REGAIN: RESTORING SUBJECT FIDELITY IN PERSONALIZATION ON SYNTHETIC IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image diffusion models are personalized to a subject by DreamBooth fine-tuning on a handful of its images. Increasingly, these images come from a diffusion model rather than a camera. We show that fine-tuning on such synthetic images degrades subject fidelity, producing oversaturated color and excess high-frequency detail. To isolate the cause, we fine-tune two models from the same base model with the same DreamBooth recipe, one on real photos of a subject and one on synthetic images of that subject generated by the first. We trace the degradation to classifier-free guidance (CFG). For the model personalized on synthetic images, the angle between the conditional and unconditional noise predictions, and with it the norm of their difference, is much larger than for the model personalized on real photos. This inflation grows toward high frequencies and also appears at other prompts semantically close to the subject, such as its class noun, but not at unrelated ones. We propose ReGain, a training-free correction applied at sampling time that measures how much each frequency band of the guidance is inflated relative to the base model and scales that band down accordingly. ReGain needs no real photos. On Stable Diffusion v1.5, ReGain closes 51–64% of the subject-fidelity gap to the model personalized on real photos, as measured by DINO, DINOv2 and CLIP-I. It also improves subject fidelity on SDXL and SD 3.5 and preserves text alignment on all three backbones.

1 INTRODUCTION

Text-to-image diffusion models (Rombach et al., 2022) are personalized to a user's subject so that text prompts can place it in novel scenes on demand. DreamBooth style (Ruiz et al., 2023) personalization methods fine-tune a diffusion model on a handful of subject images to associate it with a unique identifier. Increasingly, however, the subject images themselves come from a generative model rather than a camera, e.g. as edited variants of a real subject.

We study this change of data provenance under controlled conditions: M_{real} is personalized with DreamBooth (Ruiz et al., 2023) on real photographs, and M_{syn} is fine-tuned from the same base with an identical recipe on images generated by M_{real}, so any difference in their outputs is due to the training data alone. With the same prompt and seeds, M_{real} produces a clean image, while M_{syn}'s is oversaturated and overloaded with high-frequency texture as seen in Fig. 1 (left). Its output latents carry more power at every frequency, and the excess grows in the high bands (Fig. 1, top right). We trace this back to guidance. Classifier-free guidance (CFG) (Ho & Salimans, 2022) steers each denoising step along $\Delta = \epsilon_c - \epsilon_\varnothing$, and for M_{syn}, $\|\Delta\|$ is inflated throughout denoising, with a gap that widens as sampling proceeds (Fig. 1, bottom right).

Findings. The Δ inflation has four properties. *(1) Angle-driven.* The norms of ϵ_c and $\epsilon_\varnothing$ do not grow, but the angle between them exceeds M_{real}'s at every step so their vector difference $\|\Delta\|$ inflates*(2) Semantic.* Prompts without the identifier, using only the subject's class name (e.g., "a dog"), show similar inflation, and so do related classes (e.g. cats) never seen in fine-tuning. Unrelated prompts show none. *(3) A property of the model.* The inflation also appears away from M_{syn}'s sampling trajectories, when it denoises noisy copies of its own training images. *(4) High-frequency and time-varying.* High-frequency bands inflate several-fold over the early-to-middle trajectory, far more than low-frequency bands. Each band follows its own time profile. A lower guidance scale acts uniformly across bands and steps, so it cannot undo this.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

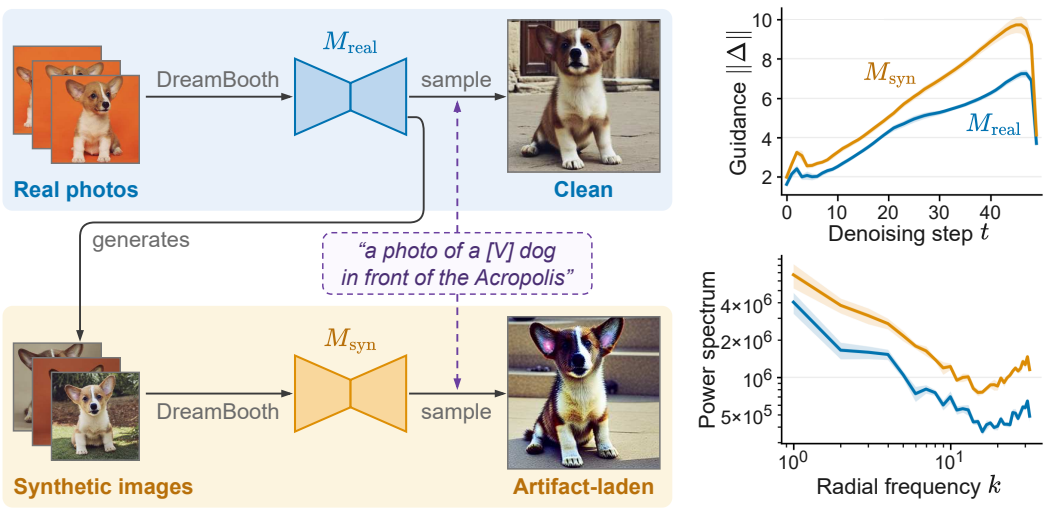

Figure 1: **Personalizing a model on synthetic images of a subject degrades its fidelity.** Left: M_{real} is DreamBooth fine-tuned from the base model M_{base} on real photographs of the subject, and M_{syn} is fine-tuned from the same base with an identical recipe on images generated by M_{real}. Both are sampled with the same prompt and the same 10 seeds. Right: compared to M_{real}, M_{syn} shows an inflated per-step $\|\Delta\|$ across denoising steps and an inflated latent power spectrum in its generated images.

Method. Based on these findings we propose ReGain, a training-free correction applied at sampling time, so it also repairs already-trained checkpoints. Correcting the inflation requires a reference for how large Δ should be. M_{real} could help, but it is available only in our analysis experiments, not in practical settings. The base model that M_{syn} was fine-tuned from, however, is freely available, and its Δ for the subject's class is uninflated. Because the inflation appears away from sampling (finding 3), ReGain measures it once, before sampling, it runs M_{syn} and the base model on noisy copies of M_{syn}'s training images and records how much larger M_{syn}'s Δ is in each frequency band and denoising step. At sampling time, it lowers the guidance of each band and step by that factor (findings 1 and 4), where the subject tokens govern the prediction (finding 2), with plain CFG elsewhere.

Contributions.

- We show that DreamBooth fine-tuning on synthetic subject images inflates the CFG guidance term Δ, and characterize this inflation in four findings.
- Using these findings, we propose ReGain, a training-free correction that lowers the guidance per frequency band and denoising step, using the public base model as the reference. It needs no real photos, retraining or per-subject tuning.
- We evaluate ReGain on the 30 subjects of the DreamBooth dataset with DreamBooth and DreamBooth-LoRA on SD 1.5 and with DreamBooth-LoRA on SDXL and SD 3.5. In all four settings, ReGain significantly improves subject fidelity (DINO, DINOv2 and CLIP-I) while preserving prompt fidelity (CLIP-T). On SD 1.5, it closes 51–64% of the subject-fidelity gap to a model personalized on real photos.

2 Related Work

Subject personalization. Perosnalization methods may bind a subject to a unique identifier by training on a handful of its images (Ruiz et al., 2023; Gal et al., 2023), updating all weights, selected layers (Kumari et al., 2023), or low-rank adapters (Hu et al., 2022; Shah et al., 2024), optionally with drift regularization (Lee et al., 2024; Kim et al., 2026). Tuning-free methods condition on the subject images through encoders or adapters (Ye et al., 2023; Tan et al., 2025; Wu et al., 2025) but reproduce fine identity detail less faithfully, so fine-tuning remains the choice when fidelity matters.

Training on synthetic data. Subject images used for personalization are increasingly produced by generative models (Avrahami et al., 2024; Kumari et al., 2025; Li et al., 2026). Generative models

retrained on their own outputs over several generations lose diversity and fidelity (Shumailov et al., 2024; Alemohammad et al., 2024; Bertrand et al., 2024), and proposed remedies mix real data back into training (Gerstgrasser et al., 2024) or guide sampling away from a model trained on self-generated data (Alemohammad et al., 2025). Yoon et al. (2025) identify the CFG guidance scale as a driver of collapse over many generations of training. In contrast to these works, our work studies (1) a single round of DreamBooth fine-tuning, not generic training (without a subject token) of a chain of models, and (2) in our case we show that the CFG guidance term (not the scale) itself is miscalibrated as the angle between conditional and unconditional predictions increases, inflating guidance in a band-specific way, and correct it without training.

