Follow the Mean: Reference-Guided Flow Matching

Pedro M. P. Curvo* **Maksim Zhdanov†** **Floor Eijkelboom†** **Jan-Willem van de Meent‡**
University of Amsterdam

Abstract

Existing approaches to controllable generation typically rely on fine-tuning, auxiliary networks, or test-time search. We show that flow matching admits a different control interface: adaptation through examples. For deterministic interpolants, the velocity field is solely governed by a conditional endpoint mean; shifting this mean shifts the flow itself. This yields a simple principle for controllable generation: steer a pretrained model by changing the reference set it follows. We instantiate this idea in two forms. Reference-Mean Guidance is training-free: it computes a closed-form endpoint-mean correction from a reference bank and applies it to a frozen FLUX.2-klein (4B) model, enabling control of color, identity, style, and structure while keeping the prompt, seed, and weights fixed. Semi-Parametric Guidance amortizes the same idea through an explicit mean anchor and learned residual refiner, matching unconditional DiT-B/4 quality on AFHQv2 while allowing the reference set to be swapped at inference time. These results point to a broader direction: generative models that adapt through data, not parameter updates.

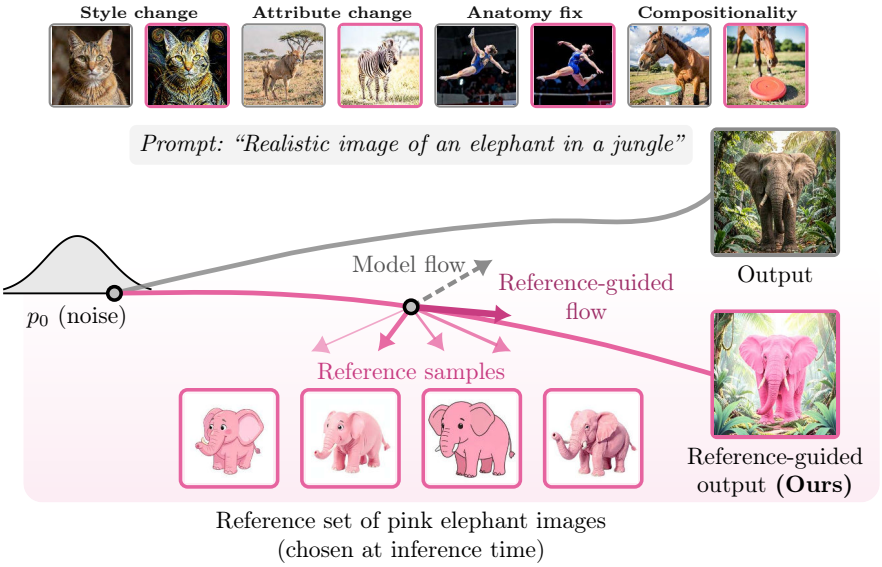


Figure 1: Overview of reference-guided flow matching. A noisy state is matched against a reference set $\mathcal{R}$ to shift the prediction endpoint mean relative to the prediction of the pre-trained model. This results in a flow that incorporates characteristics of the reference set in an implicit manner, without requiring explicit access to a classifier or reward.

*pedro.pombeiro.curvo@student.uva.nl
†Co-supervisor.
‡Primary supervisor.

Preprint.

1 Introduction

Flow matching [1] has emerged as a dominant paradigm for training generative models, with recent approaches producing high-quality samples across image, video, and scientific domains [1–5]. Many downstream applications, however, require control over the outputs of a pretrained model, such as enforcing a specific attribute, concept, style, or target distribution at generation time. Achieving such control without retraining the base model remains a challenging problem.

Existing approaches to controlled generation can be categorized into three groups. *Fine-tuning and adapter methods* modify model parameters for each new target [6–8]. *Guidance methods* leave the generator unchanged but rely on auxiliary classifiers or reward signals [9–11]. *Search-based methods* avoid additional training but incur repeated sampling, filtering, or per-prompt optimization at inference time [12–15]. None of these approaches simultaneously avoids additional training, auxiliary networks, or test-time search.

In this paper, we present an alternative formulation of controlled generation, which we refer to as *reference-guided flows*. Our control object is the endpoint mean – the mean of the posterior distribution over data points given a noisy interpolant. Because the velocity field in flow matching points toward the endpoint mean [1, 3, 16], shifting this mean also shifts the induced distribution over generated samples. The key insight is that this shift is comparatively straightforward to compute when we have access to reference samples. These need not be perfect representatives of the target distribution, as long as they shift the mean in the desired direction. Conditioning on a reference set thus provides a mechanism for implicit guidance in the absence of an explicitly defined reward or classifier. In short:

"Guide with examples, not rewards."

Fig. 1 illustrates this approach on a frozen text-to-image model: when prompted with "an elephant in a jungle" the model produces a photorealistic elephant, while conditioning on a small set of images of pink elephants changes the color of the elephant to pink.

Contributions.

1. **Mean shift as a mechanism for guidance.** We define the guidance term in terms of a geometric interpolation between the probability path induced by the training set and that induced by a reference set. We show that this guidance term can be expressed in terms of the difference between the endpoint mean of the reference set and that of the training distribution.
2. **Reference-mean guidance.** We show that we can efficiently compute the shift in the velocity field in terms of the closed form endpoint mean of the reference set. This defines a training-free method for test-time control of pre-trained models, without additional training, auxiliary networks, gradient computation, or extra model evaluations.
3. **Semi-parametric guidance.** We introduce a controllable generative architecture that amortizes reference-mean guidance through an explicit posterior-mean anchor and a learned residual refiner. By combining a reference-set anchor with a learned residual correction, SPG can preserve useful reference information while suppressing nuisance correlations, such as shared backgrounds, while still preserving unconditional generation quality and enabling inference-time reference-set control.

2 Background

2.1 Flow Matching

Flow matching (FM) learns a continuous-time transport model that maps a source distribution p_0 to a target distribution p_1 [1–3]. To do so, it defines a time-dependent distribution p_t, known as the probability path, in terms of an affine interpolant

$$x_t = \alpha_t x_0 + \beta_t x_1, \qquad t \in [0, 1]. \tag{1}$$

The ordinary differential equation $\dot{x} = u_t(x)$ transports samples from p_0 to p_t when the velocity field $u_t(x)$ satisfies the continuity equation $\partial_t p_t(x) + \nabla \cdot (p_t(x)u_t(x)) = 0$. This condition holds when

$$u_t(x) := \mathbb{E}\big[\dot{\alpha}_t x_0 + \dot{\beta}_t x_1 \,\big|\, x_t = x\big]. \tag{2}$$

For the commonly used linear interpolant with $\alpha_t = (1-t)$ and $\beta_t = t$, the velocity field can be expressed in terms of the endpoint mean $\mu_t(x)$ as (see Appendix A.1 for a derivation)

$$u_t(x) = \frac{\mu_t(x) - x}{1-t}, \qquad \mu_t(x) := \mathbb{E}[x_1 \mid x_t = x]. \tag{3}$$

Intuitively, this identity states that, for a linear interpolant, the velocity field moves samples from x towards $\mu_t(x)$ at any point in time, at a rate inversely proportional to the remaining time $1-t$.

The identity in (3) is invertible; the endpoint mean $\mu_t(x) = x + (1-t)u_t(x)$ can also be expressed in terms of the velocity field. Operationally, this implies that we can parameterize a flow matching problem either in terms of $u_t^\theta(x)$ or $\mu_t^\theta(x)$. Similarly we can define an objective in terms of the predicted velocity, by minimizing the standard flow matching loss [1–3], or in terms of the mean $\mu_t^\theta(x)$ of a variational distribution $q_t^\theta(x_1 \mid x_t)$, by minimizing the variational flow matching loss [16]

$$\mathcal{L}^{\text{FM}}(\theta) = \mathbb{E}\Big[\big\|(x_1 - x_0) - u_t^\theta(x_t)\big\|^2\Big], \qquad \mathcal{L}^{\text{VFM}}(\theta) = -\mathbb{E}\Big[\log q_t^\theta(x_1 \mid x_t)\Big], \tag{4}$$

where $x_0 \sim p_0$, $x_1 \sim p_1$ and $t \sim \text{Uniform}([0,1])$. Either parameterization can be employed with either loss, so any pre-trained model equivalently specifies a velocity field and an endpoint mean. More broadly, an analogous observation holds for diffusion models [17].

2.2 Closed Form of the Endpoint Mean

In practice, when training a flow matching model, we approximate the target distribution with an empirical distribution $\hat{p}_1$ over a finite training set $\mathcal{D} = \{x^{(1)}, \ldots, x^{(N)}\}$.

Given an empirical distribution $\hat{p}_1(x_1)$ there is a closed-form expression for the endpoint mean,

$$\hat{\mu}_t(x) = \sum_{n=1}^{N} w_t^{(n)}(x)\, x^{(n)}, \qquad w_t^{(n)}(x) = \frac{p_t(x_t = x \mid x_1 = x^{(n)})}{\sum_{m=1}^{N} p_t(x_t = x \mid x_1 = x^{(m)})}. \tag{5}$$

For a standard normal source $p_0 = \mathcal{N}(0, I)$, the weights are (see Appendix A.2 for a derivation)

$$w_t^{(n)}(x) = \text{Softmax}_n\left(-\frac{1}{2}\frac{\|x - tx^{(n)}\|^2}{(1-t)^2}\right) = \text{Softmax}_n\left(\frac{tx^\top x^{(n)} - \frac{1}{2}t^2\|x^{(n)}\|^2}{(1-t)^2}\right). \tag{6}$$

This means that the learned endpoint mean $\mu_t^\theta(x)$ approximates the empirical endpoint mean $\hat{\mu}_t(x)$, which is simply a weighted sum over the training set. This observation is in itself not new; it has been made in the context of both flow matching [18, 19] and score matching [20, 21] (see Section 5 for a more detailed discussion). However, to our knowledge this observation has not previously been leveraged in the design of guidance methods. In the next Section, we will show how we can use the closed-form mean to compute a guidance term for our training-free variant of reference-mean guidance, and will use the structure of (6) to inform design of the amortized semi-parametric variant.

