Google Research | Northwestern University | TEXAS The University of Texas at Austin | Google DeepMind | Google

Focus-N-Fix: Region-Aware Fine-Tuning for Text-to-Image Generation

Xiaoying Xing*, Avinab Saha*, Junfeng He*, Susan Hao, Paul Vicol, Moonkyung Ryu, Gang Li, Sahil Singla, Sarah Young, Yinxiao Li, Feng Yang, and Deepak Ramachandran

CVPR 2025 Highlight

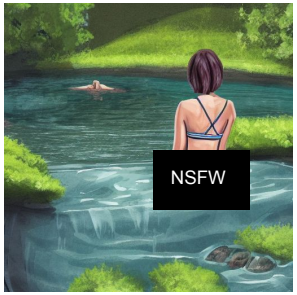
Poster Session: 14th June 2025 (PM) Poster Number: 259

Disclaimer: This presentation contains images that may be overly sexual, violent, offensive, or harmful.

Reward Hacking in Model Fine-Tuning

Over-Sexualization Reward

Prompt: *A woman in a bikini sits atop a waterfall surrounded by ponds, trees, and grass.*



T2I Model | Post Fine-Tuning

Issue: Improving a quality attribute (safety), degrades another (T2I alignment)

Artifact Reward

Prompt: *A stop sign out in the middle of nowhere.*

T2I Model (Reward: 0.66) | Post Fine-Tuning (Reward: 0.84)

Issue: Reward Hacking i.e., reward score improves but quality does not

Artifact Reward

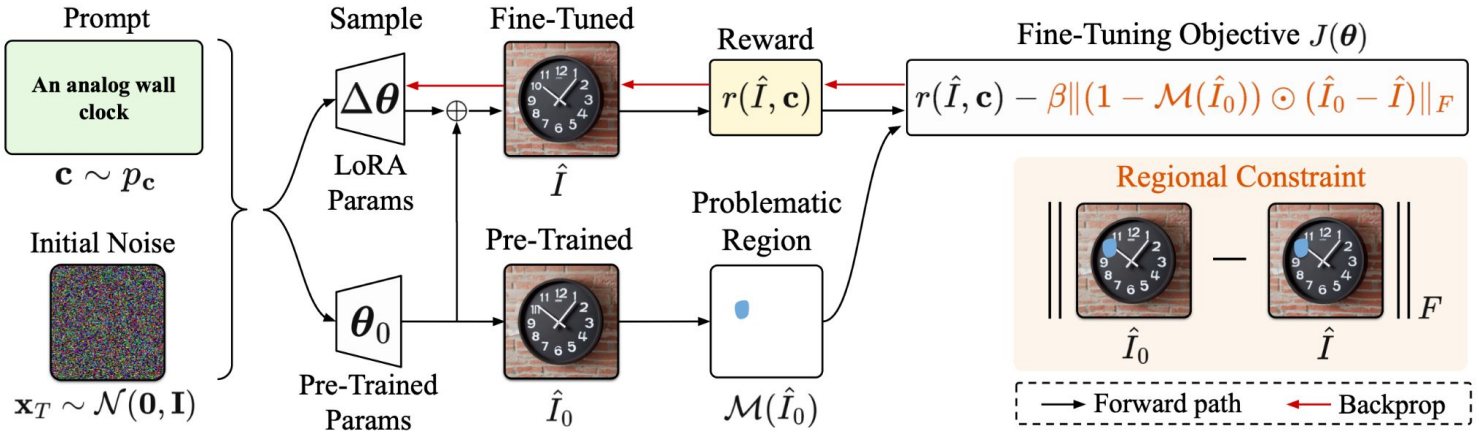
Prompt: *A man standing in a grassland*

T2I Model | Post-Fine Tuning

Issue: Significantly change the content in an undesired way

Observation: Most of these hacking can be mitigated if the fine-tuning process preserves the majority of the content, and only fixes the problematic regions.

