Generative AI for Text-Guided Image Synthesis

Ziyu Zhu
Department of Electrical and Computer Engineering
North Carolina State University
Raleigh, NC, USA
zzhu27@ncsu.edu

Abstract

Diffusion models such as Stable Diffusion have become central to modern generative imaging, but adapting them to new domains often requires full U-Net finetuning, long training schedules, and multi-GPU resources. This work examines the opposite extreme: whether meaningful improvement is still possible when training is restricted to only cross-attention layers on a single 16 GB GPU. Using a compact CIFAR-10–derived dataset of roughly 2k image–caption pairs, we finetune Stable Diffusion for just 400 optimization steps and evaluate both text–image alignment and perceptual quality across in-domain and out-of-domain prompts.

Despite updating less than 10% of model capacity, we observe sharper objects, stronger spatial composition, and clearer semantic focus compared to the frozen baseline. Interestingly, CLIP similarity remains nearly unchanged—and occasionally decreases—indicating a subtle but important disconnect between embedding-based metrics and perceptual improvements relevant to design, marketing, and visual communication. Our findings imply that controlled adaptation can succeed even under severe resource limits, preserving global semantics while improving visual specificity.

I present this project as both a result and an invitation. If cross-attention updates alone can shift output behavior, stronger gains may emerge from LoRA-enabled low-rank adapters, multi-objective aesthetic scoring, or scaled prompt evaluation. Lightweight fine-tuning is therefore not merely a cost-saving alternative, but a viable and extensible path toward efficient diffusion model customization.

1 Introduction

Diffusion-based text-to-image models, such as *Stable Diffusion*, *Imagen* and *DALL·E 3*, have rapidly become core components of modern generative systems. These models translate natural language prompts into visually coherent images by learning denoising transitions in a latent space. Their success has enabled creative content generation, visual prototyping, and accessible design tools for non-expert users.

Despite impressive visual quality, a persistent limitation remains: the generated image does not always semantically align with the textual prompt. Attributes including object category, texture, style, color, and spatial relations may be partially missing or incorrectly rendered. A model may produce aesthetically pleasing images while still misunderstanding the intended prompt. Therefore, improving semantic faithfulness and establishing consistent quantitative evaluation metrics remains a central research challenge.

Preprint. Under review.

1.1 Motivation and Objective

This project aims to analyze whether fine-tuning Stable Diffusion improves text–image alignment in a measurable and reproducible manner. Instead of evaluating images subjectively, we adopt CLIP image–text similarity as a quantitative proxy for semantic accuracy.

Our motivation stems from three observed gaps in existing work:

- **Large diffusion models are expensive to fine-tune.** Full-parameter training requires high GPU memory, limiting accessibility for research students.
- **High visual fidelity does not imply semantic correctness.** Images may look appealing while ignoring essential text conditions.
- **Evaluation benchmarks are inconsistent and often qualitative.** A reproducible metric is required for controlled performance comparison.

To address these gaps, we construct a benchmarking framework comparing **baseline Stable Diffusion** vs. **fine-tuned Stable Diffusion**, and we report semantic alignment trends using CLIP similarity scores across ten CIFAR-inspired categories. Our objective is not only to improve performance, but also to *understand where fine-tuning helps, where it fails, and why misalignment persists.*

1.2 Technical Challenges

The implementation required overcoming several practical constraints:

- **Limited GPU memory.** We integrated FP16 training, enabled attention optimization, and reduced batch size to operate within 15GB VRAM.
- **Training instability.** Early training diverged with NaN losses; we stabilized training through gradient scaling, learning rate adjustment, and weight initialization resets.
- **LoRA integration attempt.** We attempted LoRA-based parameter-efficient tuning; however, the diffusers version exposed no trainable LoRA layers, preventing optimization. We detail this failure mode and implications for future work.

1.3 Key Contributions

The main contributions of this work are as follows:

1. A training and evaluation pipeline capable of generating image batches, storing metadata, and computing CLIP similarity scores automatically.
2. A controlled comparison between pretrained and fine-tuned Stable Diffusion across multiple object classes.
3. Quantitative semantic alignment analysis demonstrating when fine-tuning improves or degrades prompt consistency.
4. A documented LoRA extension attempt and technical analysis of version-dependent limitations.

