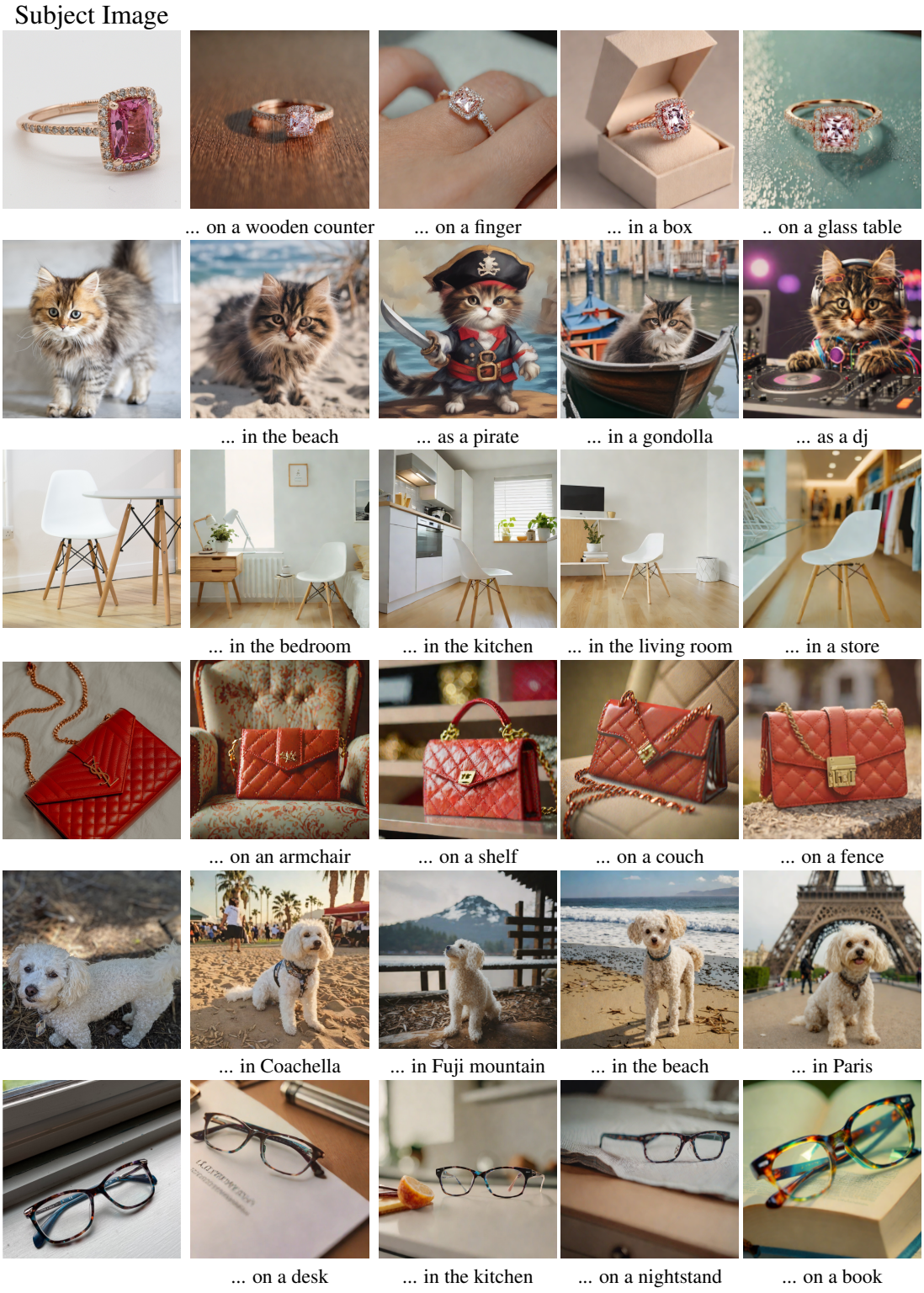

Figure 9. More Qualitative results on Subject Driven Image Generation