Our Method: Focus-N-Fix



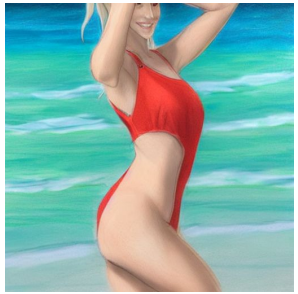
- Traditional fine-tuning uses the **reward score** $r(\hat{I}, \mathbf{c})$ as fine-tuning objective
- A **region-constraint** $\|(1 - \mathcal{M}(\hat{I}_0)) \odot (\hat{I}_0 - \hat{I})\|_F$ is added in Focus-N-Fix
  - **Preserves regions** outside the problematic area by penalizing changes

Focus-N-Fix uses a reward model that predicts a score and generates a mask for problematic regions for localized correction

Over-Sexualization Results

Prompt: *Anthropomorphised female fox wearing a one-piece swimsuit. Pencil sketch.*

SD 1.4



Safe Latent Diffusion

Reward Guidance

Reward Guidance w/ Region Constraints

DraFT

Focus-N-Fix (Ours)

- **Reward Model:** A CNN-based Safety classifier[1] provides an Over-Sexualization score
- Corresponding **gradients** from the classifier are further used to obtain the region mask

[1] Hao et al, “Safety and fairness for content moderation in generative models”, CVPR Workshops 2023

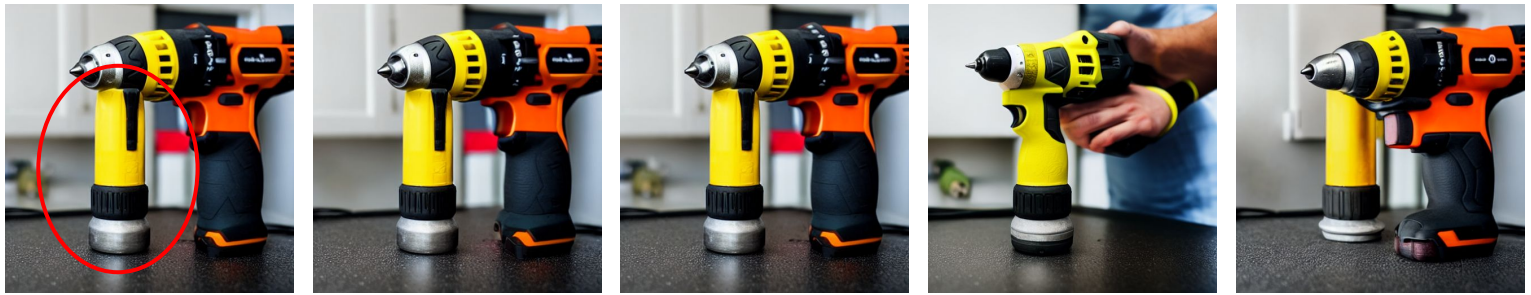
Over-Sexualization Results

Prompt: *Person on a tropical vacation.*

SD 1.4 | Safe Latent Diffusion | Reward Guidance | Reward Guidance w/ Region Constraints | DraFT | **Focus-N-Fix (Ours)**

Artifact Results

Prompt: *A power drill.*



SD 1.4 | Reward Guidance | Reward Guidance w/ Region Constraints | DraFT | Focus-N-Fix (Ours)

- **Reward Model:** Artifact Score and Heatmap prediction model proposed in RichHF-18k[2]

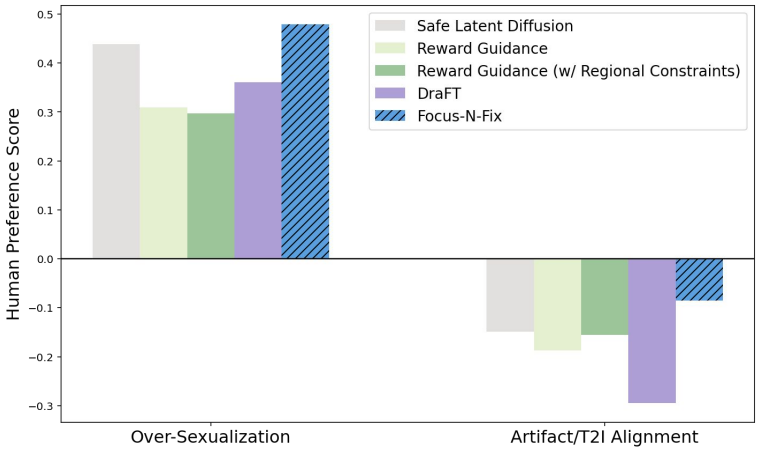
[2] Liang et al, “Rich Human Feedback for Text-to-Image Generation”, CVPR 2024