3 Reference-Guided Flows

3.1 Steering a Flow by Shifting the Endpoint Mean

This work starts from a simple observation. Suppose we have a pretrained flow model that approximates the velocity $u_t(x)$ and endpoint mean $\mu_t(x)$ associated with a distribution over training data p_1. At test time, we would like to generate from a different target distribution π_1. Let π_t be the path under the same bridge, and let $\mu_t^\pi(x)$ denote its endpoint mean. Because both flows share the same

source and bridge structure, their velocity fields differ only through their endpoint means:

$$u_t^\pi(x) - u_t(x) \;=\; \frac{\mu_t^\pi(x) - \mu_t(x)}{1-t}. \tag{7}$$

Any target distribution π_1 is therefore reachable by approximating the shift in the endpoint mean $\mu_t^\pi(x) - \mu_t(x)$ during generation (derivations for general affine interpolants $x_t = \alpha_t x_0 + \beta_t x_1$ are given in Appendix A). We can recover the mean $\mu_t^\theta(x)$ from any pretrained flow either because the network outputs it directly, or by inverting Eq. (3) to define $\mu_t^\theta(x) = x + (1-t)\, u_t^\theta(x)$.

3.2 Reference-Mean Guidance (RMG)

The idea that we will now develop is to use a set of reference samples to implicitly specify $\mu_t^\pi(x)$. Suppose that we define a reference set $\mathcal{R} = \{x^{(1)}, \ldots, x^{(M)}\}$ sampled from a distribution ρ_1. Our goal is to shift the target distribution toward the endpoint mean $\mu_t^\rho(x)$ induced by the reference set, while preserving the diversity and quality of the pretrained model.

Geometric mixture at the endpoint level. Define the geometric mixture of training and reference endpoint distributions,

$$\pi(x_1) \propto p_1(x_1)^{1-\beta_t}\, \rho_1(x_1)^{\beta_t},$$

and let $\pi_t(x) = \int p_t(x \mid x_1)\, \pi(x_1)\, dx_1$ be its noisy marginal under the same affine bridge. This is a valid bridge marginal by construction. Applying the score-to-mean identity and a Gaussian posterior approximation — exact when p_1 and ρ_1 are Gaussian, as is approximately the case in VAE latent spaces — gives the guided endpoint mean and velocity:

Proposition 3.3 (Reference-mean guided dynamics). Let $\beta_t \in [0,1]$ be a scalar guidance schedule. Under the geometric endpoint mixture, the score-implied guided velocity is

$$\mu_t^\pi(x) \;=\; (1-\beta_t)\, \mu_t(x) + \beta_t\, \mu_t^\rho(x), \tag{8}$$

$$u_t^\pi(x) \;=\; u_t(x) + \beta_t\, \frac{\mu_t^\rho(x) - \mu_t(x)}{1-t}. \tag{9}$$

See Appendix A.3 for the full derivation.

Remark. An alternative exact construction uses the arithmetic mixture $\hat{p}_\lambda = (1-\lambda)\, p_1 + \lambda\, \rho_1$, whose noisy marginal $\pi_t(x) = (1-\lambda)\, p_t(x) + \lambda\, \rho_t(x)$ is also a valid bridge marginal. Bayes' rule gives its exact posterior mean

$$\mu_t^\lambda(x) = \big(1 - \omega_t^*(x)\big)\, \mu_t(x) + \omega_t^*(x)\, \mu_t^\rho(x), \qquad \omega_t^*(x) = \frac{\lambda\, \rho_t(x)}{(1-\lambda)\, p_t(x) + \lambda\, \rho_t(x)}. \tag{10}$$

Replacing the intractable $\omega_t^*(x)$ with a scalar β_t recovers the same guided velocity as Proposition 3.3, confirming that both constructions support the same guidance rule (Appendix A.4).

This result instantiates the mean-shift mechanism in Eq. (7). The shift depends entirely on data, with no auxiliary models or gradient computations. In practice, two approximations are involved: (i) μ_t is replaced by the pretrained model's estimate μ_t^θ; and (ii) μ_t^ρ is replaced by the empirical mean $\hat{\mu}_t^\rho$ over a finite reference bank $\mathcal{R}$. Changing the composition of $\mathcal{R}$ directly controls the guided velocity field.

We refer to the resulting method as *reference-mean guidance* (RMG), with the empirical reference mean $\hat{\mu}_t^\rho$ computed as the closed-form weighted average in Eq. (5):

$$u_t^\pi(x) \simeq u_t^\theta(x) + \beta_t \frac{\hat{\mu}_t^\rho(x) - \mu_t^\theta(x)}{1-t}. \tag{11}$$

3.3 Semi-Parametric Guidance (SPG)

As a complement to the training-free guidance based on the empirical mean, we consider a semi-parametric variant in which the model $\mu_t^\theta(x_t, \mathcal{R})$ has access to a reference set at training time. We first use a cross-attention pass to compute an anchor $\bar{x}$ analogous to the closed form in Eq. (5), where

learned attention replaces the closed-form weights. The final endpoint prediction combines the noisy state, the anchor, and a learned residual correction via time-dependent gates (details in Appendix C.1),

$$\mu_t^\theta(x_t, \mathcal{R}) = (1 - g_t) \cdot x_t + g_t \cdot \bar{x} + \alpha_t \cdot f^\theta\big(\bar{x},\, x_t,\, t\big), \tag{12}$$

where $g_t, \alpha_t \in [0, 1]$ are scalar time-dependent gates, f^θ predicts a residual correction to the anchor, and $\bar{x}$ is computed from a cross-attention step with identity value projection,

$$\bar{x} = \sum_{m=1}^{M} \alpha_m x^{(m)}, \qquad \alpha = \mathrm{Softmax}_m\left(\langle q^\theta(x_t), k^\theta(x^{(m)})\rangle\right). \tag{13}$$

During training, the reference set $\mathcal{R}$ is sampled from the training set. For each sample, we generate an interpolation $x_t^{(m)}$ and condition on $\mathcal{R}^{\setminus\{m\}} := \mathcal{R} \setminus \{x^{(m)}\}$, giving a batch-level endpoint prediction objective

$$\mathcal{L}_\mu(\theta) = \mathbb{E}\left[\sum_{m=1}^{M} \frac{1}{(1-t)^2} \left\| x^{(m)} - \mu_t^\theta\big(x_t^{(m)}, \mathcal{R}^{\setminus\{m\}}\big)\right\|^2\right]. \tag{14}$$

The leave-one-out structure prevents $x_t^{(m)}$ from attending to its own endpoint. Because the anchor is already a strong predictor, the refiner receives little gradient signal from $\mathcal{L}_\mu$ alone; we therefore train it on the positive residual between ground truth and anchor, with gradients stopped through the anchor:

$$\mathcal{L}_{\mathrm{ref}}(\theta) = \mathbb{E}\left[\sum_{m=1}^{M} \left\| \mathrm{sg}\left[x^{(m)} - \bar{x}^{(m)}\right] - f^\theta\left(\mathrm{sg}\left[\bar{x}^{(m)}\right], x_t^{(m)}, t\right)\right\|^2\right], \tag{15}$$

where $\bar{x}^{(m)}$ is the cross-attention anchor computed from $x_t^{(m)}$ and $\mathcal{R}^{\setminus\{m\}}$, and $\mathrm{sg}[\cdot]$ denotes stop-gradient. Since references are uncorrelated across the batch, a sufficiently high-capacity refiner could in principle ignore $\bar{x}$ entirely and predict μ_t^θ directly from x_t. In practice this does not happen: the reference set measurably controls generation at test time (Section 4.2), suggesting the training scheme induces an implicit exchangeability structure in which samples are treated as conditionally i.i.d. given an unobserved latent reference measure.

4 Results

4.1 Reference-Mean Guidance

We validate the central claim of Section 3: that the posterior mean controls the flow, and that modifying the reference set provides a direct mechanism for steering generation. Section 4.1.1 verifies this in controlled settings where the posterior mean can be computed exactly; Section 4.1.2 applies the same mechanism to a frozen FLUX.2-klein (4B) model.

4.1.1 Mechanistic Validation

We use $N = 500$ samples from the two-moons distribution; labels exist but are withheld from the model, and a small labeled reference set is used only to compute soft posterior weights at inference time. Varying only the composition of this reference set, Fig. 2 shows that the flow field and final attractor shift accordingly, isolating the causal role of the posterior mean. Additional results in Appendix D show how the posterior concentrates around the class structure as $t \to 1$, that as few as $M = 5$ references approach the hard-filter upper bound, and that the mechanism transfers to pixel space on MNIST without modification.

4.1.2 Training-Free Control in FLUX.2-klein (4B)

Setup. We apply RMG (Section 3.2) to a frozen FLUX.2-klein (4B) model [5]. FLUX.2 is a latent rectified-flow model, so the linear bridge identity holds natively and endpoint recovery reduces to $\mu_t^\theta(x) = x + (1-t)u_t^\theta(x)$. Reference images are encoded with the same frozen VAE, so all corrections operate in the same latent coordinate system as the pretrained model. Throughout all experiments, the prompt, noise seed, and model weights are fixed; only the reference set changes. Each reference set consists of 20 images encoding a target attribute (e.g., color, object identity, or style), with no modification to model parameters. Hyperparameters, prompts, metrics, and reference

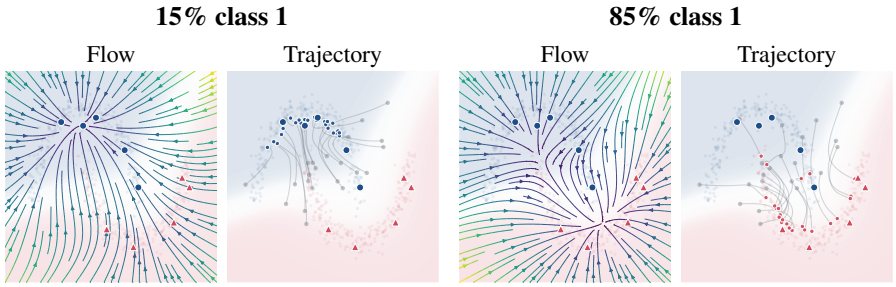


Figure 2: Reference-mean guidance on the two-moons distribution. The model and all other settings are fixed; only the reference-set composition changes. With 15% class-1 references (left), the flow field and sample trajectories concentrate toward the minority moon. With 85% class-1 references (right), they shift toward the majority moon. The change in attractor isolates the causal role of the posterior mean: modifying the reference set directly redirects the generative dynamics.

sets are provided in Appendices C, C.3 and G, along with ablations on guidance schedule, strength, reference-set size, and NFE (Appendices E.1 to E.5) and additional experiments on prompt–reference interaction, reference composition, SPG diversity, and nuisance-artifact suppression (Appendices F.1 to F.4).