Guidance analyses and fixes. Classifier-free guidance (Ho & Salimans, 2022) is the standard steering mechanism of text-to-image diffusion (Rombach et al., 2022), and many works adjust when, where, or how strongly it is applied (Kynkäänniemi et al., 2024; Sadat et al., 2024; Lin et al., 2024; Sadat et al., 2025a; Karras et al., 2024; Shen et al., 2024; Chung et al., 2025), including per frequency band (Sadat et al., 2025b) and for personalized models (Chan et al., 2024; Park et al., 2025; Jeong & Kim, 2025). These works do not consider the inflation of the guidance term itself caused by model personalization on synthetic images, which ReGain corrects before any of them are applied.

3 METHOD

3.1 SETUP AND NOTATION

To isolate what synthetic training data does to the guidance, we compare two models trained from the same base model M_{base} by the DreamBooth objective (Ruiz et al., 2023) (Figure 3A). The first, M_{real}, is fine-tuned on real photographs of a subject, binding it to a unique identifier [V]. As a running example we use subject dog6 of the DreamBooth dataset on SD 1.5. Its subject prompt c_{subj}, *"a photo of a [V] dog"*, names the subject through [V], and its class prompt c_{class}, *"a photo of a dog"*, omits it. The second model, M_{syn}, is fine-tuned by the same recipe on synthetic images of the subject generated by M_{real}. Because the two models share the base weights and the training recipe, any difference in their guidance traces to the training images alone. The two training sets are comparably diverse (see Appendix A.1), so what distinguishes them is the origin of their images . In practice, only M_{syn}, its training images, and M_{base} are available. M_{real} is used solely for the analysis in Section 3.2 and for evaluation.

At step t (counted from the noisiest state, $t = 0$), the noisy latent z_t has signal and noise variances $\bar{\alpha}_t$ and $1 - \bar{\alpha}_t$. Classifier-free guidance (Ho & Salimans, 2022) combines predictions under the prompt c and the null prompt $\varnothing$:

$$\epsilon_w \;=\; \epsilon_\varnothing + w\,(\epsilon_c - \epsilon_\varnothing), \qquad \Delta \;=\; \epsilon_c - \epsilon_\varnothing, \tag{1}$$

with guidance weight w. Superscripts denote the source model, e.g. $\Delta^{M_{\text{real}}}$ is the guidance of M_{real}.

3.2 WHY FINE-TUNING ON SYNTHETIC IMAGES INFLATES GUIDANCE

We compare the noise predictions of M_{syn} and M_{real} at every step as both sample c_{subj} from the same 10 initial noises with the CFG of equation 1. The guidance term Δ of M_{syn} is systematically larger in norm than that of M_{real}, an excess we call the Δ inflation. Its norm is controlled by three quantities: the unconditional norm $r = \|\epsilon_\varnothing\|$, the ratio $\delta = \|\epsilon_c\|/\|\epsilon_\varnothing\|$, and the angle θ between $\epsilon_\varnothing$ and ϵ_c:

$$\|\Delta\|^2 \;=\; r^2\big(1 + \delta^2 - 2\delta\cos\theta\big), \qquad \theta = \arccos\big(\langle\epsilon_\varnothing, \epsilon_c\rangle \,/\, \|\epsilon_\varnothing\|\|\epsilon_c\|\big). \tag{2}$$

We report these measurements on one DreamBooth subject, dog6, and repeat them on all 30 subjects in Appendix A.4.

(1) The Δ inflation is angle-driven. Of the three scalars, only the angle raises $\|\Delta\|$ as seen in Figure 2 (a). $\theta^{M_{\text{syn}}}$ exceeds $\theta^{M_{\text{real}}}$ at every step, by 37% on average over the 50 steps. δ stays within 1% of unity for both models, and $r^{M_{\text{syn}}}$ stays within 3% of $r^{M_{\text{real}}}$ until step 30 and then falls below it, which works against the inflation. Rewriting equation 2 as $\|\Delta\|^2 = r^2\big[(1-\delta)^2 + 4\delta\sin^2(\theta/2)\big]$, $\delta \approx 1$ gives $\|\Delta\| \approx 2r\sin(\theta/2) \approx r\theta$, so the guidance norm follows the angle: $\|\Delta^{M_{\text{syn}}}\|$ is larger at every step, by 29% on average. Since CFG adds $w\Delta$ to $\epsilon_\varnothing$, M_{syn} is over-guided at the same w.

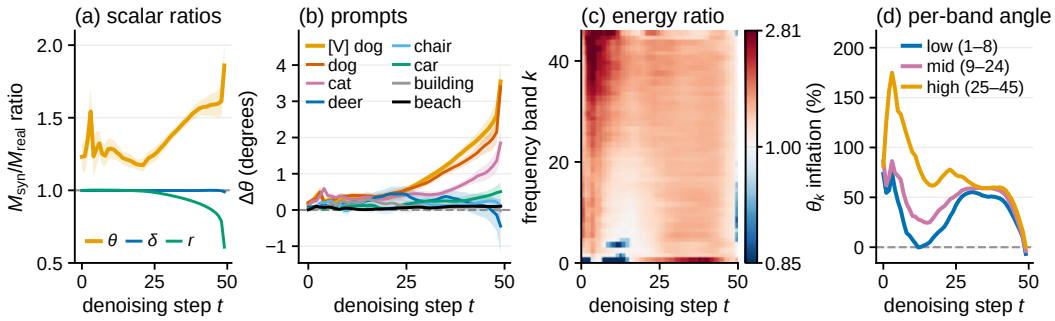

Figure 2: (a)$M_{\text{syn}}/M_{\text{real}}$ ratio of θ, δ and r. (b) Angle excess $\theta^{M_{\text{syn}}}(t) - \theta^{M_{\text{real}}}(t)$ for the subject prompt and increasingly semantically distant prompts. (c) $M_{\text{syn}}/M_{\text{real}}$ band-energy ratio of Δ per frequency band k and step t; red marks inflation. (d) θ inflation of the low, mid and high bands over steps. 10 seeds mean.

(2) The Δ inflation follows the subject's semantic neighborhood. In prompt space, the angle excess $\theta^{M_{\text{syn}}}(t) - \theta^{M_{\text{real}}}(t)$ (mean over seeds) for prompts at increasing semantic distance from the identifier splits into two tiers (Figure 2b). The class noun (*"a photo of a dog"*) inherits nearly the full inflation, and *cat*, an animal never seen in fine-tuning, about half of it. Prompts farther away (*deer*, *chair*, *car*, *building*, *beach*) show little to no inflation until the final steps.

(3) The Δ inflation persists when both models see the same input. Findings 1 and 2 compare the two models on their own sampling trajectories, which differ. Toward a fix, we test whether the inflation persists when both models are evaluated at the same inputs, namely forward-noised training images of M_{syn}, $z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, n$, with z_0 the latent encoding of a training image and $n \sim \mathcal{N}(0, I)$ drawn afresh at every step. There, where M_{syn} should be most faithful, the guidance term of M_{syn} still exceeds that of M_{real} in band energy. Retraining M_{syn} on images sampled from M_{real} at lower guidance weights leaves the inflation in place (dog6, Appendix A.2). Adding the prior-preservation loss of DreamBooth to fine-tuning does not remove it either (Appendix A.3).

(4) The Δ inflation concentrates in high frequencies and varies over time. We next ask whether the inflation is uniform across frequencies and steps, since neural networks learn different frequencies at different rates (Rahaman et al., 2019). We split Δ into K frequency bands, from the constant component ($k = 0$) to the finest detail ($k = K - 1$), with $B_k(\Delta)$ denoting the component of Δ in band k. For analysis we group the bands into low, mid and high (boundaries in Appendix C). Both models are evaluated at the states M_{syn} visits during sampling. To measure the inflation in each band, we compare the band energies of the two guidance terms, averaged over seeds, as $\sqrt{\|B_k(\Delta^{M_{\text{syn}}})\|^2 / \|B_k(\Delta^{M_{\text{real}}})\|^2}$. Figure 2c shows this ratio for every band and step. The inflation grows from the low to the high bands and peaks earliest in the high bands as seen in Figure 2 (c, d). Each band has its own time profile, so no single guidance weight describes the inflation.

3.3 ReGain: restoring the guidance from the base model

To correct the inflation at sampling time, ReGain needs a reference for how large Δ should be at each band and step. We take this reference from the base model M_{base}, which never saw the synthetic images. Concretely, ReGain has two parts as shown in Figure 3C : (1) a one-time calibration, which estimates how much M_{syn}'s guidance exceeds M_{base}'s in each band and step, as a gain $g(k, t)$, and quantizes it into a compact schedule; and (2) a correction at sampling time, which rescales each band of Δ by the scheduled gain inside the subject mask. Neither part involves training.

