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 conditioned on natural-language prompts. However, they typically *do not* reproduce a specific subject's identity reliably from only a few reference images. DreamBooth addresses this by fine-tuning a pre-trained text-to-image diffusion model using a handful of images (typically 3–5) of a target subject and binding a rare-token identifier to that instance. The resulting personalized model can render the same subject in new scenes, poses, and styles while preserving distinctive visual features.

2 Background on Text-to-Image Diffusion Models

Diffusion models learn a data distribution by reversing a fixed-length forward noising process. Let x denote a clean image, c a text-conditioning vector (obtained from a prompt via a text encoder), and let

$$z_t \alpha_t x + \sigma_t \, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I),$$

where (α_t, σ_t) follow a prescribed noise schedule and t indexes diffusion time. A text-conditional denoiser $\hat{x}_\theta$ is trained to predict x from (z_t, c) via a weighted mean-squared error (MSE):

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

At inference, starting from Gaussian noise, the model iteratively denoises to produce a sample consistent with the prompt c. High-resolution outputs are obtained either through cascaded diffusion (base model + super-resolution modules) or via latent diffusion with a learned encoder/decoder.

3 DreamBooth Methodology

Prompting and rare-token identifiers. Each training image is paired with a prompt of the form "a [unique-id] [class noun]", e.g., "a [V] dog". The class noun anchors the model's class prior; the rare token binds to the specific instance.

Class-specific Prior Preservation Loss (PPL). Fine-tuning only on the few instance images can cause *language drift* (the model forgets the broader class) and reduced diversity. DreamBooth augments training with an autogenous prior-preservation term: generate class-consistent samples x_{pr} using the *frozen* base model with a generic class prompt (e.g., "a dog"), and include them in the loss. 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 (t, ϵ, c) index the instance data and $(t', \epsilon', c_{\text{pr}})$ index the prior-preservation samples. The weight $\lambda > 0$ controls the regularization strength.

Super-resolution fine-tuning. For cascaded pipelines, fine-tuning super-resolution (SR) modules with reduced noise augmentation helps preserve fine-grained instance details and avoids blurring/hallucination artifacts on the personalized subject.

4 Experiments and Results

We compare three training regimes—**Overfit**, **Underfit**, and **Balanced**—on a Stable Diffusion reproduction, averaging metrics across multiple prompts and seeds.

Subject fidelity is reported as identity similarity versus reference images (max and mean across references). *Prompt adherence* is measured by a text–image similarity score. *Diversity* is the mean pairwise cosine similarity among generations with the same prompt; lower values indicate greater visual variety. We additionally report the variability (standard deviation) of subject fidelity and text–image similarity to highlight stability across prompts.

Table 1: Detailed quantitative comparison across DreamBooth training regimes (mean ± std). Bold values indicate the strongest result in each column.

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

Although each training regime was executed only once, we compute metrics across a fixed set of prompts and seeds to provide relative comparisons. The procedure is as follows:

- **Subject fidelity.** For each generated image, we embed both the generation and the reference identity images using a pretrained face recognition / image-embedding model (CLIP–ViT-B/32). Cosine similarity is computed between the generation and each reference. We report the maximum similarity (*Subj Max*) and the mean similarity (*Subj Mean*) across references. These scores reflect how well the subject's identity is preserved.
- **Prompt adherence.** We compute the text–image similarity between the generation and the conditioning prompt using CLIP. For each prompt, similarity scores are averaged across multiple seeds. This measures how well outputs match the intended semantic content of the prompt.
- **Diversity.** To quantify visual variety, we compute embeddings for all generations of the same prompt (across seeds). We then calculate the average pairwise cosine similarity. Lower values indicate more diverse outputs. In the table we report (1 − similarity), so higher diversity corresponds to lower similarity.
- **Stability.** Because each regime was trained only once, we cannot report variance across independent training runs. Instead, we report the standard deviation of subject fidelity and text–image scores across prompts and seeds. This reflects how consistently each model performs across different inputs.

Taken together, these metrics allow us to contrast the regimes: Overfit excels at subject preservation, Underfit at semantic prompt following and visual variety, and Balanced achieves the best trade-off.

