DreamBooth: Personalizing Text-to-Image Diffusion Models for Subject-Driven Generation

1 Introduction

Large text-to-image diffusion models can synthesize high-quality, diverse images from natural-language prompts. However, they typically struggle to reliably reproduce a specific subject's identity from only a few reference images. DreamBooth addresses this by fine-tuning a pre-trained text-to-image model using a small set of images (typically 3–6) of a target subject. By binding a rare-token identifier to that subject, the resulting personalized model can render the subject in new scenes, poses, and styles while preserving its distinctive visual features.

This report evaluates DreamBooth across three distinct subjects—a dog, a human face, and a complex object—to analyze the trade-offs between subject fidelity, prompt adherence, and output diversity.

2 Methodology

2.1 Text-to-Image Diffusion Models

Diffusion models learn a data distribution by reversing a fixed-length noising process. Given a clean image x and a text-conditioning vector c, the model is trained to denoise a noisy input $z_t = \alpha_t x + \sigma_t \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$ and (α_t, σ_t) follow a prescribed noise schedule. A text-conditional denoiser $\hat{x}_\theta$ is trained to predict x from (z_t, c) via a weighted mean-squared error (MSE) objective:

$$\mathbb{E}_{x,\, c,\, \epsilon,\, t}\Big[w_t \, \big\| \hat{x}_\theta\big(\alpha_t x + \sigma_t \epsilon,\; c\big) - x \big\|_2^2 \Big]. \tag{1}$$

2.2 DreamBooth Fine-Tuning

DreamBooth's fine-tuning process introduces two key components to preserve the model's prior knowledge while learning a new subject.

Rare-Token Identifiers. Each training image is paired with a prompt of the form "a [unique-id] [class noun]", e.g., "a photo of [V] dog". The unique identifier (a rare token like 'sks' or '[V]') becomes bound to the specific subject, while the class noun (e.g., 'dog') anchors the model to its existing semantic knowledge of that category.

Class-Specific Prior Preservation Loss (PPL). To prevent the model from forgetting the general class concept (a phenomenon known as **language drift**), DreamBooth uses a regularization term. The model generates its own "prior" images using a generic class prompt (e.g., "a photo of a dog") with its weights frozen. These autogenously generated samples are then used in a secondary loss term. The combined objective is:

$$\begin{aligned} \mathbb{E}\big[\; & w_t \, \big\| \hat{x}_\theta\big(\alpha_t x + \sigma_t \epsilon,\; c\big) - x \big\|_2^2 \\ & + \; \lambda \, w_{t'} \, \big\| \hat{x}_\theta\big(\alpha_{t'} x_{\mathrm{pr}} + \sigma_{t'} \epsilon',\; c_{\mathrm{pr}}\big) - x_{\mathrm{pr}} \big\|_2^2 \;\big], \end{aligned} \tag{2}$$

where the first term is the loss on the user-provided subject images and the second is the prior-preservation loss on the generated class images, weighted by λ.

3 Evaluation Methodology

All models were evaluated against the base Stable Diffusion v1.5 model on a fixed set of **10 prompts**, with **10 random seeds** for each prompt (totaling 100 generations per model).

Metric Backbones.

- **Text and Image Embeddings:** OpenCLIP ViT-B/32 was used for text-image similarity (Prompt Adherence).
- **Identity Embeddings:** For robust identity comparison, we used two models. For human faces, we used OpenCLIP ViT-B/32. For objects (dog, monster toy), we used DINOv2 (ViT-Base), a self-supervised model known for strong instance-level feature matching.
- **Perceptual Diversity:** LPIPS with an AlexNet backbone was used to measure perceptual distance between generated images.

Quantitative Metrics.

- **Subject Fidelity:** To measure how accurately the model reproduces the original subject, we compute the cosine similarity between a generated image and the set of original training images. We report both the **maximum similarity** (how well it matches the closest training image) and the **average similarity** (overall fidelity across all training examples).
- **Prompt Adherence:** The cosine similarity between the CLIP embedding of the prompt text and the CLIP embedding of the generated image. Higher is better.
- **Diversity (LPIPS):** The average LPIPS distance between all pairs of images generated with the same prompt across different seeds. Higher distance implies greater visual diversity.

4 Experiment 1: Training Regimes on a Dog Subject

Our first experiment compares three fine-tuning regimes on a dataset of **5 images** of a pet dog (‘sks dog‘) to find a balance between subject fidelity and editability.

4.1 Hyperparameters and Justification

The configurations for the three regimes are detailed in Table 1. We used fixed-step checkpoints for all dog models to ensure an unbiased comparison of hyperparameter effects, removing any potential for manual cherry-picking.

Table 1: Hyperparameter configurations for the three dog training regimes.