Measuring the Δ inflation against the base model. To measure how much M_{syn}'s guidance exceeds that of the base model, we evaluate both at the forward-noised training images of finding 3 with the same noise draws: M_{base} under c_{class}, since it has never seen the identifier, and M_{syn} under c_{subj}. At each band k and step t, the masked band energy of the guidance term is given by

$$e(k, t) \;=\; \frac{\left\| \, m \odot B_k(\Delta) \, \right\|^2}{|m|\, C}, \tag{3}$$

where m is a binary subject mask and $|m|$ and C are the numbers of masked latent pixels and latent channels. The mask is read from the attention maps of M_{syn} at the same state, by the segmenta-

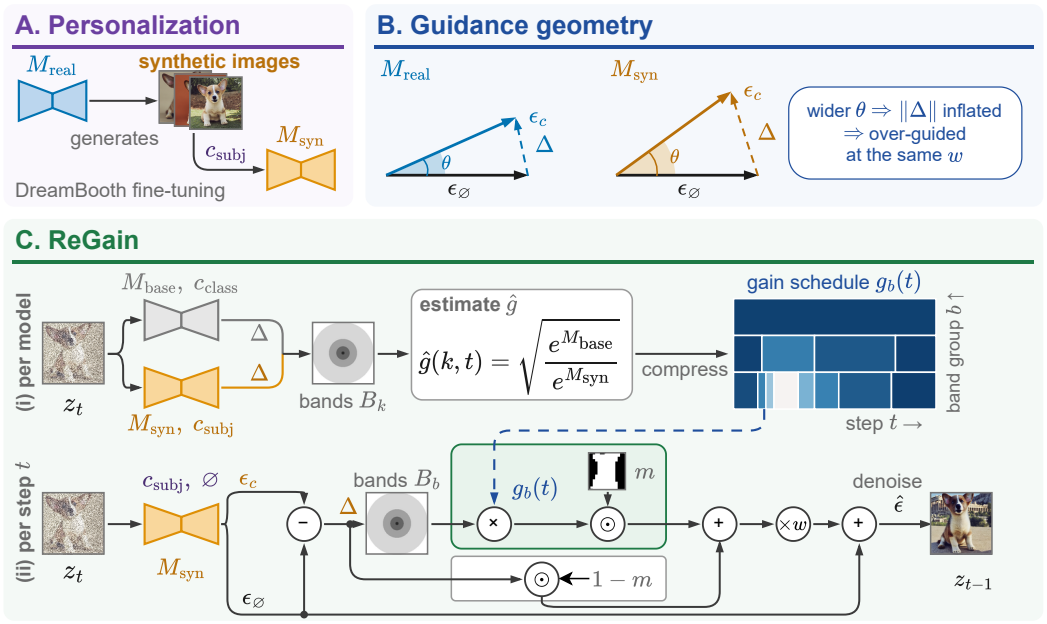

Figure 3: **Overview.** (A) M_{syn} is DreamBooth fine-tuned on synthetic subject images generated by M_{real}. (B) The angle θ between the unconditional and conditional predictions is wider for M_{syn}, so its guidance term $\Delta = \epsilon_c - \epsilon_\varnothing$ is inflated. (C) ReGain. (i) Once per model, M_{base} and M_{syn} are evaluated at forward-noised training images z_t; the square root of the ratio of their masked band energies e gives the gain $\hat{g}(k,t)$, compressed into the schedule $g_b(t)$ (darker blue attenuates more; Figure 4). (ii) At each step, Δ is rescaled per band group by $g_b(t)$ inside the subject mask m (green) and left unchanged outside it (grey), then scaled by w and added to $\epsilon_\varnothing$, equation 5.

tion stage of S-CFG (Shen et al., 2024) on U-Net models and by Seg4Diff (Kim et al., 2025) on transformer models, and is shared by both models so that their energies are compared over the same pixels. Appendix B details its construction. We define the gain as the square root of the ratio of their means,

$$\hat{g}(k,t) \;=\; \sqrt{\mathbb{E}\big[e^{M_{\text{base}}}(k,t)\big] \;/\; \mathbb{E}\big[e^{M_{\text{syn}}}(k,t)\big]}\,. \tag{4}$$

The gain schedule. The estimated gain $\hat{g}(k,t)$ has one entry per band and step, each computed from only a few training images and noise draws, so it is noisy at this resolution. To average out this noise and obtain a compact schedule, we compress it into a small number of rectangular cells of constant gain over contiguous bands and steps. The compression proceeds in two stages: the bands are first cut into the fewest contiguous band groups b, and the steps of each group are then cut into the fewest contiguous segments, such that at each stage the cells deviate from $\hat{g}$ by at most a tolerance τ in root mean square. Each stage is solved exactly by a one-dimensional dynamic program on the squared error (tolerance in Section 4.1). The DC band is exempt from this compression and keeps its measured gain. Figure 4 shows the resulting schedule for the running dog6 example.

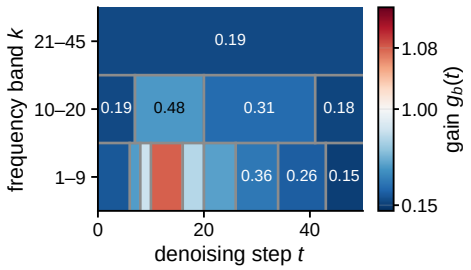

Figure 4: Gain schedule $g_b(t)$ for dog6 from the base model (equation 4), one row per band group b.

Correcting at sampling time. Because the inflation varies across bands (finding 4) and is tied to the subject (finding 2), the correction rescales each band group separately and acts only where the subject tokens govern the prediction, a choice the ablation of Table 3 validates. The bands partition the spectrum, so $\Delta = \sum_b B_b(\Delta)$ exactly for any grouping of the K bands into contiguous band groups b, with $B_b = \sum_{k \in b} B_k$. Rescaling each group by a gain $g_b(t)$ and applying the result inside

a binary subject mask m, with plain CFG outside, gives the corrected prediction at guidance weight w:

$$\hat{\epsilon} \;=\; \epsilon_{\varnothing} \;+\; w\, m \odot \sum_{b} g_b(t)\, B_b(\Delta) \;+\; w\,(1-m) \odot \Delta, \tag{5}$$

where $\odot$ is the elementwise product. The band projections $B_b(\Delta)$ are computed on the full guidance term, and the mask is applied to the result, so that each band is rescaled globally before the spatial split. At sampling, m is read at each step from M_{syn} under the sampled prompt, whose scene words serve as the background tokens; Appendix B details both constructions. When every $g_b = 1$, equation 5 reduces to plain CFG.

4 Experiments

4.1 Setup

Dataset. We evaluate on the DreamBooth dataset (Ruiz et al., 2023), 30 subjects with 4–6 real photographs each, 25 evaluation prompts per subject, 3 seeds, and one image per prompt and seed.

Baselines. For every subject we build the M_{real}–M_{syn} pair of Section 3.1, with M_{syn} fine-tuned on five images generated by M_{real}, in four settings: DreamBooth and DreamBooth-LoRA (Hu et al., 2022) on Stable Diffusion v1.5 (SD1.5) (Rombach et al., 2022), and DreamBooth-LoRA on SDXL (Podell et al., 2024) and Stable Diffusion 3.5 (SD3.5) (Esser et al., 2024), each over all 30 subjects. In each setting we compare M_{syn} with plain CFG against M_{syn} with ReGain, sampled at the same seed, prompt, sampler and guidance weight w, with the gain schedule estimated once per subject from its five training images, and report M_{real} as the reference. On SD1.5 we also compare with three guidance methods applied to M_{syn}. S-CFG (Shen et al., 2024) rescales the guidance per semantic region, CFG++ (Chung et al., 2025) renoises each step with the unconditional prediction, and FDG (Sadat et al., 2025b) guides the low and high frequencies with separate weights. Each runs with its authors' default settings at the same seed, prompt and number of sampling steps. All training and sampling hyperparameters, including those of the baselines, are listed in Appendix C.

Subject mask. The mask m of Section 3.3 is rebuilt at every sampling step from M_{syn}'s attention under the prompt being sampled. Its construction at calibration, sampling, and visualization are in Appendix B.

Metrics. Subject fidelity is the cosine similarity between embeddings of a generated image and the subject's real photographs, with CLIP ViT-B/32 (CLIP-I), DINO ViT-S/16 (DINO) and DINOv2 ViT-S/14 (DINOv2) (Radford et al., 2021; Caron et al., 2021; Oquab et al., 2024). Prompt fidelity (CLIP-T) is the cosine similarity between the CLIP embeddings of the image and of the prompt with the identifier removed. Over-guidance artifacts are measured by mean saturation, root-mean-square contrast and the high-band energy share defined in Section 4.3.