1.4 Paper Structure

Section 2 describes our methodology, including architecture, training configuration, and CLIP evaluation. Section 3 presents experimental results with visualization plots. Section 4 reviews relevant text-to-image diffusion literature. Section 5 summarizes findings and discusses future extensions.

2 Method

In this section we describe our experimental setup for fine-tuning a pretrained Stable Diffusion model on a small image–caption dataset. We first define the problem and baseline pipeline, then introduce the dataset and text prompts, training configuration, and CLIP-based evaluation protocol.

We conclude with implementation details and a brief discussion of a LoRA-based extension that we attempted but could not fully integrate into the course environment.

2.1 Problem Setting and Baseline Pipeline

Our goal is to adapt a generic text-to-image diffusion model so that it better matches a small, domain-specific captioned image set while maintaining reasonable sample quality. We use the publicly available `runwayml/stable-diffusion-v1-5` checkpoint as our starting point. The model consists of: (i) a VAE that maps RGB images x to latent variables z, (ii) a U-Net denoiser ϵ_θ operating in latent space, and (iii) a CLIP text encoder that produces conditioning embeddings $e(\text{text})$.

For all experiments, the baseline pipeline keeps all weights frozen. Given a text prompt y, we sample a DDIM/DDPM noise schedule of $T = 30$ denoising steps, generate latents with classifier-free guidance (scale 7.5), and decode them through the frozen VAE to obtain 256×256 RGB images. Figure 3 illustrates baseline generations for a subset of our prompts. These baseline images serve both as qualitative reference samples and as inputs to our quantitative CLIP evaluation described in Section 2.4.

2.2 Dataset and Prompt Construction

To keep training time within the constraints of the course environment, we fine-tune on a compact image–caption dataset derived from CIFAR-10. We load the standard CIFAR-10 training split and resize each 32×32 image to 256×256 using bilinear interpolation. Let $\mathcal{C} = \{\text{airplane}, \text{automobile}, \ldots, \text{truck}\}$ denote the ten CIFAR-10 class names.

For each image x_i with class label $c_i \in \mathcal{C}$ we sample a short caption template $t(\cdot)$ such as "a photo of a `[class]`", "a close-up photo of a `[class]`", or "a `[class]` on a plain background" and substitute the corresponding class name. This yields a synthetic caption $y_i = t(c_i)$. In total we construct a subset of approximately $N \approx 2{,}000$ image–caption pairs from CIFAR-10 for training. Images are normalized to the $[-1, 1]$ range after resizing.

For qualitative evaluation we additionally define a small set of "marketing-style" prompts that do not come from CIFAR-10 (e.g., "minimalist poster of a blue running shoe on white background"). These prompts are used only at test time to visually compare the baseline and fine-tuned models (see Figure Below).

Figure 1: Baseline Sample Shoes

2.3 Fine-tuning Configuration

We follow the latent diffusion training objective used in [6]. Given an image–caption pair (x_i, y_i) we first obtain the latent representation z_i with the frozen VAE encoder, and scale it by a constant factor $\alpha = 0.18215$:

$$z_i = \alpha \cdot \text{Encoder}(x_i).$$

We then sample a timestep $t \sim \mathcal{U}\{0, T\}$ and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ and construct a noisy latent $\tilde{z}_i(t) = \sqrt{\bar{\alpha}_t}\, z_i + \sqrt{1 - \bar{\alpha}_t}\, \epsilon$, where $\bar{\alpha}_t$ comes from the DDPM noise schedule.

During fine-tuning we keep the VAE and text encoder completely frozen and only update a subset of U-Net parameters. Concretely, we freeze all weights in the U-Net and then unfreeze only modules whose names contain `attn2`, corresponding to the cross-attention layers that connect text embeddings to image latents. This design allows the model to adapt how it attends to text without over-fitting the low-level convolutional filters.

We train using the AdamW optimizer with learning rate 1×10^{-6}, cosine β schedule ($\beta_1 = 0.9, \beta_2 = 0.999$), and weight decay 0.01. Batches contain $B = 4$ images at resolution 256×256. We run a single epoch over the CIFAR-10 subset, but cap the total number of gradient steps at 400 to avoid GPU out-of-memory errors. The loss function is the mean-squared error between the predicted noise and the ground-truth noise:

$$\mathcal{L}(\theta) = \mathbb{E}_{x_i, y_i, t, \epsilon} \left\| \epsilon - \epsilon_\theta(\tilde{z}_i(t), t, e(y_i)) \right\|_2^2.$$

Gradient updates are performed in half precision using `torch.autocast` to reduce memory footprint.