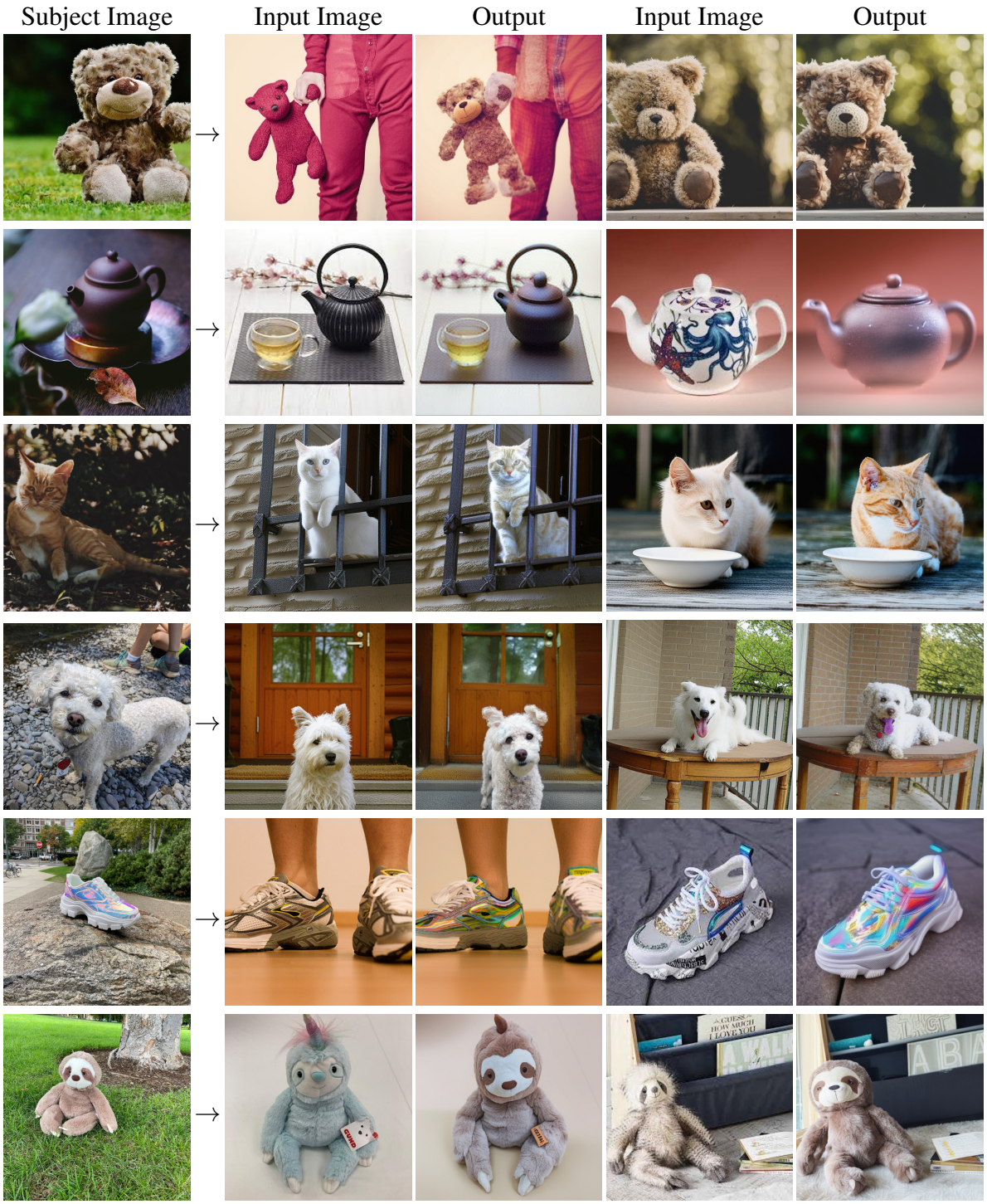

Figure 10. More qualitative results on Subject Driven Image Editing.