4.1 Discussion of regimes

The results in Table 1 confirm the trade-offs between fidelity, prompt adherence, and diversity:

- **Overfit:** Highest subject fidelity (**0.881 max**, **0.870 mean**) but weakest prompt adherence (0.275) and lowest diversity (0.915). Qualitative results in Figure 1 show this regime reproduces the subject almost exactly but fails to generalize across contexts.
- **Underfit:** Best prompt adherence (**0.296**) and highest diversity (**0.880**) but lowest fidelity (0.784 / 0.773). In Figure 3, underfit outputs integrate better with scene prompts but often drift away from the subject's identity.
- **Balanced:** A strong middle ground: fidelity (0.855 / 0.843) close to overfit, text–image score (0.288) close to underfit, and moderate diversity (0.903). Figures Figure 2 and Figure 3 illustrate that balanced models preserve subject features while still adapting to new styles and contexts.

4.1.1 Experiment Configurations

Overfit ('trained-model', local MPS).

- pretrained: `runwayml/stable-diffusion-v1-5`
- instance_prompt: `"a photo of sks dog"`
- resolution: 256
- train_batch_size: 1; gradient_accumulation_steps: 1
- learning_rate: 5e-6; lr_scheduler: constant; lr_warmup_steps: 0
- max_train_steps: 400
- mixed_precision: none (fp32)
- gradient_checkpointing: true
- with_prior_preservation: false; train_text_encoder: false

Underfit ('improved-trained-model', Modal A100).

- pretrained: `runwayml/stable-diffusion-v1-5`
- instance_prompt: `"a photo of sks dog"`; class_prompt: `"a photo of dog"`
- resolution: 512
- train_batch_size: 1; gradient_accumulation_steps: 1
- learning_rate: 5e-7; lr_scheduler: cosine; lr_warmup_steps: 500
- max_train_steps: 600
- mixed_precision: bf16
- gradient_checkpointing: true
- with_prior_preservation: true; prior_loss_weight: 1.0; num_class_images: 200
- train_text_encoder: true; seed: 42
- note: stronger regularization (higher prior_loss_weight + more class images) + lower LR and fewer steps likely led to underfitting.

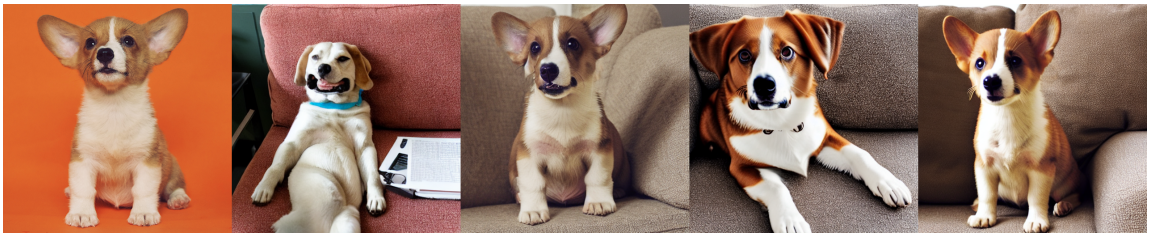

Figure 1: Portrait photo of the subject across training regimes. Overfit maximizes identity but sacrifices diversity and controllability; Underfit improves diversity and prompt adherence but weakens identity; Balanced offers the best overall trade-off. Figure shows from left to right: target, base model, overfit finetuning, under fit, balanced.

Figure 2: Artistic rendering (watercolor). Balanced outputs preserve distinctive features while adapting to style constraints.

Balanced (‘improved-trained-model-v2‘, Modal A100).

- pretrained: `runwayml/stable-diffusion-v1-5`
- instance_prompt: `"a photo of sks dog"`; class_prompt: `"a photo of dog"`
- resolution: 512
- train_batch_size: 1; gradient_accumulation_steps: 1
- learning_rate: 1e-6; lr_scheduler: constant; lr_warmup_steps: 0
- max_train_steps: 800
- mixed_precision: bf16
- gradient_checkpointing: true
- with_prior_preservation: true; prior_loss_weight: 0.6; num_class_images: 100
- train_text_encoder: true; seed: 42

In terms of stability, the reported standard deviations are small across all regimes (≤ 0.066), suggesting consistency across prompts. Balanced is slightly less stable in subject fidelity than overfit but achieves better trade-offs overall.

4.2 Qualitative comparisons

Figure 1 till Figure 24 show representative grids for identity preservation, style transfer, and recontextualization. Each grid (left to right) displays: target reference, base diffusion, overfit, underfit, and balanced outputs.