Implementation details. All models except SD3.5 are sampled with DDIM (Song et al., 2021) for $T = 50$ steps at guidance weight $w = 7.5$, and SD3.5 with its flow-matching Euler sampler for 40 steps at $w = 7.0$. The gain schedule is compressed with tolerance $\tau = 0.05$. The measurement setup of Section 3.2, the frequency bands and the full fine-tuning and sampling settings are in Appendix C.

4.2 Main comparison

Quantitative results. Table 1 reports subject and prompt fidelity in the four settings. Fine-tuning on model-generated images costs subject fidelity on every metric and in every setting, and ReGain recovers a consistent share of that loss at sampling time, with no retraining: more than half of the gap to M_{real} on SD1.5, and between a fifth and a half on SDXL and SD3.5. Prompt fidelity is left intact: CLIP-T changes by at most 0.004 relative to plain CFG in every setting. None of the three guidance baselines closes the gap. FDG recovers at most a sixth of it, none of it on DINO, and lowers CLIP-T. S-CFG and CFG++ score at or below plain CFG on all three subject fidelity metrics. The over-guidance artifacts behind the fidelity loss, which these metrics capture only indirectly, are measured in Section 4.3.

Qualitative results. Figure 5 shows four subjects, each at one prompt and seed shared by the three models. Under plain CFG, M_{syn} renders the subject with oversaturated color, harsh contrast and



Figure 5: Qualitative comparison on SD1.5 with DreamBooth. Each row is one subject, prompt and seed shared by the three models. More subjects and settings are in Appendix E.

Table 1: Subject and prompt fidelity on DreamBooth, averaged over 30 subjects, 25 prompts and 3 seeds (higher is better). The first row is the reference model M_{real}, and every other row samples from M_{syn}. Baselines, on SD1.5 only: S-CFG (Shen et al., 2024), CFG++ (Chung et al., 2025) and FDG (Sadat et al., 2025b). Bold marks the best M_{syn} result. Gap recovered is the part of the drop from M_{real} to M_{syn}, both with CFG, that ReGain wins back. For CLIP-T we report the change from M_{syn} + CFG instead.

	SD1.5, DreamBooth				*SD1.5, DreamBooth-LoRA*			
Method	CLIP-I	DINO	DINOv2	CLIP-T	CLIP-I	DINO	DINOv2	CLIP-T
M_{real} + CFG	0.814	0.681	0.661	0.300	0.791	0.630	0.606	0.306
M_{syn} + CFG	0.788	0.611	0.611	0.289	0.750	0.540	0.510	0.300
M_{syn} + S-CFG	0.781	0.605	0.604	**0.292**	0.744	0.536	0.503	0.302
M_{syn} + CFG++	0.788	0.606	0.608	0.289	0.748	0.535	0.508	0.299
M_{syn} + FDG	0.792	0.608	0.614	0.284	0.755	0.538	0.517	0.297
M_{syn} + CFG + ReGain	**0.802**	**0.647**	**0.643**	0.290	**0.775**	**0.594**	**0.568**	**0.304**
↪ Gap recovered	54%	51%	64%	+0.001	61%	60%	60%	+0.004
	SDXL, DreamBooth-LoRA				*SD3.5, DreamBooth-LoRA*			
M_{real} + CFG	0.757	0.558	0.535	0.289	0.811	0.682	0.650	0.311
M_{syn} + CFG	0.724	0.477	0.467	0.278	0.776	0.586	0.567	0.307
M_{syn} + CFG + ReGain	**0.736**	**0.502**	**0.494**	**0.279**	**0.792**	**0.605**	**0.596**	**0.307**
↪ Gap recovered	36%	31%	40%	+0.001	46%	20%	35%	0.000

excess fine texture, the over-guidance artifacts of Figure 1, and the scene loses detail around it. ReGain, from the same model and seed, brings the subject's color and texture back toward M_{real}'s while keeping the prompt's scene, and the attribute change of the last column (the dog recolored purple) is still carried out. Figure 6 shows the same behavior under DreamBooth-LoRA on SD1.5, on SDXL and on SD3.5.

4.3 Over-guidance artifact metrics

Over-guided samples carry a characteristic artifact signature: excess saturation and contrast and inflated high-frequency content, the known symptoms of over-guidance (Sadat et al., 2025a) and the same signature reported for models trained on their own high-CFG samples (Yoon et al., 2025). We quantify it with three per-image statistics, mean HSV saturation, root-mean-square grayscale contrast and the high-frequency fraction of the image's Fourier power (definitions in Appendix D), and take M_{real}'s value as the target, since the goal is to restore the reference model's image statistics. This section evaluates the SD1.5 DreamBooth setting.

Table 2: Over-guidance artifacts inside the subject region (SD1.5, DreamBooth). Bold marks the M_{syn} row closer to M_{real}.

Method	Saturation	Contrast	High band
M_{real} + CFG	0.301	0.211	0.061
M_{syn} + CFG	0.355	0.258	0.069
M_{syn} + CFG + ReGain	**0.309**	**0.234**	**0.066**
↪ Gap recovered	85%	51%	38%

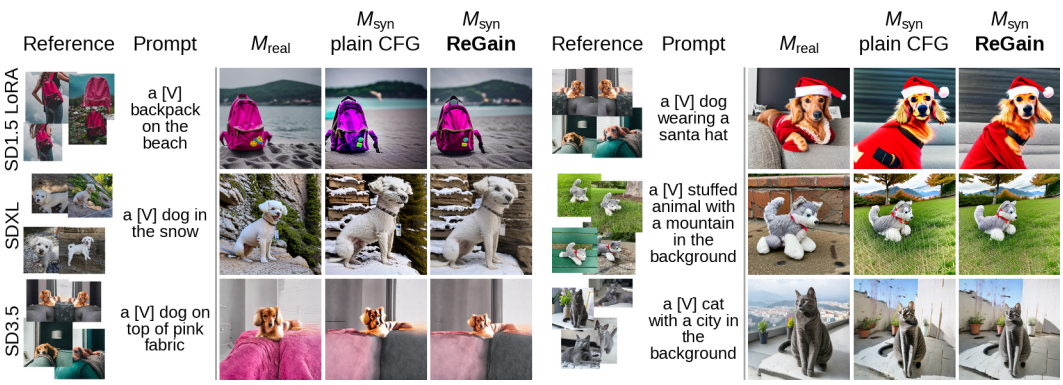

Figure 6: Qualitative comparison across fine-tuning forms and base models. Each row is one base model and each half one subject, prompt and seed shared by the three models; nothing is retuned per base model. The SD3.5 row shows the chain of Table 1.

Table 3: Subject-agnostic alternatives and the subject-mask ablation (SD v1.5, DreamBooth): a lower guidance weight w, APG (Sadat et al., 2025a), and ReGain w/o mask, which applies the gain to every pixel; w=7.5 unless stated. Saturation is inside the subject region; bold marks the M_{syn} row closest to M_{real}.

Method	DINO	CLIP-T	Saturation
M_{real}	0.681	0.300	0.301
M_{syn}, plain CFG	0.611	0.289	0.355
M_{syn}, plain CFG (w=5.0)	0.632	0.286	0.326
M_{syn}, plain CFG (w=3.0)	0.647	0.281	0.292
M_{syn}, APG	0.641	0.283	0.240
M_{syn}, ReGain	0.647	**0.290**	**0.309**
M_{syn}, ReGain w/o mask	**0.652**	0.282	0.293

Table 4: Δ inflation at the calibration states before and after the gain schedule, by band group and compression tolerance (RMS, target 1, mean $\pm$ std over 30 subjects).

		After, by tolerance			
Band	Before	0.02	**0.05**	0.1	0.2
cells		70	18	6	2
Low k	4.70 $\pm$ 1.19	1.02	**1.06**	1.13	1.29
Mid k	4.95 $\pm$ 1.19	1.01	**1.05**	1.16	1.35
High k	5.76 $\pm$ 1.38	1.03	**1.11**	1.30	1.57
All	5.31 $\pm$ 1.26	1.02	**1.08**	1.22	1.45

Where they are measured. The correction acts inside the subject mask and leaves plain CFG outside it. We therefore report the three statistics inside the mask that steered M_{syn}, over identical pixels for all three methods (Appendix D), and the full-frame values in the text.