Parameter	Overfit	Underfit	Balanced
Learning Rate	5×10^{-6}	5×10^{-7}	1×10^{-6}
Max Train Steps	400	600	800
Resolution	256	512	512
Prior Preservation	False	True	True
Prior Loss Weight (λ)	1.5	1.0	0.6
Num. Class Images	100	200	100
Train Text Encoder	False	True	True
Mixed Precision	fp32	bf16	bf16

4.2 Quantitative Results

The results in Table 2 highlight the fundamental **trade-off** inherent in fine-tuning generative models like DreamBooth. By analyzing the performance of different training regimes—Overfit, Underfit, and Balanced—across metrics for subject fidelity, prompt adherence, and diversity, we can determine the optimal balance.

Metrics.

- **Subject fidelity (Subj Max / Subj Mean).** Measures how well the generated images preserve the identity of the target subject compared to reference images. Computed using CLIP (ViT-B/32) cosine similarity between generations and training references. Subj Max reports the best match per image, while Subj Mean reports consistency across references.
- **Prompt adherence (Text–Img).** Evaluates how well outputs align with the conditioning prompt. Computed as CLIP cosine similarity between the text prompt and the generated image.
- **Diversity.** Estimates visual variety across generations for the same prompt. Calculated via LPIPS (AlexNet backbone) across different seeds; lower values indicate more diverse outputs.
- **Stability.** Reflects consistency across prompts and seeds. Reported as the standard deviation of subject fidelity and prompt adherence scores.

Table 2: Detailed quantitative comparison across DreamBooth training regimes (mean ± std). Bold values indicate the strongest result in each column. ↑ indicates higher is better, ↓ indicates lower is better.

Variant	Subj Max ↑	Subj Mean ↑	Text–Img ↑	Diversity ↓	Stability (Subj/Text)
Balanced	0.855 ± 0.066	0.843	0.288 ± 0.029	0.903	0.066 / 0.029
Overfit	**0.881** ± 0.055	**0.870**	0.275 ± 0.031	0.915	0.055 / 0.031
Underfit	0.784 ± 0.066	0.773	**0.296** ± 0.029	**0.880**	0.066 / 0.029

The findings from our quantitative evaluation are summarized below:

- **Overfit:** Achieved the highest **subject fidelity**, as indicated by Subj Max (**0.881**) and Subj Mean (**0.870**). This shows it learned the subject's features strongly, but at the cost of weaker prompt adherence and reduced diversity.

- **Underfit:** Prioritized flexibility, with the best **prompt adherence** (Text–Img **0.296**) and greatest **diversity** (lowest similarity, **0.880**). However, it struggled to reproduce the subject consistently (Subj Mean 0.773).
- **Balanced:** Provided the best **compromise**. It achieved strong subject fidelity (Subj Mean 0.843), while maintaining prompt adherence (0.288) and diversity (0.903) close to the Underfit model. This regime minimizes the trade-off effectively.

4.3 Qualitative Results

Visual results in Figure 1 confirm the quantitative findings. The balanced model successfully places the 'sks dog' in new styles and contexts while preserving its key features.

(a) Prompt: "a watercolor painting of sks dog"

(b) Prompt: "sks dog in front of the eiffel tower"

Figure 1: Qualitative comparison of dog models. For each prompt, the images show (from left): reference image, base model, overfit, underfit, and balanced. The balanced model best combines subject identity with the prompt's style and context.

5 Experiment 2: Training on a Human Face

To validate the methodology on human faces, which have strong priors in the base model, we trained on **6 images** of a person ('sks person').

5.1 Training Procedure and Justification

We used a two-stage training process. An initial training run of 300 steps showed promising contextual alignment but lacked precise facial identity (Figure 2). To improve fidelity, we continued training to **1200 steps**. This manual, iterative approach is common for subjects requiring high detail, allowing for qualitative checks to determine the optimal stopping point.

Listing 1: Configuration for human face training.

```
INSTANCE_PROMPT = "a␣photo␣of␣sks␣person"
CLASS_PROMPT    = "a␣photo␣of␣person"
MAX_TRAIN_STEPS = 1200
LEARNING_RATE   = 1e-6
PRIOR_LOSS_WEIGHT = 1.0
NUM_CLASS_IMAGES = 150
```

```
TRAIN_TEXT_ENCODER = true
```

Figure 2: Sample generations at step 300. The model captures the prompt's context (e.g., clothing, background) but has not yet converged on the subject's specific facial features.

5.2 Results

The final 1200-step model showed significant improvement in reproducing the target identity while preserving editability. Quantitative metrics in Table 3 show a substantial increase in subject fidelity (+0.184 improvement over the base model) with only a minor drop in prompt adherence (-0.012). The model also maintained good generative diversity.

Table 3: Detailed metrics for human face training (1200 steps).

Metric	Base Model	Trained Model
Subject Fidelity (Max)	0.593	**0.777**
Subject Fidelity (Avg)	0.542	**0.722**
Prompt Adherence	**0.268**	0.256
Diversity (LPIPS)	N/A	**0.598**

Figure 3: Final qualitative result for the human face model. Base Stable Diffusion generations (left half) vs. DreamBooth-trained generations (right half) for the prompt: "a side profile photo of sks man."