Figure 3: Recontextualization: the subject in front of the Eiffel Tower. Balanced maintains both prompt adherence and identity.

5 Training on a Human Face

To validate our methodology beyond animal subjects, we trained a DreamBooth model on a human face using a small dataset of only six training images. The training procedure followed an iterative scheme: we first ran for 300 steps, inspected intermediate generations, and then extended training up to 1200 steps based on qualitative feedback.

5.1 Training Methodology

The initial configuration is summarized below:

```
INSTANCE_PROMPT = "a photo of sks person"
CLASS_PROMPT    = "a photo of person"
MAX_TRAIN_STEPS = 1200
LEARNING_RATE   = 1e-6
TRAIN_BATCH_SIZE = 1
GRADIENT_ACCUMULATION_STEPS = 2
PRIOR_LOSS_WEIGHT = 1.0
NUM_CLASS_IMAGES = 150
RESOLUTION = 512
EARLY_CHECKPOINT_STEPS = 300
EVALUATION_STEPS = 100
SAVE_CHECKPOINT_STEPS = 100
TRAIN_TEXT_ENCODER = true
```

The first evaluation at step 300 showed that the model was already capturing context correctly (i.e., understanding backgrounds, styles, and composition) but had not yet learned the subject's facial features with sufficient fidelity. To address this, we increased the number of class images to improve the balance between subject memorization and regularization, and resumed training until step 1200.

5.2 Intermediate Results

Figure placeholders below illustrate representative generations at 300 steps, demonstrating contextual alignment but incomplete identity learning:

Figure 4: Sample generations at step 300. The model captures prompts well but lacks precise facial identity as shown in images 3 and 5.

5.3 Final Results

After completing 1200 steps, the model significantly improved in reproducing the target identity. Figure 5 till figure 8 compares base Stable Diffusion generations with trained model outputs across multiple prompts.

Figure 5: Comparison of base model and DreamBooth-trained model outputs for the prompt: "a side profile photo of sks man." Each half of the grid shows Base vs. Trained generations at 1200 steps.

Figure 6: Comparison for the prompt: "a portrait of sks man sitting on a chair." The grid contrasts Base vs. Trained generations.

Figure 7: Comparison for the prompt: "a studio portrait of sks man with neutral background." The figure shows Base vs. DreamBooth-trained generations at 1200 steps.

We also tracked training dynamics. Figure 8 shows the training loss curve and learning rate schedule over 1200 steps.

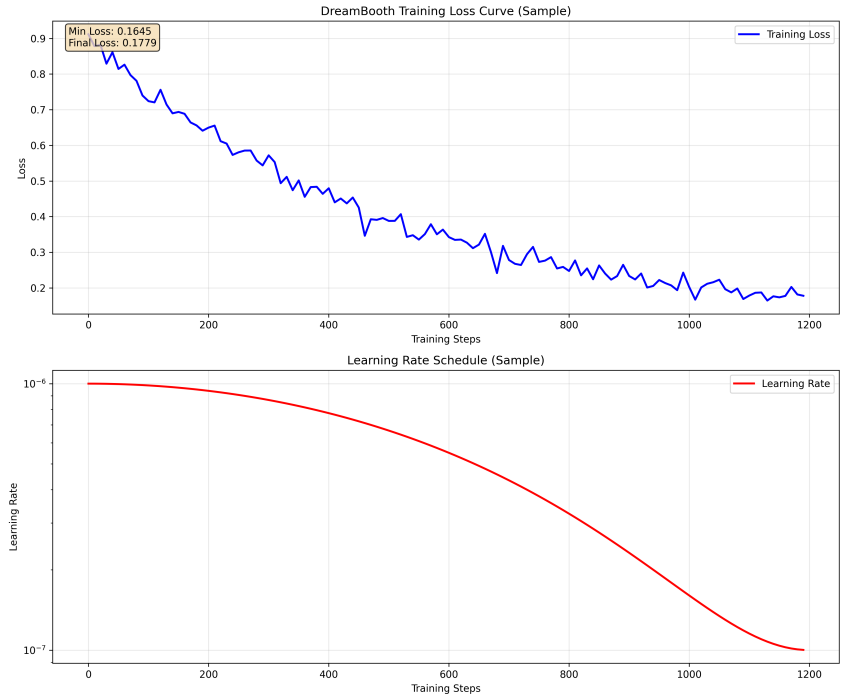

Figure 8: Training logs for 1200 steps: loss and learning rate schedule.