In this supplementary material, we present additional experiment results. The supplementary comprises the following subsections:

1. Sec. B, details the inversion method we used for image editing.
2. Sec. C, details about the user study.
3. Sec. D, details about early-stopping method used in our experiments.
4. Sec. E, details about the implementation of the baselines.
5. Sec. G, details about adaptation of SISO to various backbone models.
6. Sec. H, discussion of the limitations of DreamBooth with distilled models

A. Implementation Details

We used SDXL-Turbo [44], the distilled version of SDXL [40]. We set the loss calibration hyperparameters to $a = 1, b = 1, c = 10$, and the learning rate to $\alpha = 3e^{-4}$. The resolution in all our experiments is 512×512. For further implementation details using various backbone models, refer to Section G.

B. Diffusion Inversion

We employ ReNoise for diffusion inversion in our image editing solution. ReNoise hyperparameters include strength, calibrating noise addition, balancing reconstruction, and editability. High values harm reconstruction while improving the ability to edit, and low values hinder object changes but improve reconstruction. We tuned the default setting from 1 to 0.75 in all experiments. Although this setting slightly reduces editing potential, subject-driven editing demands changes to the subject, not the background. Thus, this value empirically proved optimal for both reconstruction and subject editing without altering the background.

C. User Study

According to the task, workers in Amazon MTurk were presented with a subject image, a condition (a prompt or an input image), and two generated images - one from SISO and the other from the baseline. The study was conducted on 100 images from the benchmark, with five workers rating each image. The method used for the study was a two-alternative forced choice, where raters must choose the preferred output between two options. In our case, the workers were presented with three questions per image. Each question requested the worker to choose between two generated images (the order of the generated images was randomly picked). For subject-driven image generation, the questions tested the following criteria: (i) object similarity (what we refer to in the paper as identity preservation), (ii) prompt alignment (what we refer to as prompt adherence), and (iii) naturalness. See Fig. 13 for an illustration of the user study interface.

D. Early Stopping

SISO generates a well-formed image at each iteration, rather than a noisy latent. This enables using the method in an interactive manner. One option is to display images from all iterations and stop the optimization process when a satisfactory result is obtained. To achieve a fully automated process, we used a simple early-stopping strategy, where the process ends if the loss has not improved by x percent on the last n iterations. Specifically, we set $x = 3$ and $n = 7$ in all of our experiments, both for generation and editing.

E. Baselines

Here, we describe how we implemented the baselines used in the paper.

Subject-driven image generation. We compared our method against three baselines: (i) DreamBooth, which fine-tunes the diffusion model parameters according to a set of reference images. We used the code given in Diffusers [54] library for all different base models (SDXL, FLUX, and Sana). (ii) AttnDreamBooth, which improves on DreamBooth with a three-stage process, optimizing a textual embedding, cross-attention layers, and the U-Net. (iii) ClassDiffusion, which utilizes a semantic preservation loss. For both AttnDreamBooth and ClassDiffusion, we used the official implementation published by the authors, using their default hyperparameters.

Subject-driven image editing. We compared our method with two baselines: (i) SwapAnything, which employs masked latent blending and appearance adaptation. (ii) TIGIC, a training-free technique that uses an attention-blending strategy during the denoising process. TIGIC was initially designed for a subject insertion, where the user wants to insert the subject into an empty area in the input image. To adapt to the subject replacement task, we used a state-of-the-art inpainting model (LaMa[1]) to remove the original object and then applied TIGIC. For both methods, we used the official implementation published by the authors, using their default hyperparameters.

F. Evaluation Metrics

We evaluate SISO using four standard quality metrics in the field and introduce one new metric to assess diversity.