Results. Table 2 reports the three statistics inside the subject region. Under plain CFG, M_{syn} is 18% over-saturated, 22% over-contrasted and carries 13% excess high-band power relative to M_{real}: the over-guidance artifacts of Figure 1, and the image-space counterpart of the spectral excess in Figure 2c. ReGain moves all three statistics back toward M_{real} and overshoots none of them. Most of the saturation excess is removed, along with about half of the contrast excess and a third of the high-band excess. On the full frame the same recoveries read 77%, 35% and 5%, diluted by the background, which covers most of the frame and is generated under plain CFG in every method.

4.4 Analysis and Ablations

All experiments in this section use DreamBooth on SD1.5 with the 30 subjects, prompts and seeds of Table 1.

Effect of subject-agnostic guidance changes. The inflation is structured in frequency and time and follows the subject tokens (findings 2 and 4, Section 3.2), so no change applied alike to every band and step should repair it. We test two such changes on M_{syn}: lowering the guidance weight to $w \in \{5.0, 3.0\}$, and adaptive projected guidance (APG) (Sadat et al., 2025a), which removes the component of the guidance term parallel to the conditional prediction. As seen in Table 3 and Figure 7, (1) lowering w removes guidance from the scene as well as from the subject: CLIP-T falls by 0.003 at w=5.0 and by 0.008 at w=3.0, the prompt's accessory or setting fades first, and neither weight recovers DINO; (2) APG over-corrects: saturation drops to 0.240, well below M_{real}'s 0.301,



Figure 7: Example generations from the baselines of Table 3, all at the same seed.

the frame is washed out, and CLIP-T falls by 0.006. ReGain is the only method that moves all three metrics toward M_{real} without overshooting any of them.

Effect of the gain schedule. Table 4 measures the excess guidance energy the schedule is built to remove: the square root of the ratio of M_{syn}'s masked band energy equation 3 to the base model's at the same forward-noised training images, the inverse of the gain equation 4, so that values above 1 are inflation. As seen in Table 4, (1) before correction the inflation rises from the low-k to the high-k group, the ordering of the per-band angle inflation of Section 3.2, and (2) after the gain schedule at $\tau = 0.05$ it lies within 10% of 1 on average (residual by step in Appendix C).

Effect of the compression tolerance τ. The tolerance sets how closely the compressed schedule follows the dense gain $\hat{g}$. We vary it over $\tau \in \{0.02, 0.05, 0.1, 0.2\}$. As seen in Table 4, a coarser tolerance leaves more residual inflation, most in the high-k group, where the gain varies fastest over time, and (2) $\tau = 0.05$ is the coarsest tolerance that keeps the inflation within 10%.

Effect of the subject mask. The mask restricts the correction to the subject. We remove it and apply the gain schedule to every pixel, the background included. As seen in Table 3, CLIP-T drops by 0.008 to 0.282, below plain CFG's 0.289: attenuating the background costs prompt fidelity. The mask is what lets ReGain leave prompt fidelity intact while correcting the subject.

Compute overhead. Calibration runs once per fine-tuned model and takes 12 minutes per subject on one RTX A4000 for SD1.5. At sampling, ReGain takes $1.1\times$ the wall-clock time of plain CFG (5.4 vs 4.9 s per image on the same GPU), mostly for building the mask from the attention maps.

5 Limitations

ReGain corrects the inflation at sampling time, using the base model's guidance as an imperfectly calibrated reference, so it cannot fully match M_{real}. A training-time correction during personalization on synthetic images could close this gap. ReGain also relies on a subject mask from existing segmentation methods, so where the mask misses part of the subject or spills onto the background, it attenuates the guidance in the wrong region. In addition, we measure the inflation only at the model's output, in its noise predictions. Mechanistically tracing where it originates inside the network, e.g., in the cross-attention in intermediate activations, is left for future work.

6 Conclusion

We studied how personalizing a text-to-image diffusion model on synthetic rather than real images of a subject affects its outputs. With the base model and the DreamBooth recipe held fixed, the change of training data alone inflates the CFG term as the angle between the conditional and unconditional noise predictions widens. This happens most strongly in high frequencies, with its strength varying at each denoising step. Building on these findings, ReGain measures the inflation once against base model (before personalization) and attenuates the guidance per frequency band and step inside the subject region. It needs no real photographs or retraining, recovers 51–64% of the subject-fidelity gap on SD1.5, improves subject fidelity on SDXL and SD3.5, while preserving prompt fidelity.

AI USE STATEMENT

In this work, we used a generative AI coding assistant to help write measurement and figure-generation code and to help draft and edit the manuscript text. All experimental designs, measurements, and claims were specified, executed, and verified by the authors; AI-assisted code was reviewed and its outputs cross-checked by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our study uses no human-subject data or personally identifiable information; subject images come from the public DreamBooth dataset, and we follow the licenses of Stable Diffusion v1.5 and that dataset.

REPRODUCIBILITY STATEMENT

All models and data are public (Stable Diffusion v1.5, DreamBooth dataset). Full implementation details, including the fine-tuning recipe, sampler settings, guidance weights, and seeds, are given in Appendix C.

REFERENCES

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard Baraniuk. Self-consuming generative models go MAD. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ShjMHfmPs0.

Sina Alemohammad, Ahmed Imtiaz Humayun, Shruti Agarwal, John Collomosse, and Richard Baraniuk. Self-improving diffusion models with synthetic data. In *Scaling Self-Improving Foundation Models without Human Supervision*, 2025. URL https://openreview.net/forum?id=FHTCV0iE06.

Omri Avrahami, Amir Hertz, Yael Vinker, Moab Arar, Shlomi Fruchter, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. The chosen one: Consistent characters in text-to-image diffusion models. In *ACM SIGGRAPH 2024 conference papers*, pp. 1–12, 2024.

Quentin Bertrand, Joey Bose, Alexandre Duplessis, Marco Jiralerspong, and Gauthier Gidel. On the stability of iterative retraining of generative models on their own data. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=JORAfH2xFd.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Kelvin CK Chan, Yang Zhao, Xuhui Jia, Ming-Hsuan Yang, and Huisheng Wang. Improving subject-driven image synthesis with subject-agnostic guidance. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6733–6742. IEEE, 2024.

Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. CFG++: Manifold-constrained classifier free guidance for diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=E77uvbOTtp.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NAQvF08TcyG.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Tomasz Korbak, Henry Sleight, Rajashree Agrawal, John Hughes, Dhruv Bhandarkar Pai, Andrey Gromov, Dan Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=5B2K4LRgmz.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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. In *International Conference on Learning Representations*, 2022.

Seulgi Jeong and Jaeil Kim. Mindiff: Mask-integrated negative attention for controlling overfitting in text-to-image personalization. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 6981–6990. IEEE, 2025.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. In *Advances in Neural Information Processing Systems*, 2024.

Chaehyun Kim, Heeseong Shin, Eunbeen Hong, Heeji Yoon, Anurag Arnab, Paul Hongsuck Seo, Sunghwan Hong, and Seungryong Kim. Seg4diff: Unveiling open-vocabulary semantic segmentation in text-to-image diffusion transformers. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=ENp2kCdYE8.

Gihoon Kim, Hyungjin Park, and Taesup Kim. Preserve and personalize: Personalized text-to-image diffusion models without distributional drift. In *International Conference on Learning Representations*, volume 2026, pp. 15591–15615, 2026.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023.

Nupur Kumari, Xi Yin, Jun-Yan Zhu, Ishan Misra, and Samaneh Azadi. Generating multi-image synthetic data for text-to-image customization. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16524–16534. IEEE, 2025.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=nAIhvNy15T.

Kyungmin Lee, Sangkyung Kwak, Kihyuk Sohn, and Jinwoo Shin. Direct consistency optimization for robust customization of text-to-image diffusion models. *Advances in neural information processing systems*, 37:103269–103304, 2024.

Yaowei Li, Xiaoyu Li, Zhaoyang Zhang, Yuxuan Bian, Gan Liu, Xinyuan Li, Jiale Xu, Wenbo Hu, yating liu, Lingen Li, Jing Cai, Yuexian Zou, Yancheng He, and Ying Shan. IC-custom: Diverse image customization via in-context learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=gv2cr8kABL.

Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Sunghyun Park, Seokeon Choi, Hyoungwoo Park, and Sungrack Yun. Steering guidance for personalized text-to-image diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15907–15916. IEEE, 2025.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pp. 5301–5310. PMLR, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22500–22510. IEEE, 2023.

Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. CADS: Unleashing the diversity of diffusion models through condition-annealed sampling. In *International Conference on Learning Representations*, 2024.

Seyedmorteza Sadat, Otmar Hilliges, and Romann M. Weber. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=e2ONKX6qzJ.