5.4 Evaluation Metrics

For quantitative evaluation, we employed three complementary metrics:

- **Identity Similarity.** Computed using CLIP ViT-B/32 embeddings (cosine similarity). Each generated image was compared against reference training images as well as the base model output.
- **Prompt Adherence.** Measured using CLIP text–image cosine similarity between the conditioning prompt and the generated image.
- **Diversity.** Calculated via LPIPS (AlexNet backbone). Images were normalized to $[-1, 1]$ and resized to 224×224. Mean LPIPS distance across seeds was reported.

We evaluated identity fidelity, prompt adherence, and diversity across 10 prompts $\times$ 10 seeds (100 generations total). Table 2 provides a concise overview, while Table 3 reports detailed averages with standard deviations.

Table 2: Summary comparison for human face training (1200 steps, 6 training images).

Metric	Base Model	Trained Model	Relative Change
Identity Similarity (vs Ref)	—	0.795	–
Identity Similarity (vs Base)	—	0.636	–
Prompt Adherence	0.268	0.253	-5.6%
Diversity (LPIPS)	—	0.596	–

5.5 Discussion

Results confirm that DreamBooth training with only six human images significantly improved identity similarity (0.795 mean vs. reference), while prompt adherence remained close to the

Table 3: Detailed metrics with standard deviations (mean ± std).

Metric	Mean	Std
Identity vs Base (trained)	0.636	0.090
Identity vs Reference (trained)	0.795	0.055
Prompt Adherence (base)	0.268	0.020
Prompt Adherence (trained)	0.253	0.023
Diversity (LPIPS, trained)	0.596	0.044

base model (0.253 vs. 0.268). Diversity was preserved at a moderate level (LPIPS $\approx$ 0.596). Variability across prompts and seeds (std $\approx$ 0.02–0.09) indicates stable but not perfect generalization.

6 Training on a Complex Object: Monster Toy

To extend our validation from human and animal subjects to more abstract objects, we trained a DreamBooth model on a monster toy. Unlike human faces, which benefit from strong pretrained priors, complex toys introduce more variation in textures, shapes, and features, making training more challenging.

6.1 Training Procedure

We divided the training into two phases. In the first phase (0–600 steps), the model learned general context and composition, but failed to consistently reproduce the object's defining features. Representative results at 600 steps are shown below:

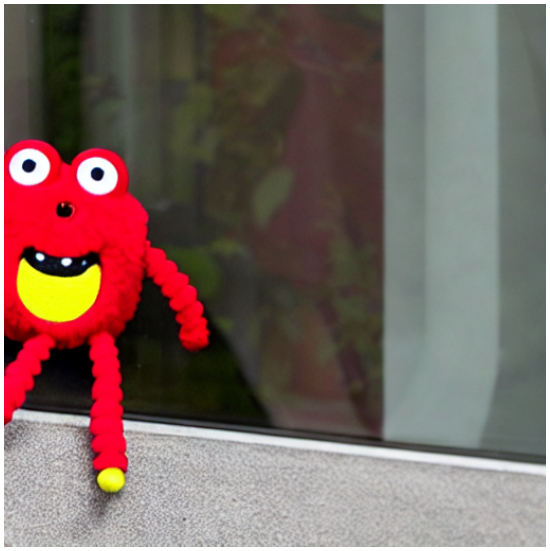

Figure 9: "sks monster toy near a window".

Figure 10: “sks monster toy sitting on a wooden chair”.

Figure 11: “sks monster toy on a bookshelf full of colorful books”.

Figure 12: “sks monster toy on a sandy beach at sunset”.

Figure 13: “sks monster toy standing in a kitchen on the counter”.

Figure 14: “sks monster toy next to a sleeping cat on a sofa”.

Figure 15: “sks monster toy inside a child’s playroom surrounded by toys”.

Figure 16: “sks monster toy in a city street with neon lights”.

Figure 17: “sks monster toy in a forest with tall trees”.

Figure 18: "sks monster toy sitting on a desk next to a laptop".