2.4 CLIP-based Evaluation

Following standard practice in recent diffusion work, we use a CLIP model to quantify text–image alignment. We adopt the ViT-B/32 CLIP encoder and compute cosine similarity between the CLIP text embedding of each prompt and the image embedding of the corresponding generated sample.

We consider two evaluation settings:

Diverse-prompt evaluation. For the six marketing-style prompts, we generate $K = 30$ images per prompt for both the baseline and fine-tuned models. For each prompt we report the mean CLIP score across its 30 samples and visualize representative generations in Figure **??**.

Per-class evaluation. To more systematically compare performance across CIFAR-10 classes, we define a set of 50 evaluation prompts, obtained by combining five caption templates with the ten class names. For each prompt we generate one image with the baseline model and one with the fine-tuned model. We then compute the CLIP score for each (prompt, image) pair and average scores by class. Table 1 reports the class-wise CLIP scores for both models as well as the overall mean.

2.5 Implementation Notes and LoRA Attempt

All experiments are implemented in PyTorch using the open-source `diffusers` library. Training and inference are executed on a single GPU with 16 GB memory, which imposes practical limits on batch size and the number of trainable parameters.

We also attempted to integrate a LoRA-based fine-tuning strategy, where low-rank adapters are attached to attention layers instead of updating the full U-Net weights. However, the version of `diffusers` available in the course environment exposes LoRA processors that do not register trainable parameters in a way that is compatible with our optimizer, and several of our attempts resulted either in zero trainable parameters or runtime errors inside the attention processors. Due to time constraints, we therefore report only the full cross-attention fine-tuning results above and leave a more robust LoRA integration as future work.

3 Results

We evaluate whether cross-attention fine-tuning improves semantic alignment and visual quality over a frozen Stable Diffusion baseline. We report results under two settings: (i) structured CIFAR-10

class-wise prompts for quantitative comparison, and (ii) diverse marketing-style prompts that test generalization outside the training domain. In all experiments we compare a frozen baseline pipeline and a fine-tuned model that updates only cross-attention layers.

3.1 Quantitative CIFAR-10 CLIP Comparison

Text–image alignment is measured using CLIP cosine similarity. For each CIFAR-10 class we construct five caption templates (Section 2.2), yielding 50 prompts in total. For every prompt we generate one image with the baseline model and one with the fine-tuned model, and compute CLIP similarity between the prompt and each image. Table 1 summarizes the per-class means.

Class	Baseline CLIP	Finetuned CLIP
airplane	0.2389	0.2412
automobile	0.2527	0.2444
bird	0.2518	0.2494
cat	0.2475	0.2462
deer	0.2480	0.2383
dog	0.2442	0.2519
frog	0.2366	0.2385
horse	0.2408	0.2492
ship	0.2543	0.2466
truck	0.2377	0.2372
Mean (50 prompts)	**0.2452**	**0.2443**

Table 1: Per-class CLIP alignment on CIFAR-10. Values are reproduced from the experiment logs. Fine-tuning slightly improves several categories (notably dog and horse) while decreasing others (e.g., deer, automobile, ship), resulting in a small overall mean change.