Seyedmorteza Sadat, Tobias Vontobel, Farnood Salehi, and Romann M Weber. Guidance in the frequency domain enables high-fidelity sampling at low cfg scales. *arXiv preprint arXiv:2506.19713*, 2025b.

Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. ZipLoRA: Any subject in any style by effectively merging LoRAs. In *European Conference on Computer Vision*, 2024.

Dazhong Shen, Guanglu Song, Zeyue Xue, Fu-Yun Wang, and Yu Liu. Rethinking the spatial inconsistency in classifier-free diffusion guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14940–14950. IEEE, 2025.

Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18682–18692. IEEE, 2025.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Youngseok Yoon, Dainong Hu, Iain Weissburg, Yao Qin, and Haewon Jeong. Model collapse in the self-consuming chain of diffusion finetuning: A novel perspective from quantitative trait modeling. In *ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models*, 2025. URL https://openreview.net/forum?id=1MIgdKsvjX.

APPENDIX

A Synthetic training images as the source of the Δ inflation **14**

A.1 Diversity of the training images 14

A.2 Guidance weight of the training images 14

A.3 Prior-preservation loss 15

A.4 The Δ inflation on all subjects 15

B Subject mask **15**

C Implementation details **16**

C.1 Evaluation prompts 18

D Over-guidance artifact metrics **18**

E Additional qualitative results **19**

A SYNTHETIC TRAINING IMAGES AS THE SOURCE OF THE Δ INFLATION

Section 3.2 attributes the Δ inflation of M_{syn} to its training images having been generated by M_{real}. Two other properties of the generated images could account for it: they could be less diverse than the real photographs M_{real} was trained on, or they could inherit the guidance weight at which M_{real} sampled them. Neither accounts for the inflation. Nor is the inflation specific to fine-tuning without the prior-preservation loss of DreamBooth.

A.1 DIVERSITY OF THE TRAINING IMAGES

M_{real} and M_{syn} are fine-tuned on different images, the subject's real photographs and the five images that M_{real} generated. The Δ inflation of Section 3.2 could therefore be attributed to the synthetic set being less diverse than the real one rather than to its origin. For each set we measure diversity as the mean, over all pairs of images in the set, of one minus the cosine similarity between their CLIP ViT-B/32 embeddings (Section 4.1), which does not depend on the number of images, and report the mean and standard deviation over the 30 subjects: 0.142 ± 0.050 for the real photographs and 0.140 ± 0.052 for the five images that M_{real} generated. The two sets are comparably diverse, so what separates M_{real} from M_{syn} is the origin of their training images and not their diversity.

A.2 GUIDANCE WEIGHT OF THE TRAINING IMAGES

Every M_{syn} in the paper is trained on images that M_{real} sampled at the guidance weight of Table 6, $w = 7.5$ on Stable Diffusion v1.5. A higher guidance weight drives each sample further along the guidance direction, so the inflation could be inherited from the weight at which the training images were sampled and would then fall at lower weights. For dog6, the subject of Figure 4, we resample the five training images from the same M_{real} at $w = 5$ and $w = 3$ with the same prompts and seeds, train M_{syn} on each set with the Stable Diffusion v1.5 DreamBooth recipe of Table 6, unchanged, and repeat the measurement of Table 4. Table 5 reports the inflation at the calibration states, before correction, by band group; the value at 7.5, 5.57, sits within the 30-subject mean of Table 4, 5.31 ± 1.26. The inflation stays large at all three weights and keeps its ordering from the low-k to the high-k group. Lowering the weight from 7.5 to 3 reduces it by about 12%, and most of that drop is already reached at 5.

Table 5: Δ inflation of dog6 against the base model, by band group, for M_{syn} trained on images sampled from M_{real} at three guidance weights (root mean square over the group's bands and all steps at the calibration states, as the Before column of Table 4). The 7.5 row is the M_{syn} used throughout the paper.

Guidance weight w	Low k	Mid k	High k	All
7.5	4.16	4.71	6.55	5.57
5.0	3.86	4.28	5.80	4.98
3.0	3.50	4.09	5.79	4.88

A.3 Prior-preservation loss

DreamBooth (Ruiz et al., 2023) can add a prior-preservation loss, which also trains the model on images of the subject's class that the base model generates under the class prompt, so that the class noun does not drift toward the subject. The models in the paper are fine-tuned without it (Table 6). To test whether it would prevent the inflation, for dog6 we fine-tune M_{real} and M_{syn} as in the main experiments with this loss added, training for 800 steps following the Diffusers recommendation, with M_{syn} again trained on five images generated by M_{real}, and repeat the measurement of Table 4. At the calibration states, before correction and computed as in Table 5, the inflation against the base model is 7.38, 7.35 and 9.47 in the low-, mid- and high-k groups and 8.41 over all bands. The prior-preservation loss therefore does not remove the inflation, which remains largest in the high-k group. Since it leaves the inflation in place, we keep the simpler setup without it.

A.4 The Δ inflation on all subjects

We repeat the measurement of finding 1 on all 30 DreamBooth subjects, each with its own M_{real} and M_{syn}, under the subject prompt with the subject's class noun and 5 seeds. Figure 8 shows the per-step ratios of θ, δ and r for every subject. Averaged over the 50 steps, $\|\Delta\|$ of M_{syn} exceeds that of M_{real} in 29 of the 30 subjects, by $48 \pm 22\%$ (mean and standard deviation over subjects), and θ by $48 \pm 22\%$, while δ and r change by $-0.1 \pm 0.1\%$ and $0.6 \pm 2.2\%$. The exception, rc_car, shows no excess in either.

B Subject mask

The subject mask m of equation 5 is built from existing segmentation methods, used unchanged.

Saliency maps. On Stable Diffusion v1.5 and SDXL we use the segmentation stage of S-CFG (Shen et al., 2024) unchanged: the cross-attention maps of the conditional pass at the two coarsest U-Net resolutions (16×16 and 8×8 on Stable Diffusion v1.5) are refined by propagation over the self-attention affinity graph (SSGC, four hops), normalized per token to unit spatial mean, averaged over layers, upsampled to the coarsest resolution and smoothed with a 3×3 Gaussian ($\sigma = 0.5$). This gives one saliency map s_j per prompt token. Stable Diffusion 3.5 has no cross-attention. There we follow Seg4Diff (Kim et al., 2025) and read the maps from the joint attention of transformer block 9, taking the image queries against the 77 CLIP text keys, with the T5 encoder omitted from this one evaluation, on the 64×64 token grid and without smoothing, at the cost of one additional conditional evaluation per sampling step.

Competition rule. The mask is a per-pixel competition between two token sets with no free parameters:

$$m(p) \;=\; \mathbb{1}\Big[\, \max_{j \in \mathcal{S}} s_j(p) \;>\; \max_{j \in \mathcal{B}} s_j(p) \,\Big], \tag{6}$$

where $\mathcal{S}$ holds the sub-word tokens of the identifier and the class noun and $\mathcal{B}$ the prompt's scene content words. Function words and generic framing words such as *photo* are excluded through a fixed stopword list: after per-token normalization their maps are nearly flat at unit height and would outcompete the subject wherever its saliency is not sharply peaked, collapsing the mask to the subject's most discriminative parts. The maximum within each set keeps the rule invariant to the number of words in it. Pixels claimed by neither set resolve to the subject, so the mask errs toward over-coverage of texture-free background. Such pixels receive the rescaled guidance term