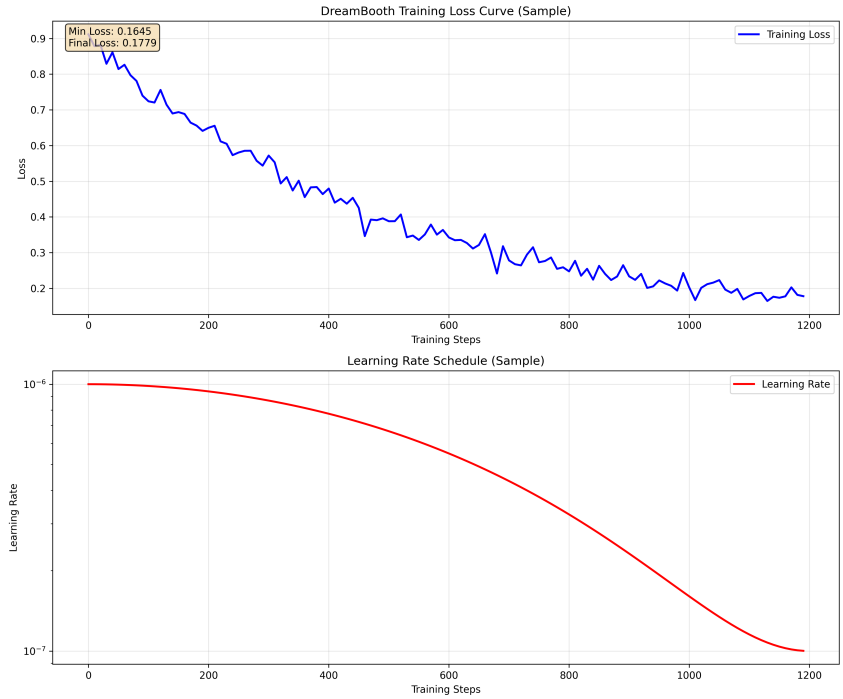

Figure 4: Training logs for 1200 steps on the human face: loss and learning rate schedule.

6 Experiment 3: Training on a Complex Object

Our final experiment tests DreamBooth on a more challenging subject: a unique monster toy with complex textures and an irregular shape, using **5 training images**. Such objects lack a strong prior in the base model.

6.1 Training Procedure and Justification

We used a two-phase training schedule. An initial 600 steps with high regularization ($\lambda = 1.0$) resulted in good prompt adherence but poor subject fidelity. To address this, we **reduced the prior loss weight to** $\lambda = 0.5$ for steps 600–1200. This adjustment reduces the influence of the class prior, allowing the model to focus more capacity on memorizing the toy's specific, unusual features.

Listing 2: Final configuration for monster toy training.

```
INSTANCE_PROMPT = "a␣photo␣of␣sks␣monster␣toy"
CLASS_PROMPT    = "a␣photo␣of␣monster␣toy"
MAX_TRAIN_STEPS = 1200
LEARNING_RATE   = 1e-6
PRIOR_LOSS_WEIGHT = 0.5   # Final value (was 1.0 for first 600 steps)
NUM_CLASS_IMAGES = 150
TRAIN_TEXT_ENCODER = true
```

6.2 Results

The adjustment in regularization proved effective. The final model at 1200 steps successfully learned the subject's identity, achieving a massive +0.345 improvement in max fidelity over the base model (Table 4). Notably, prompt adherence slightly increased, and the model maintained high diversity, indicating a well-balanced result.

Table 4: Detailed metrics for monster toy training (1200 steps).

Metric	Base Model	Trained Model
Subject Fidelity (Max)	0.264	**0.609**
Subject Fidelity (Avg)	0.209	**0.518**
Prompt Adherence	0.287	**0.290**
Diversity (LPIPS)	N/A	**0.699**

Figure 5 shows a qualitative example from the final checkpoint, demonstrating successful recontextualization of the unique toy.

Figure 5: Qualitative result for the monster toy model at 1200 steps. Base model generations (left half) vs. DreamBooth-trained generations (right half) for the prompt: "sks monster toy near a window."

7 Conclusion

Our experiments demonstrate that DreamBooth can effectively personalize a text-to-image model for a wide range of subjects. We found that:

- The trade-off between subject fidelity, prompt adherence, and diversity can be controlled by tuning hyperparameters like learning rate, training steps, and the prior-preservation weight (λ).
- A **balanced approach** with prior preservation provides the best results, preventing overfitting while ensuring the subject's identity is learned. Our fine-tuned models showed significant fidelity improvements: **+0.184** for the human face and a remarkable **+0.345** for the complex toy.
- Different subjects may require different training strategies. A simple fixed-step approach is suitable for comparing regimes, while more complex subjects benefit from **manual checkpoint inspection and adaptive regularization**.

By using a rigorous evaluation protocol, we confirmed that DreamBooth provides a powerful and flexible tool for subject-driven image generation.