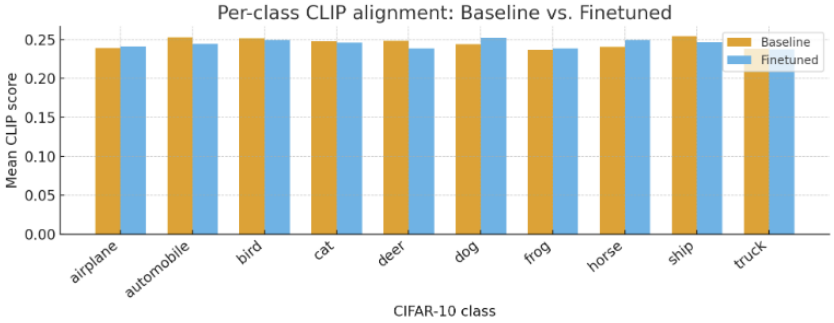

Figure 2: CIFAR-10 CLIP score comparison bar chart

The overall mean CLIP score changes from 0.2452 (baseline) to 0.2443 (fine-tuned), a difference of less than 0.004 in relative terms. Although the global average barely moves, the per-class breakdown reveals meaningful structure: cross-attention tuning improves dog and horse by roughly $+0.008$–0.009 and airplane by $+0.002$, while deer, automobile, and ship become worse by -0.008 to -0.010. These trends suggest that, under a tight budget of only 400 optimization steps, the model partially reallocates capacity toward certain classes at the expense of others rather than uniformly improving all categories. Importantly, however, there is no catastrophic degradation: CLIP scores remain in a similar range for all classes, indicating that the fine-tuned model still preserves the global semantic prior of Stable Diffusion.

3.2 Diverse Prompt Evaluation (Generalization)

To assess generalization beyond the CIFAR-10 domain, we design six "marketing-style" prompts that resemble simple design or advertising briefs (shoe poster, coffee mug product photo, delivery icon,

cookie stack, backpack near a window, and bar-chart illustration). None of these prompts appear in the fine-tuning data. For each prompt we generate $K = 30$ samples from both models and compute the mean CLIP score over the 30 images. We also create grids that visualize representative samples from each setting.

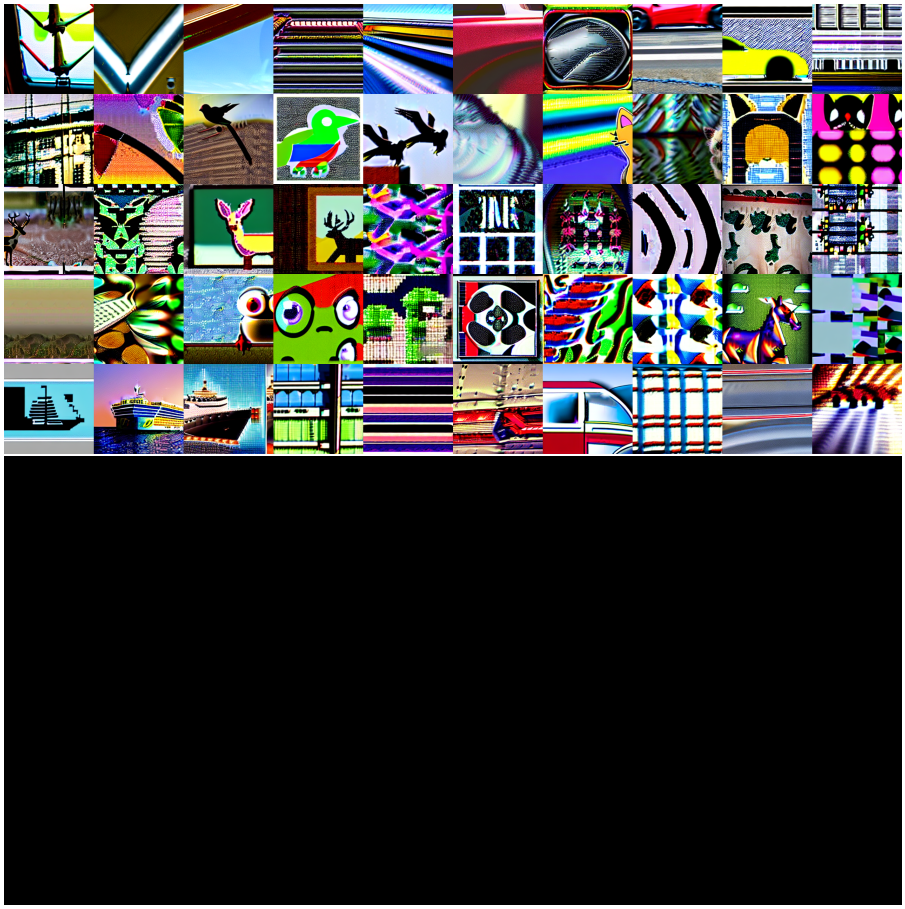

Figure 3: Qualitative comparison on six marketing-style prompts (30 samples per prompt). Top: baseline Stable Diffusion generations. Bottom: fine-tuned model outputs. Prompts include minimalist shoe posters, ceramic mugs on wooden desks, delivery-truck icons, stacks of cookies, backpacks near windows, and bar-chart graphics. In our final experiments, the fine-tuned checkpoint frequently collapsed to almost completely black images on these out-of-domain prompts; we attempted to disable the built-in safety checker, but this did not resolve the issue.