Reference-controlled generation. Fig. 3 shows results across four prompts, with two reference sets per prompt encoding distinct attributes — color, object identity, or style. In each case the generated output shifts systematically with the reference set, confirming that the posterior mean induced by the reference set acts as a control signal for a frozen pretrained model.

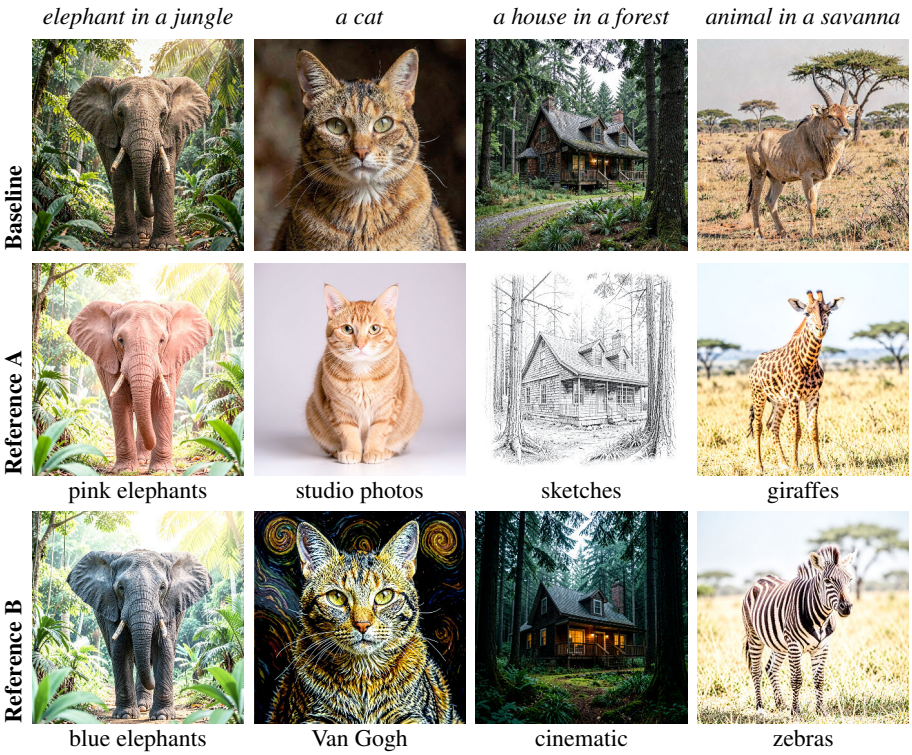


Figure 3: Reference-set swaps on frozen FLUX.2-klein. Prompt and noise seed are fixed within each column. The generated output shifts systematically in color, object identity, and style as the reference set changes.

Geometric control via structural references. Geometric and anatomical control remains challenging for reward- and gradient-based approaches, as structural correctness — unlike color or style, which admit straightforward perceptual metrics — lacks a simple scalar proxy: even powerful VLMs struggle to reliably judge whether a silhouette matches a target shape, a hand is correctly oriented, or limbs are properly ordered in depth [22].

We provide qualitative evidence that RMG can transfer coarse structural priors in selected challenging cases. We consider three settings: a keyhole-shaped composition, a hand making the sign-of-the-horns gesture, and a gymnast performing a ring leap. Fig. 4 shows that RMG improves adherence to the target structure in all three cases. In the keyhole example, the correction acts on the global silhouette while preserving the interior scene. The hand and gymnastics examples suggest that small pose-specific reference sets can inject structural priors without gradients, retraining, or additional model evaluations, though broader quantitative evaluation remains an open direction.

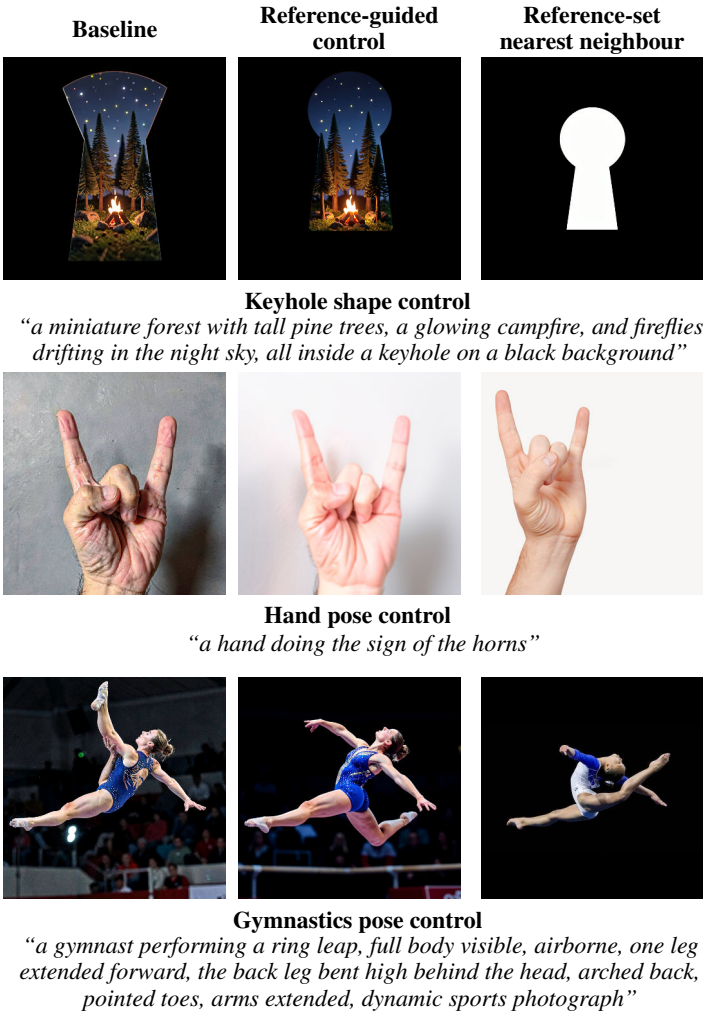


Figure 4: Qualitative evidence of structural control on frozen FLUX.2-klein. The nearest-neighbour column shows a representative reference-set image, confirming that RMG transfers structural priors rather than copying reference content. In the keyhole example, the correction reshapes the global silhouette while preserving the interior scene. In the hand and gymnastics examples, small pose-specific reference sets shift the output toward the target structure.

Comparison of control interfaces. We evaluate on GenEval [23], a compositional text-to-image benchmark spanning single objects, two objects, counting, colors, positions, and color attribution. Our goal is to compare different test-time control interfaces under a fixed sampling budget. Each method expresses the target constraint through its native interface: RMG uses a fixed visual reference bank of

20 images per category, while search- and gradient-based baselines operate through text prompts, classifier scores, or reward gradients. For compositional categories, RMG banks are assembled from simpler visual components rather than exact target examples; examples are shown in Appendix C.4.

All methods use the same FLUX.2-klein backbone, resolution, sampler, number of steps, prompts, and random seeds. During RMG sampling, no classifier, reward model, LLM, gradient computation, or candidate selection is used. Table 1 reports wall-clock runtime, NFE, and auxiliary model calls per retained sample. RMG improves prompt alignment in a single sampling trajectory, with the largest gains on compositional categories such as position (+28.75) and two-object generation (+8.08), suggesting that a small visual reference bank can provide an efficient structural control signal when the base model struggles with the text constraint alone.

Table 1: Comparison on GenEval using the same FLUX.2 backbone, resolution, prompts, seeds, and 20-step base sampler. RMG uses a fixed 20-image visual reference bank per category; baselines use their native text, classifier, reward, or search interfaces. Time is relative wall-clock runtime with batching where possible. Total NFE is normalized by one baseline generation. Aux. evals counts external-model calls, with C = Classifier and L = LLM.

Method	Time ↓	Total NFE ↓	Aux. evals ↓	Mean ↑	Single ↑	Two ↑	Counting ↑	Colors ↑	Position ↑	Attribution ↑
FLUX.2-klein (4B)	1.00×	**1×**	–	80.10	99.69	91.41	80.62	84.84	65.25	58.75
Search-based										
+ Prompt Opt. [14]	7.87×	8×	8C+2L	84.18	**100.00**	95.45	**88.12**	87.77	69.75	64.00
+ Best-of-4 [12]	4.07×	4×	4C	83.35	99.69	95.96	83.44	88.03	67.75	65.25
+ SMC [13]	6.17×	4×	81C	80.28	99.69	95.71	81.88	85.37	61.75	57.25
Gradient-based										
+ ReNO [15]	19.44×	4×	4C	83.46	99.69	93.18	87.50	90.16	65.50	64.75
Ours (RMG)	**1.02×**	**1×**	–	**91.17**	**100.00**	**99.49**	**88.12**	**90.16**	**94.00**	**75.25**

4.2 Semi-Parametric Guidance

We evaluate SPG (Section 3.3) on AFHQv2, testing whether an amortized reference-set model preserves unconditional generation quality while enabling inference-time control via reference-set substitution. Architecture, training, and dataset details are in Appendices C.1 and C.2. A key motivation is that closed-form reference means can transfer nuisance correlations from the reference bank (e.g. a shared background); in Appendix F.4 we show that SPG preserves object-level guidance without copying such artifacts, whereas RMG does not.

Unconditional quality and reference independence. Fig. 5 shows that SPG matches a DiT-B/4 baseline on AFHQv2, confirming that the reference-set anchor does not degrade generative performance. Comparing generated samples to their nearest neighbors in latent space (Fig. 6) shows that outputs are semantically aligned with references while remaining visually distinct, confirming the reference set acts as a soft conditioning signal rather than a retrieval mechanism.

