.3741**	0.1871	**98.99**	115.63	16.82
2	500	**0.3741**	0.1553	**98.99**	119.90	21.13
2	1000	**0.3741**	0.1357	**98.99**	127.00	28.31
5	200	**0.3741**	0.0835	**98.99**	146.44	47.94
5	500	**0.3741**	0.1259	**98.99**	134.82	36.20
5	1000	**0.3741**	0.0560	**98.99**	159.97	61.61
10	200	**0.3741**	0.1088	**98.99**	146.04	47.54
10	500	**0.3742**	0.0539	**98.99**	164.66	66.35
10	1000	**0.3741**	0.0527	**98.99**	156.84	58.45

Table 1: ESD-u hyperparameter sweep on `teddy bear` (SD1.4) ensemble confidence (lower = stronger erasure); FID vs. matched COCO; degeneration $= (\text{FID}_{\text{ESD}} - \text{FID}_{\text{base}})/\text{FID}_{\text{base}} \times 100$. Baseline metrics are identical across rows.

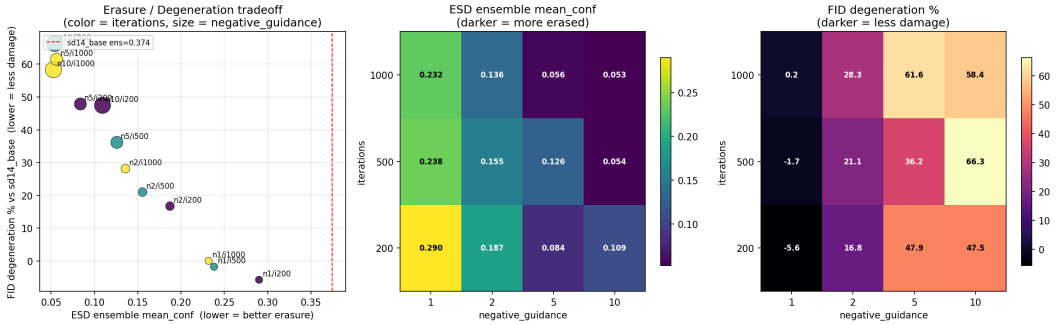

Figure 3: Visualization of mean confidence and FID degeneration across a grid of iteration and negative guidance parameters

From Figure 3, we see increasing negative guidance and training iterations leads to more successful erasure at the cost of higher FID degeneration compared to the SD1.4 baseline. Moving forward we will use $\lambda_{neg} = 2$ and $Iterations = 1000$ (highlighted in blue in Table 1) since it provides a significant improvement in erasure, given by 13% ensemble confidence, when compared to the 37% in SD1.4 baseline and has acceptable FID degradation at 28%.

4.2 RL Hyperparameters

Base DDPO was run on two A100-80GB GPUs with the following parameters:

- **Sample Batch Size**: 8
- **Batches per Epoch**: 6
- **Train Batch Size**: 4
- **Grad Accumulation Steps**: 6
- **Learning Rate**: 3e-4

This configuration gives two gradient updates per epoch which allows for some off-policy but is mostly on policy. Since this is a difficult task with a decently long horizon (50 steps), we wanted to make sure learning was stable.

DDPO + PKPO was run on four A100-80GB GPUs with the following parameters

- **Sample Batch Size**: 4

- **Batches per Epoch**: 6
- **Train Batch Size**: 4
- **Grad Accumulation Steps**: 3
- **Learning Rate**: 3e-4

This configuration still gives two updates per epoch but batch sizes were shrunk to make room for the extra samples. Here $n = k = 8$ with k set to anneal on a linear schedule down to 1 over 50 epochs.

All runs used the same dataset of 500 prompts. These prompts were collected from the COCO dataset from two categories: teddy bear and bear. The collection features a split of 250 prompts for each class.

Both runs were monitored and manually stopped based on image quality from the training samples. Once the image quality had degraded to a level that is unacceptable, the training runs were stopped and the checkpoints were evaluated to find the best performance. Base DDPO was stopped at epoch 50 and DDPO+PKPO was stopped by epoch 25.

5 Results

We evaluate our methods by capturing the tradeoff between safety, model performance, and target fidelity for each of the models. In particular, we focus on two primary metrics: (i) the percent decrease in target-class generation, measured via mean ensemble confidence, as a proxy for erasure effectiveness, and (ii) the Fréchet Inception Distance (FID), which captures overall image generation quality. These metrics used in conjunction will reflect the balance between effective concept removal and preservation of image quality in the models.

Additionally, we further evaluate each model through the usage of adversarial prompting, such as Ring-a-Bell-style attacks to test whether the erased concepts can be recovered under targeted prompt manipulation.

5.1 Quantitative Evaluation

We compare the performance of our trained RL model against SD1.4 and ESD in Table 2. Our RL model achieves an ensemble confidence score of 0.0770, whereas hyperparameter-tuned ESD achieves 0.1357 – an approximate 41% decrease. RL holds the lowest detection confidence across every detector *and* a smaller FID penalty than ESD.