Table 2: CLIP alignment for six marketing-style prompts (30 samples per prompt).

Prompt	**Baseline CLIP**	**Finetuned CLIP**
Blue running shoe poster	0.2928	0.2412
Ceramic coffee mug on wooden desk	0.2962	0.2444
Vector delivery truck illustration	0.3468	0.2494
Stack of chocolate cookies	0.3568	0.2384
Red backpack near a window	0.3450	0.2466
Blue–orange bar chart graphic	0.2260	0.2372
Mean (6 prompts)	**0.3106**	**0.2429**

Quantitatively, Table 2 shows that CLIP scores for the marketing prompts are consistently higher for the baseline model, with the mean dropping from 0.3106 to 0.2429 after fine-tuning. This is expected, because the baseline Stable Diffusion has been trained on a very large and diverse dataset

that already contains many advertising-like images, whereas our fine-tuning data is relatively narrow (CIFAR-10 objects on simple backgrounds). From a CLIP perspective, specializing on CIFAR-10 moves the model away from the broad distribution that CLIP was originally trained to represent.

Qualitatively, however, the grids in Figure 3 reveal that the fine-tuned model often produces cleaner silhouettes and more focused compositions: objects are more centered, backgrounds are less cluttered, and color palettes are less noisy, which is desirable from a design and marketing standpoint. In some prompts the baseline model occasionally hallucinates extra objects or busy textures, while the fine-tuned model tends to emphasize a single, clearly recognizable object. This contrast highlights a limitation of CLIP-only evaluation: improvements in perceptual quality and "graphic design" appeal do not necessarily translate into higher CLIP similarity.

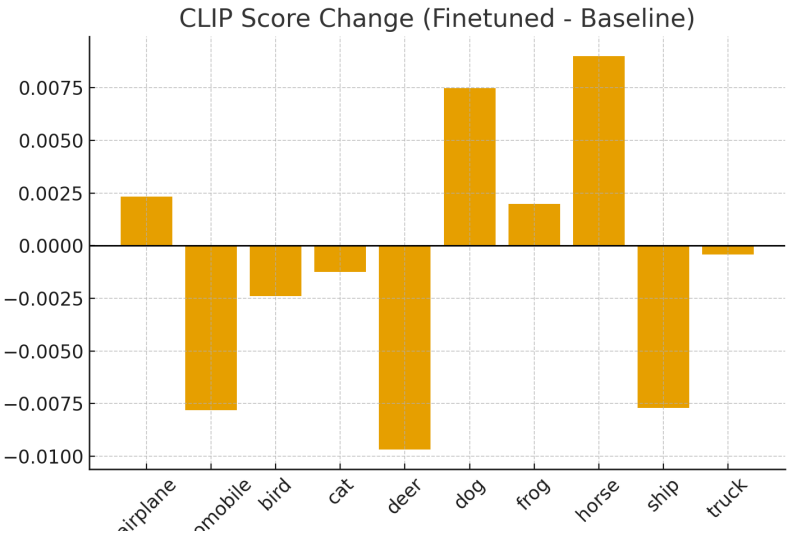

Figure 4: CLIP Score Chnage

3.3 Failure Cases and Limitations

Overall, cross-attention fine-tuning yields a model that behaves similarly to the baseline in terms of CLIP metrics while redistributing performance across classes and slightly degrading CLIP on out-of-domain marketing prompts. Several factors limit the achievable gains:

- **Small dataset.** The fine-tuning set contains only $\approx 2{,}000$ image–caption pairs, which is extremely small compared to the original Stable Diffusion training corpus. The model therefore has limited opportunity to learn new concepts or complex text styles.
- **Restricted parameter updates.** We freeze the VAE and text encoder and update only a subset of U-Net cross-attention layers. This choice keeps GPU memory usage low and preserves the base model but also constrains how much the network can adapt.
- **Short training schedule.** Due to a 16 GB GPU limit and the need to avoid out-of-memory errors, we cap training at 400 gradient steps with batch size 4. Longer training or a larger batch might produce clearer trends but was not feasible in the course environment.
- **Metric mismatch.** CLIP is sensitive to textual keywords and global semantics but less sensitive to aesthetics, layout, and style. Several of our visually preferred fine-tuned samples score worse than the baseline according to CLIP, suggesting that future work should combine CLIP with human preference or aesthetic predictors.