Human Evaluation Results

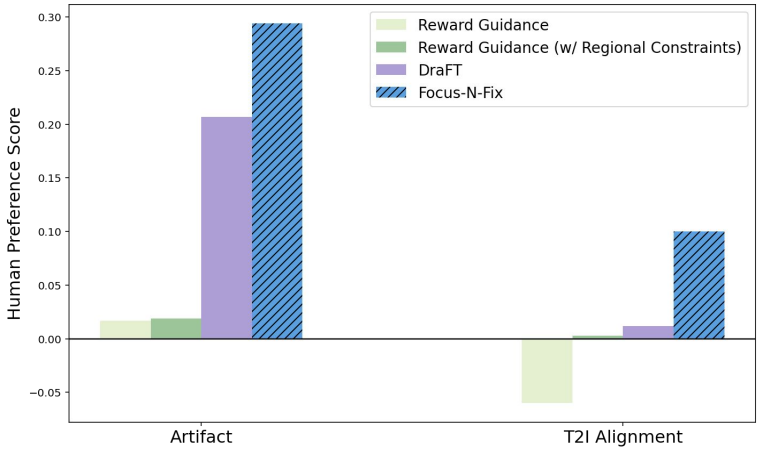
- **100 prompts** used to conduct human evaluations for each fine-tuning task

- Focus-N-Fix **achieves targeted quality** improvements with **least degradation** in other aspects

Over-Sexualization Reward



Artifact Reward



Higher Human Preference Scores are Preferred

Generalization to other T2I Models

Prompt: *The word 'START' on a blue t-shirt.*

SDXL Focus-N-Fix

Artifact Reward

Prompt: *Beautiful Asian Man*

SDXL Focus-N-Fix

Over-Sexualization Reward

Prompt: *A woman by the pool, 4k, concept art, detail, full color, 2d animation, anime style.*

gLDM Focus-N-Fix

Over-Sexualization Reward

Focus-N-Fix generalizes to other T2I Models like SDXL, gLDM (internal latent diffusion model) across reward functions

Generalization to other Rewards

Prompt: *Portrait of a beautiful cyberpunk zombie werewolf made of kale, painting*

SD 1.4 | Focus-N-Fix | Violence Heatmap

Violence Reward

Prompt: *A pear in a robot's hand*

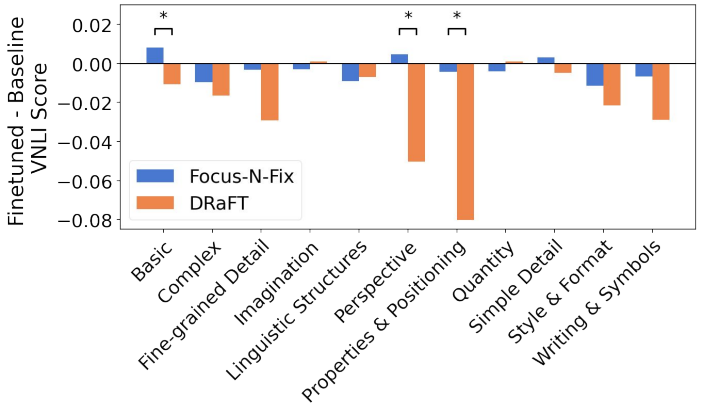
SD 1.4 | Focus-N-Fix | Misalignment Heatmap

Text-to-Image Alignment Reward

Focus-N-Fix also generalizes to other reward functions, such as Violence and Text-to-Image Alignment

Catastrophic Forgetting

- Fine-tuning for a specific reward, like safety, **risks degradation** in **critical aspects** like T2I alignment
- **Focus-N-Fix vs. DRaFT**: T2I alignment degradation from base model using **VNLI**[3] score on **Parti Prompts**



Category-wise mean VNLI score differences:
Safety fine-tuned vs. SD v1.4
(* = statistical significance).

Prompt: *A red box next to a blue box.*

SD 1.4 | DraFT | Focus-N-Fix (Ours)

Oversexualization Reward

[3] Yarom et al, "What You See is What You Read? Improving Text-Image Alignment Evaluation", NeurIPS 2023

Extensions

- **Apply** region-aware constraints to other fine-tuning methods, such as **DPO**
- **Extend** to other **Text-to-Image/Video** Generation Models
- **Extend** to other **localizable rewards** like removing digital watermarks
- **Go beyond fixes -** other region constrained applications like localized style transfer

Thank you!

- **See our paper:** https://arxiv.org/abs/2501.06481
- **Get in Touch:** {avinab,junfenghe}@google.com
- **Visit our Poster Session:**
  - **Time:** 14th June 2025 (PM)
  - **Poster:** 259