Identity Preservation. We follow the protocol of [42] and compute the mean cosine similarity between the generated image and five real images of the subject from the

[1] https://github.com/advimman/lama

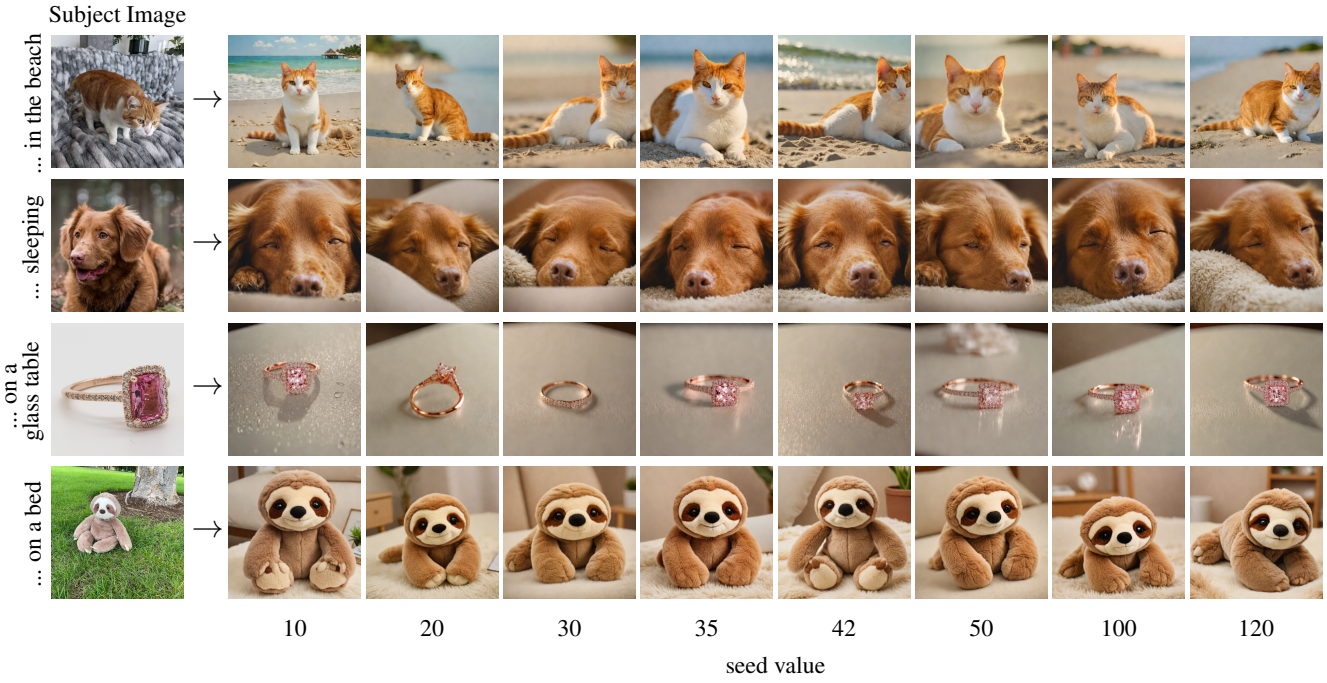

Figure 11. We show the stability of our method across eight seeds for Subject Driven Image Generation.