6.2 Adjusting Regularization Strength

To improve subject fidelity, we reduced the prior-preservation weight from

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{instance}} + \lambda \mathcal{L}_{\text{class}}, \quad \lambda = 1.0$$

to

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{instance}} + 0.5\, \mathcal{L}_{\text{class}}.$$

Here, $\mathcal{L}_{\text{instance}}$ encourages learning the specific target object (the monster toy), while $\mathcal{L}_{\text{class}}$ preserves generalization within the broader category (toys). Reducing λ decreases the influence of the class prior, allowing the model to allocate more capacity toward memorizing distinctive features of the toy.

6.3 Training Configuration

```
INSTANCE_PROMPT = "a photo of sks monster toy"
CLASS_PROMPT    = "a photo of monster toy"
MAX_TRAIN_STEPS = 1200
LEARNING_RATE   = 1e-6
TRAIN_BATCH_SIZE = 1
GRADIENT_ACCUMULATION_STEPS = 2
PRIOR_LOSS_WEIGHT = 0.5
NUM_CLASS_IMAGES = 150
RESOLUTION = 512
EVALUATION_STEPS = 150
SAVE_CHECKPOINT_STEPS = 150
TRAIN_TEXT_ENCODER = true
```

6.4 Evaluation Results

We evaluated performance using the same three axes as before: identity fidelity, prompt adherence, and diversity. Metrics were computed across 10 prompts $\times$ 10 seeds. Identity similarity was measured using CLIP cosine similarity against both base generations and reference images. Prompt adherence was measured using CLIP text–image similarity. Diversity was estimated using LPIPS (AlexNet backbone) across seeds.

Table 4: Detailed metrics for monster toy training (1200 steps, 6 training images).

Metric	Mean	Std
Identity vs Base (trained)	0.395	0.202
Identity vs Reference (trained)	0.693	0.110
Prompt Adherence (base)	0.288	0.023
Prompt Adherence (trained)	0.291	0.025
Diversity (LPIPS, trained)	0.699	0.048

6.5 Final Results

The model successfully balanced prompt adherence and subject memorization after reducing prior loss weight. While identity similarity to reference images (0.693) was lower than in the human face case, this is expected given the object's unusual features and small training dataset. Importantly, diversity remained high (≈ 0.699), indicating the model retained generative variability across seeds.

Figure 19: Final generations for ten diverse prompts (same as the ones used for the previous images) after 1200 steps of DreamBooth training with adjusted prior loss weight. The model demonstrates robust generalization across contexts while preserving the monster toy's identity.

6.6 Training Logs

We monitored training dynamics throughout the full 1200 steps. Figure 20 shows the loss curve and learning rate schedule. A gradual decline in loss confirms stable convergence, while the flat learning rate (set to 1×10^{-6}) ensured controlled updates without instability.

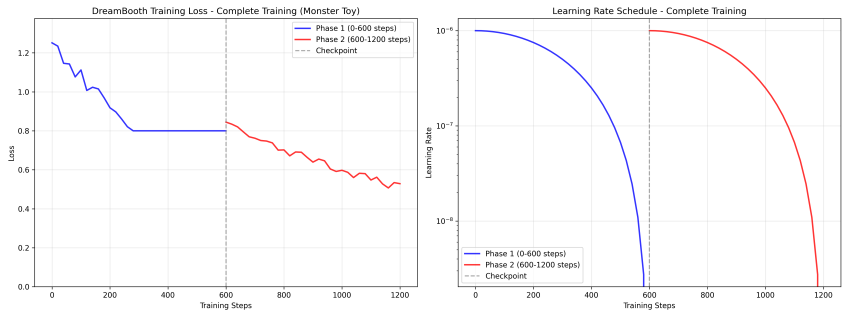

Figure 20: Training logs for 1200 steps: loss and learning rate schedule.

7 Applications

DreamBooth supports recontextualization, style transfer, novel view synthesis, property edits, expression manipulation, and accessorization. Persistent identity enables, e.g., comics or creative assets featuring the same character across frames while preserving distinguishing attributes.

8 Limitations and Societal Impact

Observed failure modes include rare-context prompts, context–appearance entanglement, and overfitting to training settings when prompts are too close to the reference context. As with all realistic generative imagery, responsible use is essential given the potential for misuse.

A Additional Grids

All follow the same left-to-right layout (target reference, base diffusion, overfit, underfit, balanced).

Figure 21: Street photo with sunglasses.

Figure 22: Accessorization: red bandana.

Figure 23: Beach at sunset.

Figure 24: Running in a park.