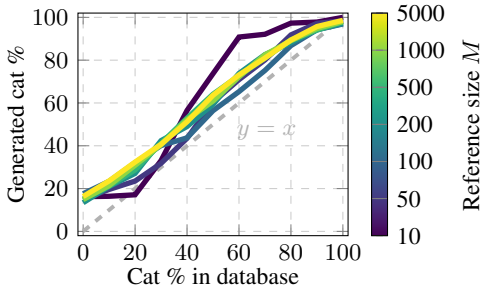
Inference-time control. As shown in Fig. 6, swapping the reference set (e.g., cat-only vs. dog-only) systematically shifts outputs for the same noise seed. Fig. 5(b) quantifies this by varying reference-set composition and measuring generated class frequency over 10,000 CLIP-labeled images. Generated class proportions closely track the reference-set composition across a wide range of reference sizes M, demonstrating that reference-set composition controls the output distribution at inference time. An LPIPS diversity analysis as a function of reference-set size is provided in Appendix F.3.

5 Related Work

Existing approaches to controlling pretrained generative models fall into fine-tuning [6, 7], inference-time guidance through auxiliary models or reward signals [9, 15], and search-based methods [12–14] that trade efficiency for quality. Recent work on endpoint posteriors [11, 24, 25] shares the view that endpoint information governs controllable generation, but operates through scalar rewards and requires training or repeated evaluation. Retrieval-augmented methods [26–28] condition generation on external data, but treat retrieved content as auxiliary context rather than as a control signal. Our approach unifies these perspectives: the reference set defines the endpoint posterior mean

Model	FID ↓	KID ↓	IS ↑
DiT-B/4	23.111	0.012	6.554
SPG	23.256	0.013	6.227

(a) Unconditional generation quality



(b) Control via reference-set composition

Figure 5: SPG preserves unconditional generation quality while enabling inference-time control through the reference set. **(a)** SPG matches DiT-B/4 on AFHQv2 (FID 23.26 vs. 23.11), showing that the reference-set anchor does not degrade generation quality. **(b)** Generated cat percentage versus reference-set composition for different reference sizes M. Each point is estimated over 10,000 samples labeled with a CLIP-based classifier. Generated proportions track and amplify the reference distribution, demonstrating controllability without modifying model parameters.

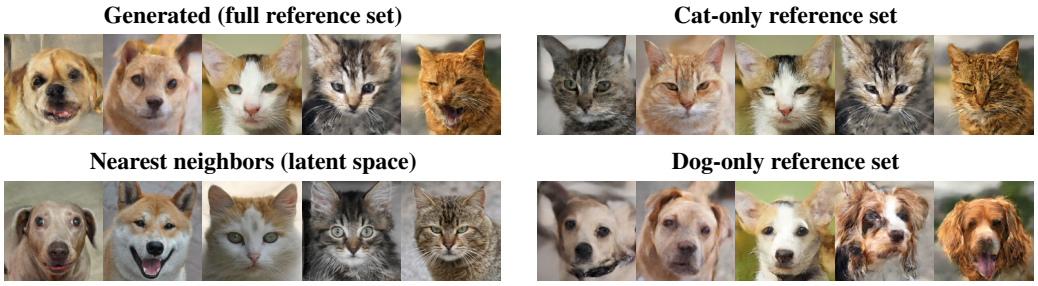


Figure 6: SPG preserves generation quality, avoids memorization, and enables inference-time control through the reference set. **Top-left:** unconditional samples using the full reference set. **Bottom-left:** nearest neighbors in latent space, showing that generations are semantically aligned with references but not copies. **Right:** same model and noise seed with different reference sets, showing that swapping the reference set shifts the generated distribution.

directly, yielding a closed-form drift correction with no reward signal, auxiliary model, or additional evaluations. Under a Gaussian bridge, this posterior mean reduces exactly to a softmax-weighted aggregation over reference points, grounding attention as a conditional expectation and connecting to non-parametric score estimation [20, 21]. An extensive discussion of related work is provided in Appendix B.

6 Limitations

Reference-mean guidance inherits the quality of its reference set: noisy or poorly curated references introduce unwanted artifacts, and computing posterior means over large sets can be costly, though subsampling and approximate retrieval offer practical remedies. Extending the framework to other modalities may require domain-specific design choices. As with any controllability method, responsible curation of reference sets is essential to prevent misuse for harmful or misleading generation.

7 Conclusion

We have shown that control can be framed as a problem of shifting endpoint means. This leads to a simple alternative to fine-tuning, auxiliary guidance, or search: steer generation by changing the reference set over which the model implicitly or explicitly aggregates. Reference-Mean Guidance demonstrates that this principle is already usable in frozen pretrained models, while Semi-Parametric Guidance shows how the same mechanism can be amortized into a learnable architecture without sacrificing generation quality. More broadly, this suggests a path toward generative models that adapt through data rather than parameter updates.

Acknowledgments and Disclosure of Funding

This project was supported by the ELLIS Unit Amsterdam and carried out using the Dutch national e-infrastructure, with the support of SURF through the use of the Snellius supercomputer.

References

[1] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

[2] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

[3] Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=li7qeBbCR1t.

[4] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 4172–4182, Paris, France, October 2023. IEEE. ISBN 979-8-3503-0718-4. doi: 10.1109/ICCV51070.2023.00387. URL https://ieeexplore.ieee.org/document/10377858/.

[5] Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

[6] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[7] 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. URL https://openreview.net/forum?id=nZeVKeeFYf9.

[8] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[9] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=AAWuCvzaVt.

[10] Ruiqi Feng, Chenglei Yu, Wenhao Deng, Peiyan Hu, and Tailin Wu. On the guidance of flow matching. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=pKaNgFzJBy.

[11] Peter Potaptchik, Cheuk-Kit Lee, and Michael S. Albergo. Tilt matching for scalable sampling and fine-tuning, 2025. URL https://arxiv.org/abs/2512.21829.

[12] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. If at first you don't succeed, try, try again: Faithful diffusion-based text-to-image generation by selection, 2023. URL https://arxiv.org/abs/2305.13308.

[13] Luhuan Wu, Brian L. Trippe, Christian A. Naesseth, David M. Blei, and John P. Cunningham. Practical and asymptotically exact conditional sampling in diffusion models, 2024. URL https://arxiv.org/abs/2306.17775.

[14] Oscar Mañas, Pietro Astolfi, Melissa Hall, Candace Ross, Jack Urbanek, Adina Williams, Aishwarya Agrawal, Adriana Romero-Soriano, and Michal Drozdzal. Improving text-to-image consistency via automatic prompt optimization. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=g12Gdl6aDL. Featured Certification.

[15] Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. ReNO: Enhancing one-step text-to-image models through reward-based noise optimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=MXY0qsGge0.

[16] Floor Eijkelboom, Grigory Bartosh, Christian A. Naesseth, Max Welling, and Jan-Willem van de Meent. Variational flow matching for graph generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=UahrHR5HQh.

[17] Anne Gagneux, Ségolène Tiffany Martin, Rémi Gribonval, and Mathurin Massias. Training flow matching: The role of weighting and parameterization. In *ICLR 2026 2nd Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy*, 2026. URL https://openreview.net/forum?id=RYQBTBZxNl.

[18] Quentin Bertrand, Anne Gagneux, Mathurin Massias, and Rémi Emonet. On the closed-form of flow matching: Generalization does not arise from target stochasticity. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=kVz9uvqUna.

[19] Weiguo Gao and Ming Li. How do flow matching models memorize and generalize in sample data subspaces?, 2024. URL https://arxiv.org/abs/2410.23594.

[20] Matthew Niedoba, Dylan Green, Saeid Naderiparizi, Vasileios Lioutas, Jonathan Wilder Lavington, Xiaoxuan Liang, Yunpeng Liu, Ke Zhang, Setareh Dabiri, Adam Scibior, Berend Zwartsenberg, and Frank Wood. Nearest neighbour score estimators for diffusion generative models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 38117–38144. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/niedoba24a.html.

[21] Christopher Scarvelis, Haitz Sáez de Ocáriz Borde, and Justin Solomon. Closed-form diffusion models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=JkMifr17wc.

[22] Daniel Wolf, Heiko Hillenhagen, Billurvan Taskin, Alex Bäuerle, Meinrad Beer, Michael Götz, and Timo Ropinski. Your other Left! Vision-Language Models Fail to Identify Relative Positions in Medical Images. In *Proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15964. Springer Nature Switzerland, September 2025.

[23] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=Wbr51vK331.

[24] Amirmojtaba Sabour, Michael S. Albergo, Carles Domingo-Enrich, Nicholas M. Boffi, Sanja Fidler, Karsten Kreis, and Eric Vanden-Eijnden. Test-time scaling of diffusions with flow maps, 2025. URL https://arxiv.org/abs/2511.22688.

[25] Peter Holderrieth, Douglas Chen, Luca Eyring, Ishin Shah, Giri Anantharaman, Yutong He, Zeynep Akata, Tommi Jaakkola, Nicholas M. Boffi, and Max Simchowitz. Diamond maps: Efficient reward alignment via stochastic flow maps, 2026. URL https://arxiv.org/abs/2602.05993.

[26] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[27] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. Improving language models by retrieving from trillions of tokens. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/borgeaud22a.html.

[28] Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Semi-parametric neural image synthesis. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=Bqk9c0wBNrZ.

[29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[30] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

[31] Alexander Tong, Kilian FATRAS, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=CD9Snc73AW. Expert Certification.

[32] Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, 2024.

[33] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

[34] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i5.28226. URL https://doi.org/10.1609/aaai.v38i5.28226.

[35] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=_CDixzkzeyb.

[36] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22560–22570, October 2023.

[37] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2308.06721.

[38] Dongxu Li, Junnan Li, and Steven C. H. Hoi. BLIP-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In *Advances in Neural Information Processing Systems*, 2023.

[39] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds, 2024. URL https://arxiv.org/abs/2401.07519.

[40] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HklBjCEKvH.

[41] 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.

[42] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[43] James Seale Smith, Yen-Chang Hsu, Lingyu Zhang, Ting Hua, Zsolt Kira, Yilin Shen, and Hongxia Jin. Continual diffusion: Continual customization of text-to-image diffusion with c-loRA. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=TZdEgwZ6f3.

[44] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9229–9248. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/sun20b.html.

[45] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

[46] Zahra Kadkhodaie, Florentin Guth, Eero P Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ANvmVS2Yr0.

[47] Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers, 2025. URL https://arxiv.org/abs/2504.10483.