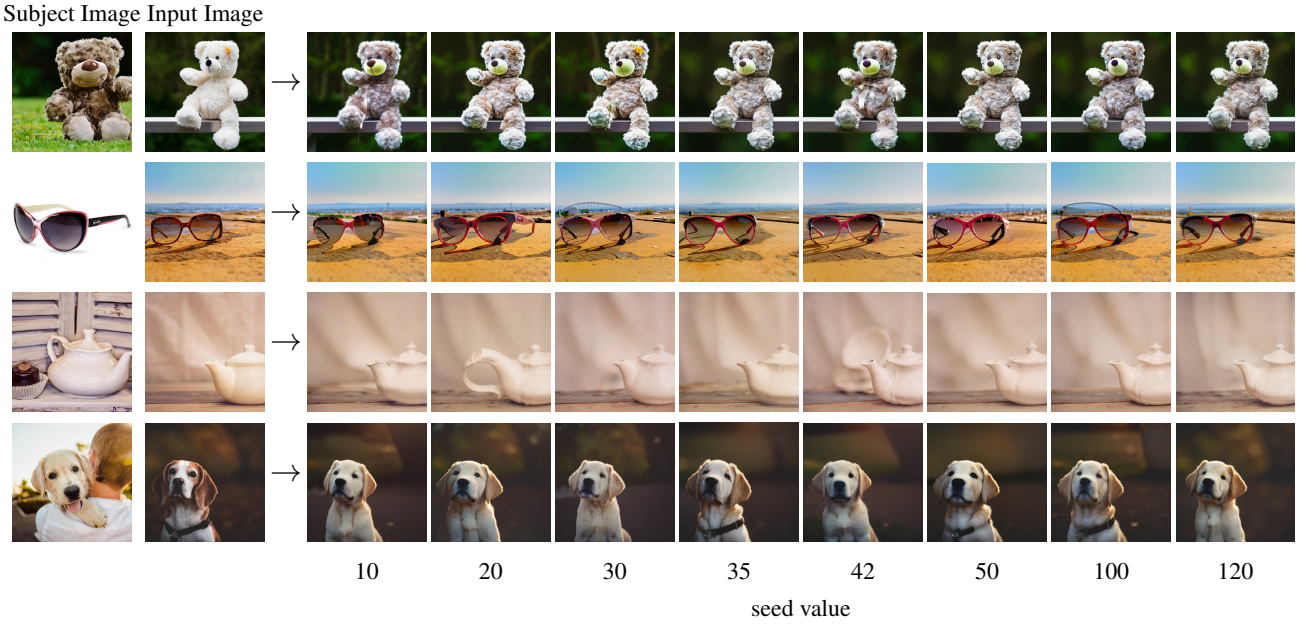

Figure 12. We show the stability of our method across eight seeds for Subject Driven Image Editing.

DreamBooth dataset, using DINO (effective for instance-level similarity) and IR features (effective for item-level similarity) [46, 58].

Diversity. Single-image concept learning often leads to overfitting, causing the model to generate images that closely resemble the reference. This inflates identity scores and masks poor generalization. To reveal this issue, we assess diversity by computing the mean squared error (MSE) between images generated from three different seeds.

Naturalness. Following [34, 52], we evaluate the real-

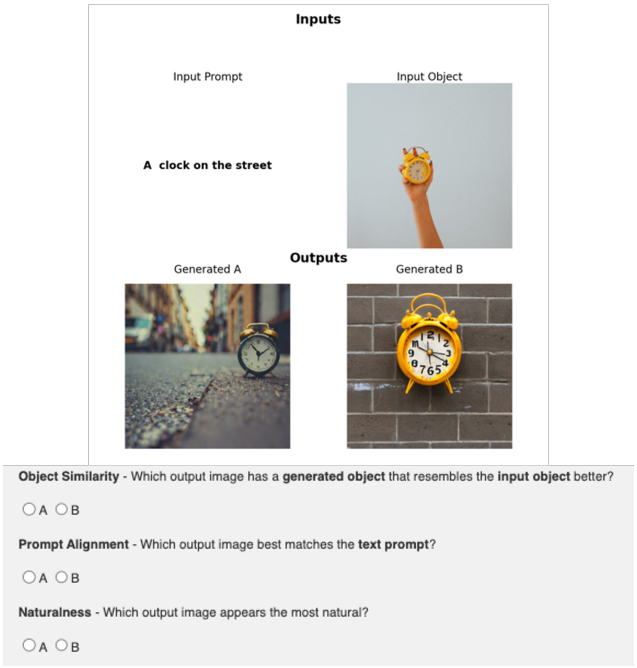

Figure 13. Illustration of the user study interface for the Subject-driven image generation task.

ism of generated images using FID [19] and KID [2]. We also compute CMMD [23] for semantically richer CLIP-based evaluation, as used in [4].