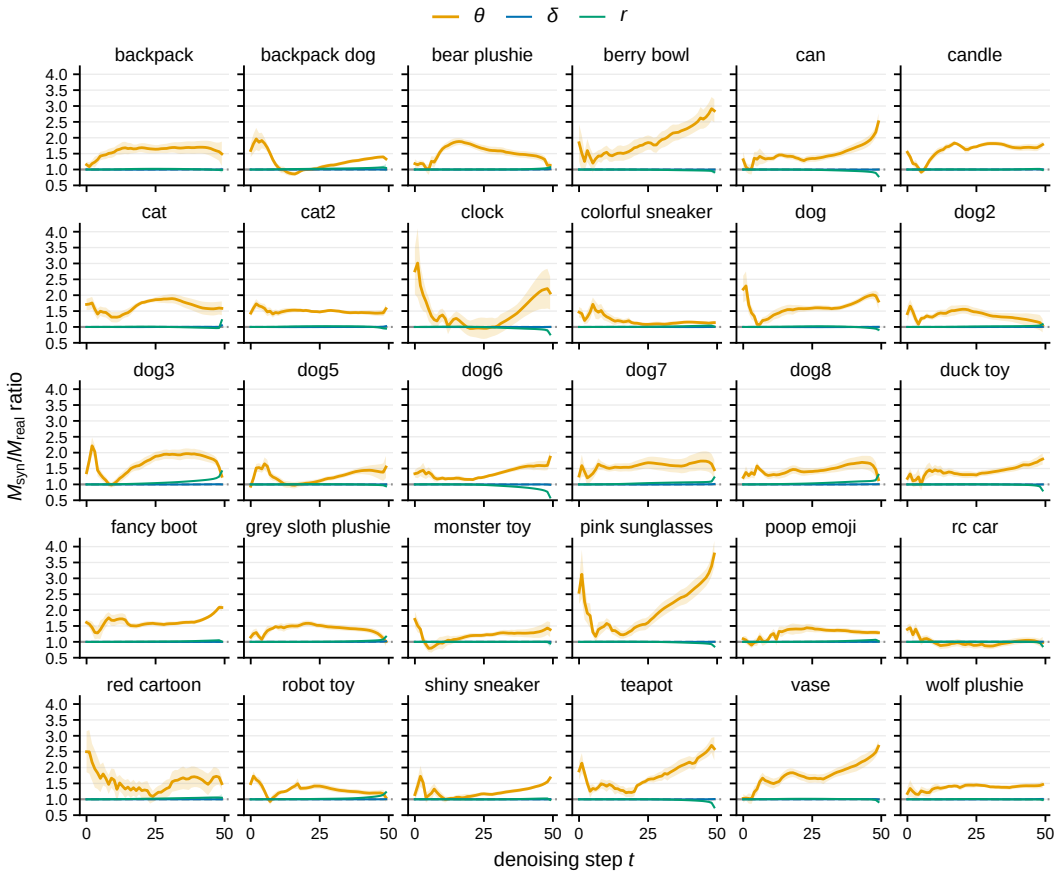

Figure 8: Per-step $M_{\text{syn}}/M_{\text{real}}$ ratio of θ, δ and r for the 30 DreamBooth subjects, as in Figure 2a. Mean over 5 seeds, shading is the standard error.

of equation 5. The same construction is used at every sampling step, across seeds, prompts and subjects (Figure 9).

Mask at calibration. The calibration prompts of Section 3.3 name no scene content, so $\mathcal{B}$ would be empty. The mask is therefore read from one auxiliary evaluation of M_{syn} at the same forward-noised state under the subject prompt extended with the fixed phrase "in a scene", whose single content word fills $\mathcal{B}$. The rule equation 6 is otherwise unchanged and no caption of the training images is used. Where the captions are known they permit a check: masks built with the generic phrase agree with masks built from each training image's own caption at IoU 0.86, averaged over the five training images and all 50 steps. At sampling, the same auxiliary evaluation supplies the mask when the sampled prompt names no scene content.

C Implementation details

Table 6 lists the fine-tuning and sampling parameters of the four settings of Table 1. Within a setting, M_{real} and M_{syn} are trained with the same procedure and hyperparameters and differ only in their training images: M_{real} is trained on the subject's real photographs and M_{syn} on the five images that M_{real} generated. The identifier [V] is the token *monadikos* and the fine-tuning prompt is *"a photo of monadikos ⟨class⟩"* throughout.

Synthetic training set. M_{real} generates one image under each of five prompt templates shared by all subjects and disjoint from the evaluation prompts: a close-up photo of the subject, and *"a photo of monadikos ⟨class⟩"* completed by one of *outdoors in a backyard on a sunny day*, *on a couch*, *on a staircase* and *in front of a brick wall*.



Figure 9: **Subject mask m during sampling.** The mask of equation 6 (red) at steps 0, 10, 20 and 40 of a 50-step plain-CFG trajectory of M_{syn}, drawn on the clean image predicted at that step, and the mask of the last step on the final sample. Stable Diffusion v1.5, DreamBooth, seed 100. Each row gives the subject and the scene phrase of its evaluation prompt *"a [V] ⟨class⟩ . . . "* (Appendix C.1).

Table 6: Fine-tuning and sampling settings. One column per setting of Table 1. LoRA adapters use a scaling factor α equal to the rank and are merged into the base model weights before sampling. All text encoders are frozen and no prior-preservation loss is used. The synthetic training set is sampled with each base model's default scheduler and evaluation uses DDIM, except on Stable Diffusion 3.5, which uses its flow-matching Euler scheduler (shift 3.0) in both cases.

	SD v1.5 DreamBooth	SD v1.5 DreamBooth-LoRA	SDXL 1.0 DreamBooth-LoRA	SD 3.5 Medium DreamBooth-LoRA
Fine-tuning				
Trained weights	full U-Net	U-Net attention	U-Net attention	MMDiT
LoRA rank	–	16	4	4
Learning rate (constant)	5×10^{-6}	10^{-4}	5×10^{-4}	4×10^{-4}
Steps	500	500	500	600
Batch size	1	1	1	1
Gradient accumulation	1	1	1	4
Resolution	512	512	1024	512
Precision	fp32	fp32	fp16	bf16
Subjects	30	30	30	30
Synthetic training set of M_{syn} (generated by M_{real})				
Images	5	5	5	5
Sampler, steps	PNDM, 50	PNDM, 50	Euler, 50	flow Euler, 40
Guidance weight	7.5	7.5	7.5	7.0
Resolution	512	512	1024	1024
Evaluation sampling (all models)				
Sampler, steps	DDIM, 50	DDIM, 50	DDIM, 50	flow Euler, 40
Guidance weight w	7.5	7.5	7.5	7.0
Resolution	512	512	1024	1024
Seeds	100–102	100–102	100–102	100–102

Guidance baselines. The three baselines of Table 1 are sampled from the same M_{syn} on both Stable Diffusion v1.5 settings, with DDIM for 50 steps at the seeds and prompts of ReGain and with the settings of their official implementations. S-CFG (Shen et al., 2024) uses $w = 7.5$ and rescales the guidance of each attention region at every step, with the rate clipped to $[0.8, 3.0]$. CFG++ (Chung et al., 2025) uses $\lambda = 0.6$, the value its authors match to $w = 7.5$ at 50 steps. FDG (Sadat et al., 2025b) splits the guidance with a one-level Laplacian pyramid and uses $w = 7.5$ on the high band and $w = 3$ on the low band, the setting its authors report for Stable Diffusion 2.1.

Trajectory measurements. The measurements of Section 3.2 are made on one dog subject (dog6), with M_{real} and M_{syn} obtained by DreamBooth fine-tuning of Stable Diffusion v1.5. Both models are sampled with DDIM for 50 steps at $w = 7.5$, with 10 seeds shared by both models and one image per prompt and seed. For the band-resolved comparison, M_{real} is evaluated at the states that M_{syn} visits. The frequency bands are computed on the 64×64 latent, which gives $K = 46$ bands, and the low, mid and high band groups are $k = 1$ to 8, 9 to 24 and 25 to 45.

Schedule estimation. The estimator equation 4 is evaluated at the five training images, each forward-noised with 10 independent noise draws at every step, the same draws for both models. The estimated gain $\hat{g}(k, t)$ has one band per ring of the model's latent grid, 46 on the 64×64 latent of Stable Diffusion v1.5 and 91 on the 128×128 latent of SDXL and Stable Diffusion 3.5, and one column per sampling step (50, or 40 on Stable Diffusion 3.5). The compression uses the root-mean-square tolerance $\tau = 0.05$ at each of its two stages. The number of cells is determined by the compression and varies by subject: dog6 has 3 band groups and 15 cells (Figure 4), and the mean over the 30 subjects of Table 4 is 18 cells. The DC band is kept as measured. At the calibration states, the residual inflation after the compressed schedule (Table 4) sits in the last five steps, where the schedule's time cells are coarsest. Figure 10 shows the schedules of three further subjects. Across all 30 subjects, the highest band group is attenuated at every step, with gains between 0.11 and 0.79 (median 0.21).

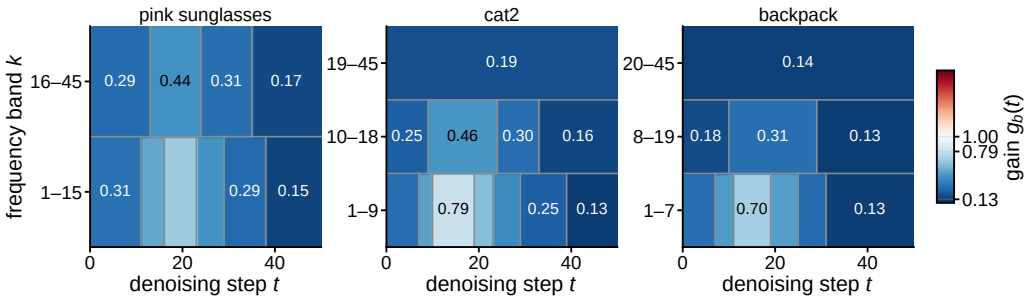

Figure 10: **Estimated gain schedules of three further subjects.** The compressed schedule $g_b(t)$ at $\tau = 0.05$ of the pink sunglasses, cat2 and backpack subjects of Figure 5, Stable Diffusion v1.5 DreamBooth, in the layout of Figure 4: one row per band group, one column per sampling step, and color and label giving the gain.