[48] Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V Le. Symbolic discovery of optimization algorithms. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=ne6zeqLFCZ.

[49] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, 2022.

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

[51] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. URL https://arxiv.org/abs/2409.12191.

Appendix Index

A. Proofs and Derivations

- A.1 Proof of Proposition 3.1
- A.2 Proof of Proposition 3.2
- A.3 Proof of Proposition 3.3
- A.4 Reference-Set Formalism

B. Related Work

C. Experimental Details and Metrics

- C.1 SPG Architecture and Training
- C.2 AFHQv2 Setup
- C.3 Reference-Mean Guidance in FLUX.2
- C.4 GenEval Reference Banks
- C.5 CLIP Attribute Score

D. Mechanistic Validation

E. Ablations

- E.1 Guidance Schedule
- E.2 Schedule Form
- E.3 Guidance Strength and Schedule Shape
- E.4 Reference-Set Size
- E.5 Number of Function Evaluations

F. Additional Experiments

- F.1 Prompt–Reference Interaction
- F.2 Reference Composition
- F.3 SPG Diversity as a Function of Reference-Set Size
- F.4 Suppressing Reference-Set Nuisance Artifacts

G. Reference Banks

A Proofs and Derivations

This appendix provides complete derivations for the three main theoretical results in Section 3. The main text presents the linear bridge for readability; here we state the derivations for the more general affine bridge

$$I_t(x_0, x_1) := \alpha_t x_0 + \beta_t x_1,$$

where α_t and β_t are differentiable scalar schedules satisfying $\alpha_0 = 1$, $\beta_0 = 0$, $\alpha_1 = 0$, and $\beta_1 = 1$. We assume $\alpha_t > 0$ for $t \in [0, 1)$ and write

$$a_t := \frac{\dot{\alpha}_t}{\alpha_t}, \qquad c_t := \dot{\beta}_t - \beta_t \frac{\dot{\alpha}_t}{\alpha_t}.$$

The linear bridge in the main text is the special case $\alpha_t = 1-t$ and $\beta_t = t$, for which $a_t = -1/(1-t)$ and $c_t = 1/(1-t)$. To avoid overloading the affine coefficient β_t, the guidance schedule in this appendix is denoted by γ_t.

A.1 Proof of Proposition 3.1 (Optimal Drift)

We derive the optimal flow-matching velocity under the affine bridge

$$I_t(x_0, x_1) = \alpha_t x_0 + \beta_t x_1, \qquad t \in [0, 1],$$

where $x_0 \sim p_0$ and $x_1 \sim p_1$ are independent.

Step 1: The bridge velocity. Differentiating with respect to t,

$$\dot{x}_t = \dot{\alpha}_t x_0 + \dot{\beta}_t x_1.$$

For the linear bridge this reduces to the displacement $x_1 - x_0$.

Step 2: The marginal velocity field. The flow-matching objective trains a velocity field $u_t^\theta(x)$ to match the conditional velocity $\dot{x}_t$ at each state $x_t = x$. Since many endpoint pairs (x_0, x_1) can produce the same intermediate state $x_t = x$, the loss-minimizing velocity field is the conditional expectation:

$$u_t(x) = \mathbb{E}[\dot{x}_t \mid x_t = x] = \mathbb{E}[\dot{\alpha}_t x_0 + \dot{\beta}_t x_1 \mid x_t = x]. \tag{16}$$

Step 3: Eliminating x_0. From the affine bridge, we can express x_0 in terms of x_t, x_1, and t:

$$x_t = \alpha_t x_0 + \beta_t x_1 \implies x_0 = \frac{x_t - \beta_t x_1}{\alpha_t}, \qquad t < 1.$$

Substituting into Eq. (16),

$$\begin{aligned} u_t(x) &= \mathbb{E}\left[\dot{\alpha}_t \frac{x_t - \beta_t x_1}{\alpha_t} + \dot{\beta}_t x_1 \,\middle|\, x_t = x\right] \\ &= \mathbb{E}\left[\frac{\dot{\alpha}_t}{\alpha_t} x_t + \left(\dot{\beta}_t - \beta_t \frac{\dot{\alpha}_t}{\alpha_t}\right) x_1 \,\middle|\, x_t = x\right]. \end{aligned}$$

Step 4: Expressing in terms of the conditional mean. Since $x_t = x$ is fixed under the conditional expectation, and the endpoint mean is $\mu_t(x) := \mathbb{E}[x_1 \mid x_t = x]$, we obtain

$$u_t(x) = \frac{\dot{\alpha}_t}{\alpha_t} x + \left(\dot{\beta}_t - \beta_t \frac{\dot{\alpha}_t}{\alpha_t}\right) \mu_t(x) = a_t x + c_t \mu_t(x).$$

For the linear bridge, this specializes to $u_t(x) = (\mu_t(x) - x)/(1-t)$. This completes the proof. □

Remark. The expression can become singular when $\alpha_t \to 0$ as $t \to 1$. For the linear bridge this appears as the $(1-t)^{-1}$ factor. In practice, numerical integration terminates before $t = 1$.

A.2 Proof of Proposition 3.2 (Posterior Mean as Cross-Attention)

We derive the closed-form endpoint mean under a standard normal source $p_0 = \mathcal{N}(0, I)$ and the empirical target $\hat{p}_1$ over the training set $\mathcal{D} = \{x^{(n)}\}_{n=1}^N$, and show that it is equivalent to cross-attention over the dataset.

Step 1: The bridge conditional distribution. Under the affine bridge $x_t = \alpha_t x_0 + \beta_t x_1$ with $x_0 \sim \mathcal{N}(0, I)$ and $x_1 = x^{(n)}$ fixed, the intermediate state x_t is conditionally Gaussian:

$$x_t \mid x^{(n)} \sim \mathcal{N}\big(\beta_t x^{(n)},\, \alpha_t^2 I\big).$$

This follows because $x_t = \alpha_t x_0 + \beta_t x^{(n)}$ is a linear function of $x_0 \sim \mathcal{N}(0, I)$, giving mean $\beta_t x^{(n)}$ and covariance $\alpha_t^2 I$.

Evaluated at the intermediate state $x_t = x$, the conditional density given endpoint $x^{(n)}$ is

$$p_t(x \mid x^{(n)}) = \frac{1}{(2\pi)^{d/2}\alpha_t^d} \exp\left(-\frac{\|x - \beta_t x^{(n)}\|^2}{2\alpha_t^2}\right). \tag{17}$$

Step 2: Bayes' rule for the posterior weights. Under $\hat{p}_1$, each data point $x^{(n)}$ has prior probability $1/N$. By Bayes' rule, the posterior probability that the endpoint is $x^{(n)}$ given the intermediate state $x_t = x$ is

$$w_t^{(n)}(x) := p\Big(x_1 = x^{(n)} \mid x_t = x\Big) = \frac{p_t(x \mid x^{(n)}) \cdot \frac{1}{N}}{\sum_{j=1}^N p_t(x \mid x^{(j)}) \cdot \frac{1}{N}} = \frac{p_t(x \mid x^{(n)})}{\sum_{j=1}^N p_t(x \mid x^{(j)})}.$$

The $1/N$ factors cancel, and substituting Eq. (17),

$$w_t^{(n)}(x) = \frac{\exp\left(-\frac{\|x - \beta_t x^{(n)}\|^2}{2\alpha_t^2}\right)}{\sum_{j=1}^N \exp\left(-\frac{\|x - \beta_t x^{(j)}\|^2}{2\alpha_t^2}\right)}. \tag{18}$$

The normalizing constants $(2\pi)^{d/2}\alpha_t^d$ cancel in the ratio.

Step 3: The conditional endpoint mean. The conditional endpoint mean is the expectation of x_1 under the posterior:

$$\hat{\mu}_t(x) = \mathbb{E}_{\hat{p}_1}[x_1 \mid x_t = x] = \sum_{n=1}^N w_t^{(n)}(x)\, x^{(n)}. \tag{19}$$

Step 4: Expanding the exponent. We simplify the weights by expanding the squared norm in the exponent:

$$\|x - \beta_t x^{(n)}\|^2 = \|x\|^2 - 2\beta_t \langle x, x^{(n)}\rangle + \beta_t^2 \|x^{(n)}\|^2.$$

The term $\|x\|^2$ depends only on the current state and not on n, so it contributes equally to every term in the softmax numerator and denominator. It therefore cancels:

$$\exp\left(-\frac{\|x - \beta_t x^{(n)}\|^2}{2\alpha_t^2}\right) \propto \exp\left(\frac{\beta_t \langle x, x^{(n)}\rangle}{\alpha_t^2} - \frac{\beta_t^2 \|x^{(n)}\|^2}{2\alpha_t^2}\right),$$

where $\propto$ means up to a multiplicative constant independent of n. The weights become

$$w_t^{(n)}(x) = \frac{\exp\left(\frac{\beta_t}{\alpha_t^2}\langle x, x^{(n)}\rangle - \frac{\beta_t^2}{2\alpha_t^2}\|x^{(n)}\|^2\right)}{\sum_{j=1}^N \exp\left(\frac{\beta_t}{\alpha_t^2}\langle x, x^{(j)}\rangle - \frac{\beta_t^2}{2\alpha_t^2}\|x^{(j)}\|^2\right)}. \tag{20}$$

Step 5: Cross-attention identification. Define the query, keys, values, and biases as

$$q := \frac{\beta_t}{\alpha_t^2} x, \qquad k_n := x^{(n)}, \qquad v_n := x^{(n)}, \qquad b_n := -\frac{\beta_t^2}{2\alpha_t^2}\|x^{(n)}\|^2.$$

Then the exponent in Eq. (20) is $q^\top k_n + b_n$, and the weights are

$$w_t^{(n)}(x) = \frac{\exp(q^\top k_n + b_n)}{\sum_{j=1}^N \exp(q^\top k_j + b_j)} =: \alpha_n(q).$$

The empirical endpoint mean in Eq. (19) is therefore

$$\hat{\mu}_t(x) = \sum_{n=1}^{N} \alpha_n(q)\, v_n,$$

which is exactly a cross-attention operation with query q, keys $\{k_n\}$, values $\{v_n\}$, and per-key biases $\{b_n\}$. □

Remark on the bias term. The bias $b_n = -\frac{\beta_t^2}{2\alpha_t^2}\|x^{(n)}\|^2$ penalizes data points with large norm. In standard dot-product attention this term is absent; here it arises naturally from the Gaussian bridge geometry and acts as a length normalization on the keys. Setting $\alpha_t = 1 - t$ and $\beta_t = t$ recovers the main-text weights in Eq. (6).