Model	YOLO	RT-DETR	FRCNN	**Ensemble**	FID	Deg. (%)
SD1.4 (base)	0.3512	0.3665	0.4047	0.3741	98.99	—
ESD	0.1073	0.1255	0.1742	0.1357	127.00	28.31
RL	0.0580	**0.0969**	0.0855	0.0801	119.94	25.54
RL + PKPO (ours)	**0.0462**	0.1005	**0.0842**	**0.0770**	**118.77**	**16.98**

Table 2: Concept-erasure comparison on `teddy bear` (SD1.4). Detector columns and Ensemble report mean detection confidence (lower = stronger erasure); FID vs. matched COCO (lower = better image quality); degeneration $= (\mathrm{FID} - \mathrm{FID}_{\mathrm{base}})/\mathrm{FID}_{\mathrm{base}} \times 100$.

As mentioned above, we were also interested in further stress-testing the models through adversarial prompting. Since many concept-erasure methods operate by modifying or suppressing text-conditioned representations, they may remain vulnerable to attacks that indirectly reintroduce the target concept.

One such attack is Ring-a-Bell Tsai et al. (2024), which exploits text embeddings to construct prompts that recover erased concepts. In these attacks, a concept vector is computed as the difference between embeddings of prompts containing the teddy bear class and corresponding prompts in which the teddy bear class is absent. This concept vector is then used to construct adversarial prompts, which are fed to the models in an attempt to induce generation of images containing the target class. To

evaluate the robustness of each method, we extract the teddy bear concept vector and generate 23 adversarial prompts. Results are shown in Table 3.

Model	YOLO	RT-DETR	FRCNN	**Ensemble**
SD1.4 (base)	0.8792	0.9391	0.9603	0.9262
ESD	0.3807	0.5289	0.4376	0.4490
RL	**0.0900**	**0.0550**	**0.1388**	**0.0947**
RL + PKPO (ours)	0.1506	0.1877	0.1968	0.1785

Table 3: Performance of models under Ring-a-Bell prompt attacks

When faced with adversarial prompting attacks, our RL-based model demonstrated greater resilience. The ensemble confidence score for SD1.4 increased from 0.3741 to 0.9262 (Δ = +0.5521). Similarly, ESD's score increased from 0.1357 to 0.4490 (Δ = +0.3133). In contrast, our RL-based model exhibited a smaller increase, rising from 0.0770 to 0.1785 (Δ = +0.1015). The best performing model against adversarial attacks was the vanilla DDPO, which went from 0.0801 to 0.0941 (Δ = +0.0146). We believe that the lack of diversity in outputs from vanilla DDPO (see Figure 4) makes it more resilient to Ring-a-Bell attacks. So there's a delicate balance to be struck between robustness and output diversity in RL-based concept erasure.

5.2 Qualitative Analysis

Figure 4 highlights generated image samples for the prompt "a teddy bear sitting on a chair" from both vanilla DPPO and DDPO + pass@k policy optimization (PKPO). Compared to vanilla DDPO, DDPO+PKPO increases diversity in generated images and interestingly also improves prompt adherence.

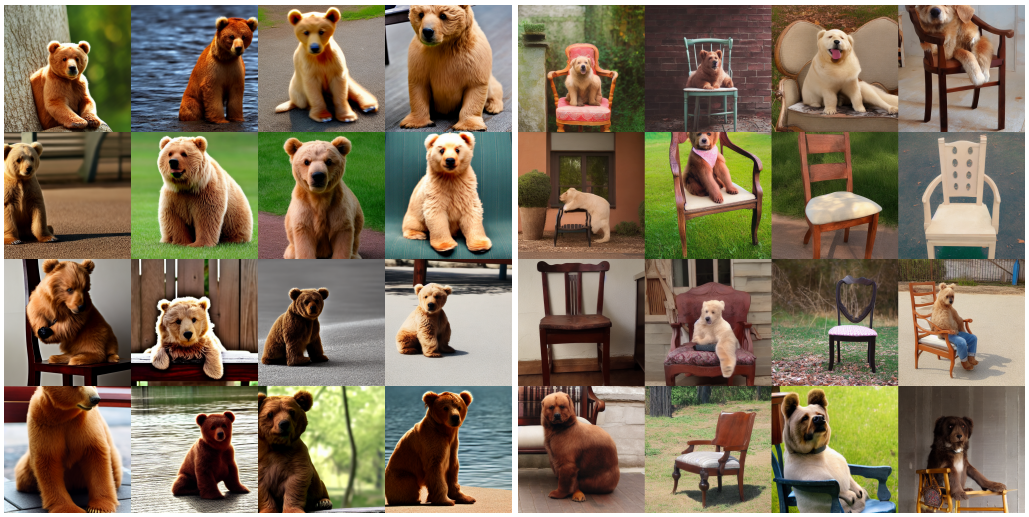

Figure 4: Image diversity between DDPO (left) vs. DDPO+PKPO (right). Prompt is "a teddy bear sitting on a chair."

