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

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.

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 7 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.

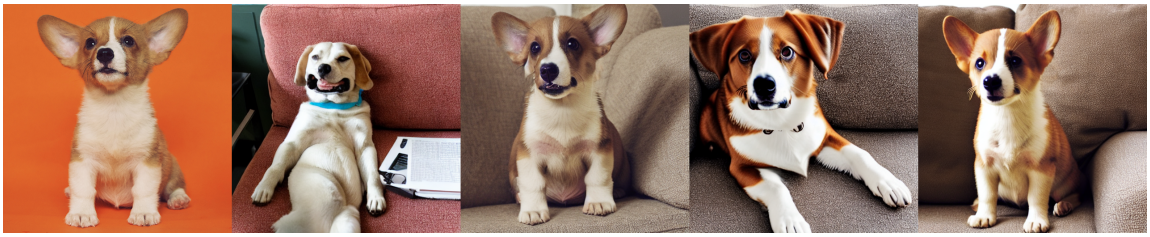

Figure 1: Portrait photo of the subject across training regimes. Overfit maximizes identity but sacrifices diversity and controllability; Underfit improves diversity and prompt adherence but weakens identity; Balanced offers the best overall trade-off. Figure shows from left to right: target, base model,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

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

5 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.

6 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 3: Recontextualization: the subject in front of the Eiffel Tower. Balanced maintains both prompt adherence and identity.

Figure 4: Street photo with sunglasses.

Figure 5: Accessorization: red bandana.

Figure 6: Beach at sunset.

Figure 7: Running in a park.