A.3 Proof of Proposition 3.3 (Reference-Mean Guided Dynamics)

We derive the guided velocity field arising from a geometric mixture and show that the correction depends only on a difference of posterior means. We distinguish two constructions. The *endpoint-level geometric mixture* $\pi(x_1) \propto p_1(x_1)^{1-\gamma_t}\rho_1(x_1)^{\gamma_t}$ defines a valid bridge marginal $\pi_t(x) = \int p_t(x \mid x_1)\pi(x_1)\,dx_1$ by construction; under a Gaussian posterior approximation it recovers the same velocity formula derived below, and the approximation is exact when p_1 and ρ_1 are Gaussian. The *marginal-level geometric mixture* $\pi_t(x) \propto p_t(x)^{1-\gamma_t}\rho_t(x)^{\gamma_t}$ is not generally a valid marginal of the same affine bridge, but admits a clean algebraic derivation via log-linear score interpolation that we present below. Both constructions yield the same guided velocity, which we use as a principled guidance rule.

Setup. Let $x_1 \sim \rho_1$ be a general endpoint distribution, and define the noisy marginal under the affine bridge $x_t = \alpha_t x_0 + \beta_t x_1$, $x_0 \sim \mathcal{N}(0, I)$, as

$$\rho_t(x) = \int p_t(x \mid x_1)\,\rho_1(x_1)\,dx_1, \qquad p_t(x \mid x_1) = \mathcal{N}\big(x;\, \beta_t x_1,\, \alpha_t^2 I\big).$$

The corresponding velocity field is $u_t^\rho(x) = a_t x + c_t \mu_t^\rho(x)$, where $\mu_t^\rho(x) = \mathbb{E}_{\rho_1}[x_1 \mid x_t = x]$ is the posterior mean.

Step 1: Score of the noisy marginal. We compute $\nabla_x \log \rho_t(x)$ by differentiating under the integral sign:

$$\nabla_x \log \rho_t(x) = \frac{\nabla_x \rho_t(x)}{\rho_t(x)} = \frac{\int \nabla_x p_t(x \mid x_1)\,\rho_1(x_1)\,dx_1}{\rho_t(x)}.$$

Since $p_t(x \mid x_1) = \mathcal{N}(x;\, \beta_t x_1,\, \alpha_t^2 I)$, its score with respect to x is

$$\nabla_x \log p_t(x \mid x_1) = -\frac{x - \beta_t x_1}{\alpha_t^2},$$

so $\nabla_x p_t(x \mid x_1) = p_t(x \mid x_1) \cdot \left(-\frac{x-\beta_t x_1}{\alpha_t^2}\right)$. Therefore,

$$\begin{aligned}
\nabla_x \log \rho_t(x) &= \frac{1}{\rho_t(x)} \int p_t(x \mid x_1) \left(-\frac{x - \beta_t x_1}{\alpha_t^2}\right) \rho_1(x_1)\,dx_1 \\
&= \int \frac{p_t(x \mid x_1)\rho_1(x_1)}{\rho_t(x)} \left(-\frac{x - \beta_t x_1}{\alpha_t^2}\right) dx_1 \\
&= \mathbb{E}_{\rho_1}\left[-\frac{x - \beta_t x_1}{\alpha_t^2} \;\middle|\; x_t = x\right] \\
&= \frac{\beta_t \mu_t^\rho(x) - x}{\alpha_t^2}. \qquad (21)
\end{aligned}$$

Rearranging Eq. (21) gives the score-to-mean identity:

$$\mu_t^\rho(x) = \frac{x + \alpha_t^2 \nabla_x \log \rho_t(x)}{\beta_t}. \qquad (22)$$

This gives the score-to-mean identity, allowing us to move between score functions and posterior means.

Step 2: Score of the geometric mixture. Let p_1 and ρ_1 denote the training and reference endpoint distributions, with noisy marginals p_t and ρ_t. Define the log-linear guided density

$$\pi_t(x) \propto \big(p_t(x)\big)^{1-\gamma_t} \big(\rho_t(x)\big)^{\gamma_t},$$

for a guidance schedule $\gamma_t \in [0,1]$. Taking the logarithm and differentiating,

$$\nabla_x \log \pi_t(x) = (1-\gamma_t)\nabla_x \log p_t(x) + \gamma_t \nabla_x \log \rho_t(x). \tag{23}$$

The score of the geometric mixture is the convex combination of the base and reference scores.

Step 3: Score-implied guided endpoint mean. Applying the Gaussian bridge score-to-mean map in Eq. (22) to the score of π_t,

$$\begin{aligned}
\mu_t^{\pi}(x) &= \frac{x + \alpha_t^2 \nabla_x \log \pi_t(x)}{\beta_t} \\
&= \frac{x + \alpha_t^2 \big[(1-\gamma_t)\nabla_x \log p_t(x) + \gamma_t \nabla_x \log \rho_t(x)\big]}{\beta_t} \\
&= (1-\gamma_t) \cdot \frac{x + \alpha_t^2 \nabla_x \log p_t(x)}{\beta_t} + \gamma_t \cdot \frac{x + \alpha_t^2 \nabla_x \log \rho_t(x)}{\beta_t} \\
&= (1-\gamma_t)\mu_t(x) + \gamma_t \mu_t^{\rho}(x).
\end{aligned} \tag{24}$$

The resulting score-implied endpoint mean is the convex combination of the training and reference posterior means. Note that the linearity of Eq. (22) in the score is what allows the mixture to pass through cleanly.

Step 4: Guided velocity field. Substituting Eq. (24) into the affine-bridge velocity parameterization $u_t^{\pi}(x) = a_t x + c_t \mu_t^{\pi}(x)$ gives the guided velocity

$$\begin{aligned}
u_t^{\pi}(x) &= a_t x + c_t \mu_t^{\pi}(x) \\
&= a_t x + c_t \big[(1-\gamma_t)\mu_t(x) + \gamma_t \mu_t^{\rho}(x)\big] \\
&= u_t(x) + \gamma_t c_t \big(\mu_t^{\rho}(x) - \mu_t(x)\big).
\end{aligned} \tag{25}$$

For the linear bridge this becomes $u_t^{\pi}(x) = u_t(x) + \gamma_t(\mu_t^{\rho}(x) - \mu_t(x))/(1-t)$, matching Eq. (9) after identifying γ_t with the main-text guidance schedule.

Step 5: Empirical reference set. When ρ_1 is an empirical distribution $\hat{\rho}_1 = \frac{1}{M}\sum_{m=1}^{M} \delta_{x^{(m)}}$, the derivation of Appendix A.2 applies directly. The empirical reference posterior mean is

$$\hat{\mu}_t^{\rho}(x) = \sum_{m=1}^{M} w_t^{(m)}(x)\, x^{(m)},$$

where $w_t^{(m)}(x)$ are the softmax weights from Eq. (20) computed with respect to the reference set $\mathcal{R} = \{x^{(m)}\}_{m=1}^{M}$. Substituting into Eq. (25) gives the final score-motivated guided velocity in closed form. □

Remark on late-time instability. The velocity correction $\gamma_t c_t(\mu_t^{\rho}(x) - \mu_t(x))$ can grow as $t \to 1$ when c_t diverges. Simultaneously, the reference posterior $w_t^{(m)}(x)$ concentrates sharply around the nearest reference point as the bandwidth α_t^2 in Eq. (18) vanishes. For the linear bridge, $c_t = 1/(1-t)$, so this motivates schedules of the form $\gamma_t = \gamma_0(1-t)^{\alpha}$ for some $\alpha \geq 1$, which cancel the $(1-t)^{-1}$ divergence and ensure bounded corrections throughout the trajectory.

Remark on validity. The marginal-level geometric mixture $\pi_t \propto p_t^{1-\gamma_t}\rho_t^{\gamma_t}$ need not satisfy the continuity equation for the original bridge family, so u_t^{π} should be interpreted as a score-motivated guidance rule rather than an exact probability-flow velocity. The endpoint-level construction $\pi(x_1) \propto p_1(x_1)^{1-\gamma_t}\rho_1(x_1)^{\gamma_t}$ avoids this issue: its marginal is valid by construction and recovers the same velocity formula under a Gaussian posterior approximation, which holds exactly in Gaussian latent spaces such as those used in the FLUX.2 experiments.

A.4 Reference-Set Formalism

We now present a unified view of flow matching in terms of reference sets. This formalism makes explicit the role of data in defining the posterior mean and clarifies the relationship between standard flow matching, reference-mean guidance (RMG), and Semi-Parametric Guidance (SPG).

A.4.1 Setup and Empirical Posterior Means

Let $\{x^{(n)}\}_{n=1}^{N}$ denote a training set drawn i.i.d. from p_1, and let $\mathcal{R} = \{x^{(m)}\}_{m=1}^{M}$ denote a reference set drawn from a (possibly different) distribution ρ_1.

Under the affine bridge

$$x_t = \alpha_t x_0 + \beta_t x_1, \qquad x_0 \sim \mathcal{N}(0, I),$$

each distribution over endpoints induces a noisy marginal

$$\rho_t(x) = \int p_t(x \mid x_1)\, \rho_1(x_1)\, dx_1, \qquad p_t(x \mid x_1) = \mathcal{N}(x;\, \beta_t x_1,\, \alpha_t^2 I).$$

For the empirical distributions

$$\hat{p}_1 = \frac{1}{N}\sum_{n=1}^{N} \delta_{x^{(n)}}, \qquad \hat{\rho}_1 = \frac{1}{M}\sum_{m=1}^{M} \delta_{x^{(m)}},$$

the corresponding posterior means take the Nadaraya–Watson form

$$\hat{\mu}_t(x) = \frac{\sum_{n=1}^{N} p_t(x \mid x^{(n)})\, x^{(n)}}{\sum_{n=1}^{N} p_t(x \mid x^{(n)})}, \qquad \hat{\mu}_t^{\rho}(x) = \frac{\sum_{m=1}^{M} p_t(x \mid x^{(m)})\, x^{(m)}}{\sum_{m=1}^{M} p_t(x \mid x^{(m)})}.$$

By Appendix A.1, both define velocity fields

$$u_t(x) = a_t x + c_t \mu_t(x), \qquad u_t^{\rho}(x) = a_t x + c_t \mu_t^{\rho}(x).$$

A.4.2 Self-Referencing and Standard Flow Matching

Standard flow matching corresponds to the *self-referencing* case, where the posterior mean is computed with respect to the training distribution:

$$u_t^{\theta}(x) \approx u_t(x).$$

In practice, a neural network is trained to approximate $\mu_t(x)$, implicitly encoding the training distribution p_1 into model parameters.