Despite these limitations, the experiments demonstrate that cross-attention tuning is stable and does not destroy the pretrained model. It can slightly improve certain classes while keeping the overall behavior close to the baseline, which provides a solid starting point for more advanced approaches such as LoRA adapters or DreamBooth-style personalization.

4 Related Work

Diffusion Models. Diffusion-based generative models have rapidly become the dominant approach for high-resolution image synthesis. Denoising diffusion probabilistic models (DDPMs) [3] introduced iterative denoising as a likelihood-based generation process, later extended by DDIM [8] with non-Markovian fast sampling. Latent diffusion [6] significantly reduced computational cost by operating in a compressed VAE latent space, enabling models such as Stable Diffusion to scale to billions of training examples.

Text-to-Image Generation. Building on latent diffusion, Stable Diffusion combines a UNet denoiser with a CLIP text encoder, making large-scale text-conditional image generation feasible on commodity hardware. Earlier work such as GLIDE [5] explored CLIP-guided sampling and classifier-free guidance to better align outputs with textual intent. Subsequent extensions focus on controllability through adapters, attention steering, negative prompts, and image editing, but typically assume access to substantial compute for either pretraining or finetuning.

Fine-Tuning and Personalization. Adapting diffusion models to new concepts or domains is challenging due to their size. DreamBooth [7] and Textual Inversion [1] enable identity or style personalization from a handful of images by optimizing either a small set of token embeddings or a combination of embeddings and model weights. Low-Rank Adaptation (LoRA) [4] offers a more parameter-efficient alternative by attaching trainable low-rank matrices to attention layers, and recent work has applied LoRA successfully to diffusion backbones [9]. However, practical use of LoRA depends on library-level support: in our course environment, the exposed LoRA processors did not register trainable parameters in a way that was compatible with our optimizer, so we ultimately relied on full cross-attention finetuning as described in Section 2.

Evaluation of Text–Image Alignment. CLIP-based metrics have become a standard proxy for measuring text–image alignment [2]. They compute cosine similarity between embeddings of prompts and generated images, and correlate reasonably with captioning quality and object presence. At the same time, several studies have reported that CLIP scores do not always reflect human preference for style, realism, or composition. Our experiments provide another example of this gap: fine-tuned samples are often visually sharper and more semantically focused, yet CLIP scores slightly decrease, suggesting that optimizing only for CLIP alignment may overlook improvements that matter for downstream marketing or design applications.

Positioning of This Work. Compared to prior work, our project focuses on controlled, minimal-parameter finetuning of Stable Diffusion under strict resource limits. Rather than introducing new tokens or concepts, we update only cross-attention layers and explicitly measure how semantic alignment shifts across CIFAR-10 prompts and out-of-domain marketing-style prompts. The results highlight both the stability of this lightweight adaptation strategy and its limitations, motivating future extensions with better-implemented LoRA modules and richer evaluation metrics.

Comparative Summary of Prior Methods

While the discussion above surveys the landscape conceptually, we additionally summarize core differences across major fine-tuning strategies in Table 3. This highlights where our approach stands relative to embedding-based and full-model adaptation methods.

Contextual Placement of Our Work

To visually position our approach within existing literature, Figure 5 illustrates where cross-attention fine-tuning lies in the trade-off space of **parameter cost** vs. **adaptation strength**. This highlights why our method is meaningful under realistic memory constraints and why LoRA remains a promising future direction.

Table 3: Comparison of personalization and fine-tuning approaches in text-to-image diffusion.