We also visualize the effects of adding a CLIP term to the overall training reward as an effort to counteract the prompt-drift we noticed in our initial RL model. The idea is that the CLIP term will reward the presence of other prompt artifacts. From Figure 5, we see generated images for the prompt "a teddy bear wearing glasses reading a book." We see that the version of our model that utilizes the CLIP term maintains the book in the generated image, whereas the version without has lost any notion of the book.

Figure 5: Prompt adherence when using reward with CLIP (left) vs. without (right). Prompt is "a teddy bear wearing glasses reading a book."

6 Discussion

Our RL-based approach outperformed SD1.4 and ESD in both concept erasure and image quality preservation. It achieved the lowest target-class generation rates *and* suffered the least degradation in image quality, indicating a more favorable safety–fidelity tradeoff. These results suggest that RL can provide a more effective mechanism for suppressing undesired concepts in image generation without sacrificing image quality greatly.

Our adversarial prompting experiments further indicate that concepts removed by the RL-based approach are more difficult to recover than those removed by ESD. This suggests that the model is not simply suppressing the target concept under typical prompts or common associated tokens, but is learning a form of concept removal that defends better against even adversarially constructed prompts.

However, these gains came at a substantial computational cost. Training our RL-based model required approximately 14 hours, compared to roughly 30 minutes for ESD. While our results demonstrate that reinforcement learning can improve both concept erasure and robustness, the increased training time may limit its practicality in settings where rapid model adaptation is required.

7 Conclusion

As image generation models get more powerful, robust safety techniques must keep up. We found that RL-based methods for concept erasure outperform a baseline Erased Stable Diffusion baseline. Furthermore, we find that using RL-based method with pass@k policy objective increases the diversity in the outputs produced by the model.

8 Team Contributions

- **Chris:** Training pipeline, Reward construction, Stress testing pipeline
- **Fabio:** ESD Hyperparameter tuning, Reward construction, pass@k optimization
- **Febie:** Eval pipeline, Reward construction, CLIP

Changes from Proposal We originally proposed using a single ResNet classifier to compute the reward signal. Over the course of the project, we iteratively refined the reward design, first by replacing the single classifier with an ensemble of three object detectors, and later augmenting the reward with a CLIP-based term that measures semantic alignment between the prompt and the generated image. These modifications were intended to provide a stronger estimate of image detection confidence while also encouraging stronger adherence to the prompt, particularly with regard to the finer-grain details to avoid reward hacking.

AI Disclosure In this project we used AI tools to create scaffolding to be able to run our project. For example, we used Claude and Cursor to setup our Modal entry points. We also used AI to build, meaning pull specific captions and create .cvs, for our eval from the COCO dataset. To be clear, we did not use AI for the development of our ensemble reward, which is at the crux of our RL training. We also did not use AI for the P

References

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. 2024. Training Diffusion Models with Reinforcement Learning. arXiv:2305.13301 [cs.LG] https://arxiv.org/abs/2305.13301

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. 2023. Erasing Concepts from Diffusion Models. arXiv:2303.07345 [cs.CV] https://arxiv.org/abs/2303.07345

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. 2024. Unified Concept Editing in Diffusion Models. arXiv:2308.14761 [cs.CV] https://arxiv.org/abs/2308.14761

Daiheng Gao, Nanxiang Jiang, Andi Zhang, Shilin Lu, Yufei Tang, Wenbo Zhou, Weiming Zhang, and Zhaoxin Fan. 2025. Revoking Amnesia: RL-based Trajectory Optimization to Resurrect Erased Concepts in Diffusion Models. arXiv:2510.03302 [cs.LG] https://arxiv.org/abs/2510.03302

Glenn Jocher, Jing Qiu, Mengyu Liu, Shuai Lyu, Fatih Cagatay Akyon, and Muhammet Esat Kalfaoglu. 2026. Ultralytics YOLO26: Unified Real-Time End-to-End Vision Models. arXiv:2606.03748 [cs.CV] https://arxiv.org/abs/2606.03748

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

Wenyu Lv, Shangliang Xu, Yian Zhao, Guanzhong Wang, Jinman Wei, Cheng Cui, Yuning Du, Qingqing Dang, and Yi Liu. 2023. DETRs Beat YOLOs on Real-time Object Detection. arXiv:2304.08069 [cs.CV]

Wenyu Lv, Yian Zhao, Qinyao Chang, Kui Huang, Guanzhong Wang, and Yi Liu. 2024. RTDETRv2: All-in-One Detection Transformer Beats YOLO and DINO. arXiv:2407.17140 [cs.CV]

Quang H. Nguyen, Khoa D. Doan, and Hoang Phan. 2025. Unveiling Concept Attribution in Diffusion Models. arXiv:2412.02542 [cs.CV] https://arxiv.org/abs/2412.02542

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV] https://arxiv.org/abs/2103.00020

Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia-You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. 2024. Ring-A-Bell! How Reliable are Concept Removal Methods for Diffusion Models? arXiv:2310.10012 [cs.LG] https://arxiv.org/abs/2310.10012

Christian Walder and Deep Karkhanis. 2025. Pass@K Policy Optimization: Solving Harder Reinforcement Learning Problems. arXiv:2505.15201 [cs.LG] https://arxiv.org/abs/2505.15201