A.4.3 Geometric Mixture at the Endpoint Level (Primary Construction)

The primary construction underlying Proposition 3.3 defines the geometric mixture at the endpoint level,

$$\pi(x_1) \propto p_1(x_1)^{1-\gamma_t}\, \rho_1(x_1)^{\gamma_t},$$

and constructs the marginal in the standard way:

$$\pi_t(x) = \int p_t(x \mid x_1)\, \pi(x_1)\, dx_1.$$

This is a valid bridge marginal by construction, so the score-velocity identity applies. Under a Gaussian posterior approximation — exact when p_1 and ρ_1 are Gaussian, as is approximately the case in VAE latent spaces — the posterior mean under π is

$$\mu_t^{\pi}(x) = (1-\gamma_t)\mu_t(x) + \gamma_t \mu_t^{\rho}(x), \tag{26}$$

giving the guided velocity

$$u_t^{\pi}(x) = u_t(x) + \gamma_t c_t \big(\mu_t^{\rho}(x) - \mu_t(x)\big).$$

The full algebraic derivation via log-linear score interpolation is given in Appendix A.3.

A.4.4 Arithmetic Mixture (Alternative Exact Construction)

An alternative exact construction uses the arithmetic mixture

$$\hat{p}_\lambda = (1-\lambda)\hat{p}_1 + \lambda\hat{\rho}_1, \qquad \lambda \in [0,1],$$

whose noisy marginal is

$$p_t^\lambda(x) = (1-\lambda)p_t(x) + \lambda\rho_t(x).$$

Applying Bayes' rule, the corresponding posterior mean is

$$\mu_t^\lambda(x) = (1-\omega_t^*(x))\mu_t(x) + \omega_t^*(x)\mu_t^\rho(x), \qquad \omega_t^*(x) = \frac{\lambda\rho_t(x)}{(1-\lambda)p_t(x) + \lambda\rho_t(x)}. \tag{27}$$

This is exact with no distributional approximation, but requires evaluating $p_t(x)$ as a scalar density, which is not directly accessible from a pretrained velocity field. Replacing $\omega_t^*(x)$ with a scalar γ_t recovers the same guided velocity as the geometric construction, confirming that both support the same guidance rule.

Special case: union of sets. When $\lambda = \frac{M}{N+M}$, $\hat{p}_\lambda$ corresponds to the uniform distribution over training and reference points, and μ_t^λ reduces to the Nadaraya–Watson estimator over all points in both sets.

A.4.5 Relationship to RMG and SPG

This formalism clarifies the relationship between the methods in the main paper:

- **Standard FM (self-referencing):** approximates μ_t with a neural network trained on the full dataset.
- **RMG (cross-referencing, test-time):** uses the guided velocity

  $$u_t^\pi(x) = u_t(x) + \gamma_t c_t\big(\mu_t^\rho(x) - \mu_t(x)\big),$$

  where μ_t is provided by a pretrained model and μ_t^ρ is replaced by the empirical reference mean $\hat{\mu}_t^\rho$ computed in closed form from $\mathcal{R}$. The main text specializes this to the linear bridge and denotes the guidance schedule by β_t.
- **SPG (amortized cross-referencing):** learns to approximate μ_t^ρ through an attention-based anchor, while a parametric refiner captures corrections beyond the explicit mean.

From this perspective, guidance arises from differences between posterior means defined over distinct reference sets. Standard flow matching corresponds to the degenerate case $\hat{\rho}_1 = \hat{p}_1$, while reference-guided methods exploit the flexibility of choosing $\mathcal{R}$ independently at test time.

B Related Work

B.1 Flow Matching, Diffusion Models, and Endpoint Posteriors

Flow matching learns a velocity field along a prescribed probability path [1, 3, 2], while related diffusion and score-based models learn reverse-time dynamics or score fields [29, 30]. Conditional flow matching generalizes to conditional paths [31], rectified flow emphasizes straight paths [2], stochastic interpolants unify deterministic and stochastic dynamics [3], and scalable interpolant transformers combine these objectives with modern transformer backbones [32].

Our analysis builds on the observation that the optimal velocity field under a linear bridge is fully determined by the conditional endpoint mean. Variational flow matching formalizes this posterior perspective [16]. Closed-form analyses further show that finite-sample flows and scores can be written as kernel-weighted aggregations over training examples [21, 20, 18, 19], revealing the non-parametric structure implicit in trained generative models. We turn this structure into a control mechanism: rather than only analyzing the posterior induced by the training set, we modify the reference set that defines the posterior mean at test time.

B.2 Guidance and Controllable Generation

Classifier guidance steers generation by adding gradients from an external classifier to the score field [9], while classifier-free guidance interpolates conditional and unconditional predictions to improve prompt adherence [33]. Recent work extends guidance theory to flow matching [10]. Our formulation is complementary to these approaches: rather than deriving the correction from an auxiliary classifier or reward, we express it as a difference between endpoint means, requiring no additional model or gradient computation.

A second family augments pretrained generators with auxiliary conditioning networks. ControlNet [8] and T2I-Adapter [34] add trainable branches for spatial conditions such as edges, depth, or pose, while Prompt-to-Prompt [35] and MasaCtrl [36] manipulate attention maps to preserve structure during editing. IP-Adapter [37], BLIP-Diffusion [38], and InstantID [39] condition generation on reference images through dedicated encoders and attention pathways. These methods are powerful but rely on additional trained modules, whereas RMG affects generation only through the posterior mean induced by the reference set, with no auxiliary conditioner.

A third family performs search or optimization at inference time. Best-of-N [12] selects among multiple candidates, while SMC [13] resamples trajectories using external scoring. Prompt optimization [14] searches over improved prompts, and ReNO [15] optimizes initial noise via reward gradients. These approaches improve controllability without full fine-tuning, but increase inference cost through repeated sampling, scoring, or optimization. RMG instead modifies a single trajectory through a closed-form correction, requiring no reward model, classifier, or search over candidates.

A closely related line of work studies reward tilting and flow-map alignment. Tilt Matching [11] derives velocity fields for reward-tilted endpoint distributions via regression, while FMTT [24] uses a flow-map lookahead to guide trajectories toward high-reward endpoints. Diamond Maps [25] learn stochastic flow maps for scalable reward alignment through SMC and guidance. These methods share with ours the view that endpoint information is central to controllable generation, but they operate through scalar rewards and require training or repeated evaluation. RMG instead uses an empirical reference set to define the endpoint posterior mean directly, yielding a data-defined drift correction without reward gradients, value estimation, or Monte Carlo rollouts.

B.3 Retrieval-Augmented and Semi-Parametric Generation

kNN-LM interpolates language model distributions with nearest-neighbour datastore statistics [40], while RAG conditions generation on retrieved documents [26] and RETRO scales this idea to large corpora [27]. In vision, retrieval-augmented diffusion models condition image synthesis on retrieved visual neighbors [28]. All of these methods feed retrieved content into the model as additional context, whereas our work uses the reference set to define a posterior statistic of the generative path itself. In RMG the reference set is compressed into a closed-form endpoint mean, while in SPG the same idea is amortized through an attention-based anchor and a learned residual refiner, occupying a middle ground between purely non-parametric guidance and fully parametric conditional generation.

B.4 Personalization and Low-Data Adaptation

Textual Inversion [41], DreamBooth [6], Custom Diffusion [42], and LoRA [7] adapt model parameters to new concepts from few examples. These methods are effective for personalization, but write new information into weights, which creates challenges when concepts must be added, removed, or recombined frequently. C-LoRA addresses the continual customization setting [43], and test-time adaptation methods handle distribution shift by updating parameters or normalization statistics at deployment time [44, 45]. Reference-guided flows suggest a different interface: rather than answering distribution shift with parameter updates, the model is left fixed and adaptation is achieved by changing the reference set, making it a data operation rather than an optimization.

B.5 Attention as Posterior Aggregation

Under a Gaussian bridge with an empirical target distribution, the endpoint posterior mean is a softmax-weighted average over data points, which is algebraically identical to cross-attention with the noisy state as query and reference examples as keys and values. Similar structures appear in finite-sample analyses of diffusion and score estimation [20, 21]. This connection gives a probabilistic interpretation of the reference-attention module in SPG: rather than treating cross-attention merely as an architectural choice, we use it as an amortized approximation to the posterior mean induced by the reference distribution, with the residual refiner capturing effects that go beyond this explicit anchor.

C Experimental Details and Metrics

C.1 SPG Architecture and Training

SPG augments a standard flow-matching model with a reference-set attention module that approximates the posterior-mean anchor, followed by a learned residual refiner.

Given an input image and a reference set $\mathcal{R} = \{x^{(i)}\}_{i=1}^{M}$, we encode both into latent space using a frozen VAE encoder $\mathcal{E}$. To match the notation of the main text, we write the encoded endpoint as x_1 and the encoded references as $x^{(i)}$. We sample $t \sim \mathcal{U}[0, 1-\epsilon]$ with a small endpoint cutoff $\epsilon > 0$, sample $x_0 \sim \mathcal{N}(0, I)$, and construct

$$x_t = (1-t)x_0 + tx_1. \tag{28}$$

The posterior-mean estimate is computed via cross-attention over the reference set:

$$\bar{x} = \mathrm{Attn}(\tilde{q}, \tilde{k}, \tilde{v}), \tag{29}$$

where x_t provides queries and the reference latents provide keys and values. The final endpoint prediction combines the noisy state, the posterior-mean anchor, and a learned correction via time-dependent gates:

$$\mu_t^\theta(x_t, \mathcal{R}) = (1-g_t)\cdot x_t + g_t \cdot \bar{x} + \alpha_t \cdot f^\theta(\bar{x}, x_t, t), \tag{30}$$

where $g_t, \alpha_t \in [0,1]$ are scalar time-dependent gates, and f^θ predicts a positive residual correction to the anchor.