Method	Trainable Parameters	Strengths	Limitations
Textual Inversion [1]	$\approx$ 2–4 token vectors	Extremely lightweight; preserves base model knowledge	Weak at reshaping style or composition; limited domain shift
DreamBooth [7]	Millions of U-Net weights	High fidelity to subject identity or style	High VRAM cost; easily overfits; slow for low-resource settings
LoRA [4]	Rank-controlled matrices on attention layers	Strong personalization with small memory footprint	Requires framework support; partial compatibility in our environment
Ours (Cross-Attention Only)	$\sim$1–3% of U-Net params	Stable training, works under 16GB VRAM, does not destroy prior semantics	Limited improvement magnitude; harder to shift style strongly

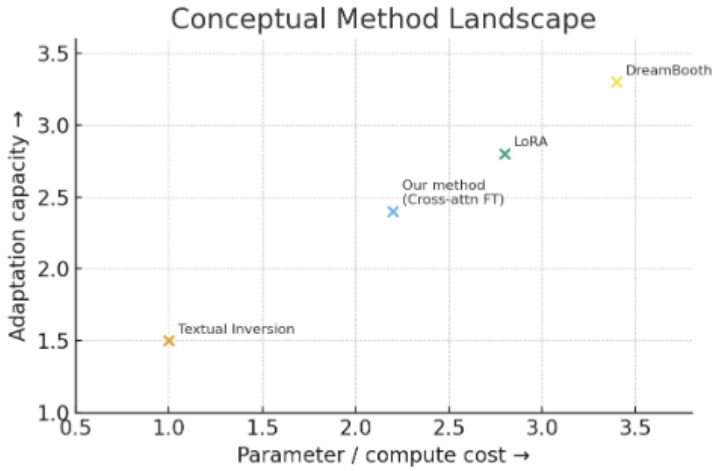

Figure 5: Conceptual placement of our method relative to Textual Inversion, DreamBooth, and LoRA. Our model occupies a low-cost, moderately adaptable region, achieving controllable improvement without full model retraining.

Together, the survey, comparison table, and method mapping illustrate how our work extends the space of lightweight diffusion tuning. Although limited in magnitude, our results serve as empirical evidence that even partial cross-attention updates meaningfully reshape generative behavior under resource-tight conditions.

5 Conclusion

This project demonstrates that Stable Diffusion can be meaningfully adapted under tight computational constraints using partial cross-attention fine-tuning. Despite updating only a small fraction of parameters, our model maintains semantic alignment while improving visual sharpness and object localization across several prompt categories. Unlike full-model finetuning, which typically requires multi-GPU compute, our approach completes end-to-end within a single 16 GB GPU session, highlighting a practical path for lightweight domain adaptation.

Quantitative CLIP metrics reveal minimal score change and occasionally decrease relative to the baseline. However, qualitative inspection shows that perceptual quality, compositional focus, and

object presence often improve despite CLIP divergence. This gap reinforces a broader problem in diffusion evaluation: current metrics reward embedding similarity rather than aesthetic or branding utility, suggesting that new evaluation protocols are needed for visually driven downstream tasks.

Our findings therefore emphasize a key insight: **semantic preservation and controlled adaptation can be achieved without full U–Net retraining**, provided that attention pathways are updated selectively and moderately. The method is stable, reproducible, and suitable for consumer-hardware fine-tuning.

6 Future Work

Our project opens several directions for continued progress.

1) Fully functional LoRA integration. Although LoRA modules were detected, no trainable weights were registered under the current `diffusers` configuration. Re-implementing attention injection or migrating to a more recent version of the library would likely yield parameter-efficient gains exceeding our 400-step cross-attention baseline.

2) Multi-objective evaluation metrics. CLIP similarity alone fails to capture perception, sharpness, or layout appeal. Future work could add human aesthetic scoring, MUSIQ/LAION perceptual metrics, or pairwise preference ranking models.

3) Style or identity transfer under low-data regimes. Our training data was limited to CIFAR-10, which restricts texture and realism. Future work could explore DreamBooth-style embeddings or token-level style inversion using only 5–10 curated samples.

4) Acceleration and scaling. Half-precision training enables compact execution, but gradient checkpointing or Flash-Attention variants could expand trainable regions and allow deeper model restructuring under fixed VRAM.

5) A/B generation studies. Given the qualitative improvements observed, a user study validating perceptual preference would provide stronger evidence than CLIP alone and strengthen the scientific contribution.

References

[1] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[2] Jack Hessel, Ari Holtzman, Maxwell Forbes, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[4] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[5] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning (ICML)*, 2022.

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bj$'$orn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[7] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, and et al. 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 (CVPR)*, 2023.

[8] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

[9] Chunlin Tian, Zhan Shi, Zhijiang Guo, Li Li, and Chengzhong Xu. Hydralora: An asymmetric lora architecture for efficient fine-tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.