C.1 Evaluation prompts

Evaluation uses the official DreamBooth prompt lists (Ruiz et al., 2023) verbatim, instantiated as *"a monadikos ⟨class⟩ ..."* with each subject's official class noun. Object subjects use 25 prompts: *in the jungle, in the snow, on the beach, on a cobblestone street, on top of pink fabric, on top of a wooden floor, with a city in the background, with a mountain in the background, with a blue house in the background, on top of a purple rug in a forest, with a wheat field in the background, with a tree and autumn leaves in the background, with the Eiffel Tower in the background, floating on top of water, floating in an ocean of milk, on top of green grass with sunflowers around it, on top of a mirror, on top of the sidewalk in a crowded street, on top of a dirt road, on top of a white rug*, and the property modifications *red*, *purple*, *shiny*, *wet* and *cube shaped*. Live subjects use 25 prompts: the first ten recontextualization prompts above, the accessorization prompts *wearing a red hat, wearing a santa hat, wearing a rainbow scarf, wearing a black top hat and a monocle, in a chef outfit, in a firefighter outfit, in a police outfit, wearing pink glasses, wearing a yellow shirt, in a purple wizard outfit*, and the same five property modifications.

D Over-guidance artifact metrics

Definitions. Saturation is the mean of the HSV S channel, and root-mean-square contrast is the standard deviation of ITU-R 601 grayscale intensity, both on $[0, 1]$. The high-band fraction is the share of Fourier power, with the zero-frequency term removed, beyond a radial cutoff at a quarter of the Nyquist frequency, the pixel-space image of the low-band boundary of Section 3.2. Removing

the mean matters: with the DC term left in the denominator it dominates the total power and scales with brightness and variance, so the statistic would then track contrast as well as spectral shape.

Measurement region. The region of Table 2 is the mask that steered M_{syn}, taken once per subject, prompt and seed and applied to all three methods so that they are compared over identical pixels. Over the evaluation set it covers 43% of the frame. The high-band fraction needs a rectangular grid and is measured on the mask's bounding box.

E ADDITIONAL QUALITATIVE RESULTS

Figures 11 and 12 repeat the comparison of Figure 5 in the DreamBooth-LoRA setting on Stable Diffusion v1.5 and SDXL, Figure 13 adds further subjects and prompts, and Figure 14 adds further rows of Figure 7.



Figure 11: **DreamBooth-LoRA, Stable Diffusion v1.5.** The block of Figure 5 in the DreamBooth-LoRA setting: the subject's reference photos, the prompt, and then M_{real}, M_{syn} under plain CFG, and the same M_{syn} under its estimated gain schedule (ReGain, ours), at one seed and w=7.5.

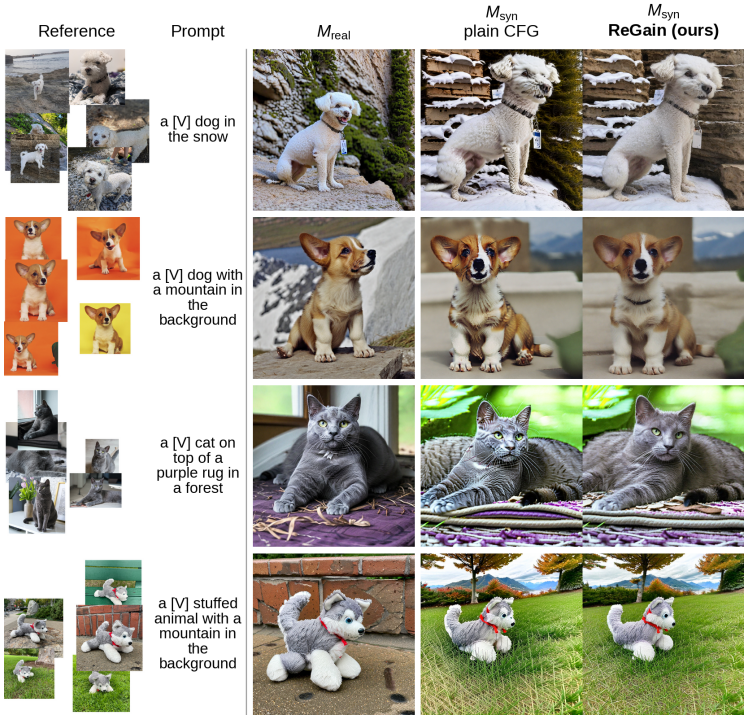

Figure 12: **DreamBooth-LoRA, SDXL base 1.0.** The same block on SDXL at 1024 $\times$ 1024. Columns and protocol follow Figure 11.

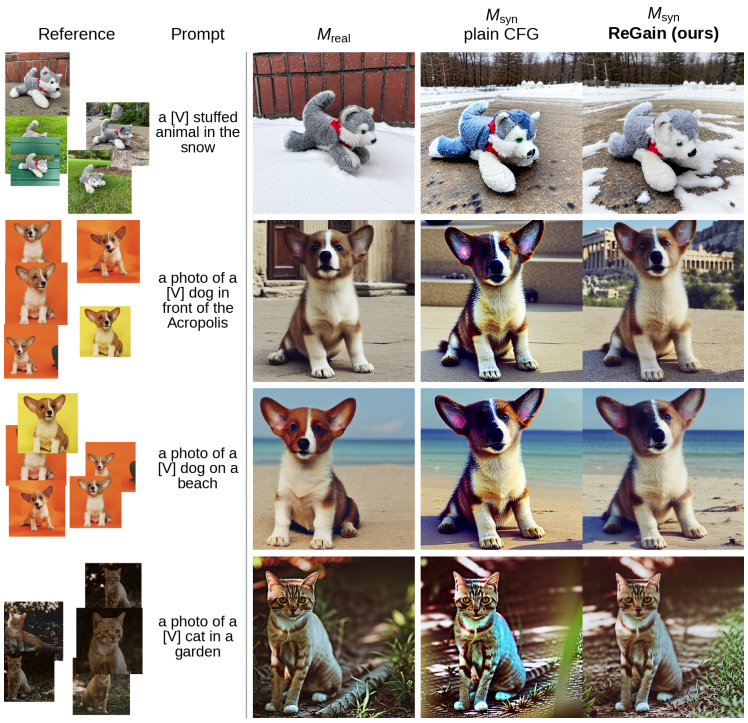

Figure 13: **Additional subjects and prompts.** Four further subject and prompt combinations, shown in the same block at the same seed and guidance weight (w=7.5, 50 steps): the subject's reference photos, the prompt, then M_{real}, M_{syn} under plain CFG, and the same M_{syn} under its estimated gain schedule (ReGain, ours).

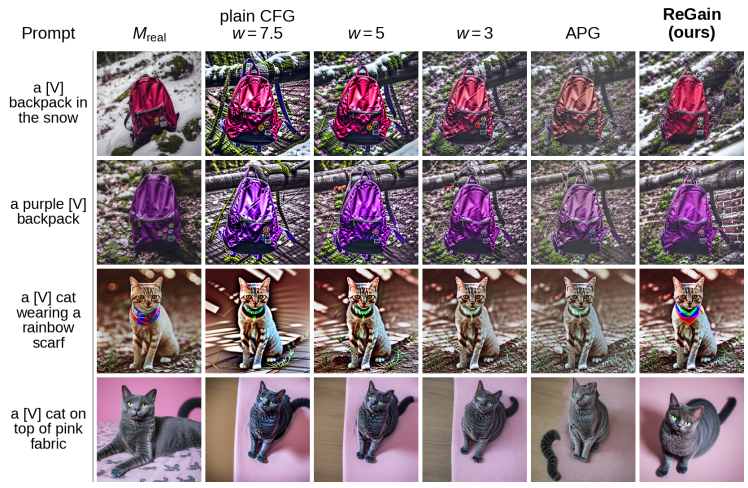

Figure 14: **Subject-agnostic alternatives, further rows.** Four more rows in the layout of Figure 7, showing M_{real}, M_{syn} under plain CFG at w=7.5, 5.0 and 3.0, under APG at w=7.5, and under ReGain, at the same seed across columns.