Prompt Adherence. Following [11, 42], we measure alignment between the generated image and input prompt using CLIP-T.

Background Preservation. For image editing, we compute LPIPS [68], which measures perceptual similarity between images [34]. We exclude edited regions by masking with Grounding DINO and SAM [25] before computing LPIPS.

G. Adaptation to Various Backbone Models

A key advantage of SISO is its ability to be used with different backbone models with limited adaptation. In this section, we will describe the main differences in implementation between the different backbones we used (SDXL-Turbo, FLUX schnell, Sana). First, SDXL-Turbo and FLUX schnell are distilled versions, meaning that they generate images using a small number of steps (1-4). Sana, on the other hand, does not have a distilled version and requires 20 steps to generate a high-quality image. We found that when using distilled versions, backpropagating through the final denoising step is sufficient. However, when using a non-distilled version, like Sana, it may be beneficial to backpropagate through more than one denoising step. Specifically, we set the number of steps to backpropagate through to 3. Also, even when using a distilled version, the number of denoising steps used in each iteration may be important, and different models behave differently in this context. We will denote this number as t. SDXL-Turbo is less noisy for different values of t, but FLUX schnell showed a significant difference when using various values of t. More specifically, setting $t >= 2$ resulted in low-quality generated images, even when trying to backpropagate through more denoising steps (see Fig. 14). However, FLUX schnell generates blurred images when used with one denoising step. A naive approach to overcome the blurriness is to use a model trained for upscaling resolution. However, this requires loading another model, which may complicate the process. We solved the issue using the training simplification (Sec. 3.3 in the paper). Although the weights were optimized using $t = 1$, they can be used in inference with different values of t, thus producing high-quality images.

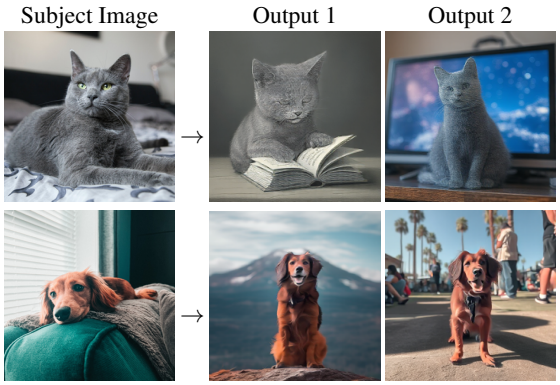

Figure 14. Optimizing on FLUX schnell using four denoising steps results in low quality images.

H. DreamBooth limitation on Distilled Models

In our work, we observed that applying DreamBooth during a fine-tuning phase, by continuing the diffusion model's noise prediction objective with MSE, often fails when only a small number of denoising steps are used for learning the new objective, rather than directly copy-pasting. We note that even when three reference images are provided, the resulting diversity remains low (see Tab. 6).

What we find is that the copy-paste mechanism behaves adaptively depending on the scenario. For example, when the subject is a girl holding a bag, the learned representation captures both the girl and the bag; this becomes evident when generating an image of the subject in Paris, where the girl fits into the context more naturally. However, in other cases, the model relies on the background information from the reference image. For instance, when trained with an image of the bag in a natural setting, it tends to produce results that emphasize nature when prompted for natural places (see Fig. 15).

	DINO↑	IR↑	CLIP-T↑	MSE↑
DreamBooth	**0.62**	**0.65**	0.31	0.07
SISO (SDXL-Turbo)	0.52	0.60	0.31	**0.11**

Table 6. DreamBooth vs. SISO with three reference images on SDXL-Turbo. DreamBooth's stronger identity scores coincide with low MSE, consistent with copy-like behavior. Low MSE indicates near-reconstruction (copying) of the reference.

Figure 15. Qualitative results for subject-driven image generation with DreamBooth three reference images and distilled model.