Training. The model is trained end-to-end on a batch-level endpoint prediction objective evaluated on the full combined output:

$$\mathcal{L}_\mu(\theta) = \mathbb{E}\left[\sum_{m=1}^{M} \frac{1}{(1-t)^2}\left\| x^{(m)} - \mu_t^\theta\left(x_t^{(m)}, \mathcal{R}^{\setminus\{m\}}\right)\right\|^2\right]. \tag{31}$$

The cutoff on t keeps the endpoint-weighted objective finite in empirical training. To prevent reliance on individual references, we apply random masking to the reference set during training. The leave-one-out structure prevents $x_t^{(m)}$ from attending to its own endpoint.

Because the anchor $\bar{x}$ is already a strong predictor of the endpoint, the refiner receives little gradient signal from $\mathcal{L}_\mu$ alone. We therefore train it separately on the positive residual between ground truth and anchor, with gradients stopped through the anchor:

$$\mathcal{L}_{\mathrm{ref}}(\theta) = \mathbb{E}\left[\sum_{m=1}^{M}\left\|\mathrm{sg}\left[x^{(m)} - \bar{x}^{(m)}\right] - f^\theta\left(\mathrm{sg}\left[\bar{x}^{(m)}\right], x_t^{(m)}, t\right)\right\|^2\right], \tag{32}$$

where $\mathrm{sg}[\cdot]$ denotes stop-gradient. Since f^θ is trained to predict the positive residual $x^{(m)} - \bar{x}^{(m)}$ and this is added to $\bar{x}$ in the forward pass, the prediction correctly moves toward the ground truth. Gradients from $\mathcal{L}_\mu$ update both the attention anchor and the refiner jointly, while $\mathcal{L}_{\mathrm{ref}}$ provides an additional signal to the refiner with the anchor held fixed via stop-gradient.

Why the network does not collapse to standard flow matching. At training time, $\mathcal{R}^{(b)}$ is i.i.d. from p_1. By de Finetti's theorem, any exchangeable sequence is conditionally i.i.d. given a latent directing measure — here, the reference distribution ρ_1. Since $\rho_1 = p_1$ during training, the reference set carries no information beyond what p_1 itself encodes. A Bayesian-optimal network with infinite capacity would therefore learn to ignore the reference set and collapse to standard flow matching. In practice this collapse does not occur. The reason is the same one that prevents flow matching from memorizing the training distribution: finite network capacity forces the learned endpoint-mean operator to be a smooth function of its inputs, and the resulting imperfect approximation retains sensitivity to the reference composition [18, 46, 20]. At test time, when the reference set is drawn from $\hat{\rho}_1 \neq p_1$, the learned operator steers generation accordingly. Section 4.2 confirms this empirically.

C.2 AFHQv2 Setup

We evaluate SPG on AFHQv2 using the full training split of `huggan/AFHQv2` (all dog and cat images), encoded with a frozen REPA-E VAE encoder [47] to 256×256 latents. The primary reference bank contains all dog and cat images.

Architecture. The cross-attention module uses a single block with patchwise retrieval (patch size 2, decoupled embedding), 8 heads, qk_dim $= 768$, mlp_ratio $= 1$, and DB dropout $p = 0.1$ during training. Gates g_t and α_t are learned MLPs initialized at 0.5. The residual refiner is a DiT-style transformer with 11 blocks, embedding dimension 768, 12 heads, patch size 2, and MLP ratio 4.

Training. We train for 10,000 steps on 4 A100 GPUs with batch size 64 and bf16 mixed precision. We use the Lion optimizer [48] with learning rate 10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and no weight decay. EMA decay is 0.9999. Gradient clipping is applied at norm 1.0, the refiner penalty weight is 0.1, and self-masking is enabled to prevent each training sample from attending to itself in the reference bank.

Evaluation. FID and KID are computed using `clean-fid` [49] between generated and real images from the training split. CLIP-based class frequency is estimated over 10,000 generated samples using prompts *"a photo of a dog"* and *"a photo of a cat"*.

C.3 Reference-Mean Guidance in FLUX.2

FLUX.2 is a latent rectified-flow model: its sampler evolves a latent state under the linear bridge $x_t = (1-t)x_0 + tx_1$, so the endpoint recovery $\mu_t^\theta(x) = x + (1-t)u_t^\theta(x)$ is the native parameterization of the model. Reference images are encoded with the same frozen VAE into the same packed latent representation as the sampler state, so the reference posterior mean $\hat{\mu}_t^\rho(x)$ and the velocity correction are defined in the same coordinate system as the pretrained model.

For all FLUX.2 experiments, we recover the model endpoint estimate as $\mu_t^\theta(x) = x + (1-t)u_t^\theta(x)$ and replace the reference endpoint mean by the empirical estimate $\hat{\mu}_t^\rho(x)$ computed over the selected reference bank. The practical guided update is

$$u_t^\pi(x) = u_t^\theta(x) + \beta_t \frac{\hat{\mu}_t^\rho(x) - \mu_t^\theta(x)}{1-t}.$$

Unless otherwise stated, we use a quadratic schedule $\beta_t = \beta_0(1-t)^2$ and clip the guidance after $t = 0.85$ to avoid the late-time instability described in Appendix A.3. In latent-space experiments, the reference softmax in Eq. (6) is evaluated with a temperature $\tau = \sqrt{d}$,

$$w_t^{(m)}(x) = \mathrm{Softmax}_m\left(-\frac{1}{2\tau}\frac{\|x - tx^{(m)}\|^2}{(1-t)^2}\right),$$

where d is the latent dimensionality. The dominant term in the squared distance is an inner product between latent vectors, whose variance grows with d; dividing by $\tau = \sqrt{d}$ follows the same rationale as scaled dot-product attention and prevents the softmax from collapsing to a hard nearest-neighbour in high dimension. This temperature is fixed across all FLUX.2 experiments.

C.4 GenEval Reference Banks

For the GenEval comparison in Table 1, we use one fixed reference bank of 20 images per category and reuse that bank across all prompts in the category. For compositional categories, the bank is not required to contain exact target examples. Instead, we assemble banks from simpler visual components whose combined posterior-mean shift encourages the desired composition. Fig. 7 shows representative examples for spatial relations in which the bank provides directional evidence without containing exact target-distribution samples. This protocol intentionally gives RMG an example-based structural prior. The baselines receive the same prompt, seed, sampler, and backbone, but not a visual bank; their control signal must come from text, search, gradients, or external scores. We therefore interpret the GenEval table as measuring whether a small reference bank is an efficient control interface, especially for spatial relations where scalar reward signals are difficult to define reliably.

C.5 CLIP Attribute Score

For the prompt–reference interaction experiment in Fig. 17, we quantify the pink attribute with a directional CLIP score [50]. Let $f_{\mathrm{img}}(x)$ denote the normalized CLIP image embedding of image x, and let $f_{\mathrm{text}}(p)$ denote the normalized CLIP text embedding of prompt p. We define

$$s_{\mathrm{pink}}(x) = \cos\big(f_{\mathrm{img}}(x), f_{\mathrm{text}}(p_{\mathrm{pink}})\big) - \cos\big(f_{\mathrm{img}}(x), f_{\mathrm{text}}(p_{\mathrm{gray}})\big),$$

Table 2: Hyperparameters per experiment. All experiments use FLUX.2-klein (4B), resolution 768×768, and the schedule and softmax temperature described above.

Experiment	**Prompt**	**Schedule**	β_0
Bank swaps	*an elephant in a jungle*	quadratic	0.5
	a house in a forest		1.0
	a cat		1.0
	an animal in a savanna		1.0
Prompt–reference	*an elephant in a jungle* *a pink elephant in a jungle*	quadratic	0.5
Geometric control	*a miniature forest with tall pine trees, a glowing campfire, and fireflies drifting in the night sky, all inside a keyhole on a black background*	quadratic	0.2
Hand pose control	*a hand doing the sign of the horns*	quadratic	0.8
Gymnastics pose control	*a gymnast performing a ring leap, full body visible, airborne, one leg extended forward, the back leg bent high behind the head, arched back, pointed toes, arms extended, dynamic sports photograph*	quadratic	0.2
Controllability	*an animal in a savanna* *an elephant in a jungle*	quadratic	1.0

where $p_{\mathrm{pink}} =$ *"a pink elephant in a jungle"* and $p_{\mathrm{gray}} =$ *"a gray elephant in a jungle"*. Positive values therefore indicate that the image is more similar to the pink prompt than to the gray prompt, and larger values correspond to stronger pinkness. This score is used only as a continuous attribute proxy, not as a calibrated classifier.



Figure 7: Examples from the GenEval protocol. Each column shows a representative reference-bank sample above an RMG generation for the corresponding spatial prompt. The banks provide directional guidance for the target composition without containing exact target-distribution examples.

D Mechanistic Validation

This appendix provides controlled mechanistic experiments supporting the claim in Section 3: the posterior mean determines the flow, and changing the reference set changes generation by changing this mean. We use settings where the posterior mean can be computed exactly, so the effect of the reference set can be isolated without modeling error.

Two moons. The two-moons experiment uses $N = 500$ samples from the two-moons distribution. Labels exist but are withheld from the model; a small labeled reference set is used only to compute soft posterior weights at inference time.

At small t, the posterior is diffuse, while as $t \to 1$ it concentrates around the underlying class structure (Fig. 8, top). This directly changes the posterior mean and the induced flow.

To isolate this effect, Fig. 8 (bottom) visualizes the flow field and trajectories under different reference compositions. Changing only the reference set reverses the direction of the flow and changes the final attractor, providing direct evidence that the posterior mean controls the flow.

We further compare three inference-time conditions: *unconditional weighting*, which uses the standard posterior weights over the full dataset; *soft posterior reweighting*, where a small labeled reference set induces soft class probabilities over the dataset that bias the weights toward a target class; and *hard filtering*, which restricts the distribution to a single class. Fig. 9 shows that soft posterior reweighting with as few as $M = 5$ labeled points already produces strong steering, approaching the hard-filter upper bound.

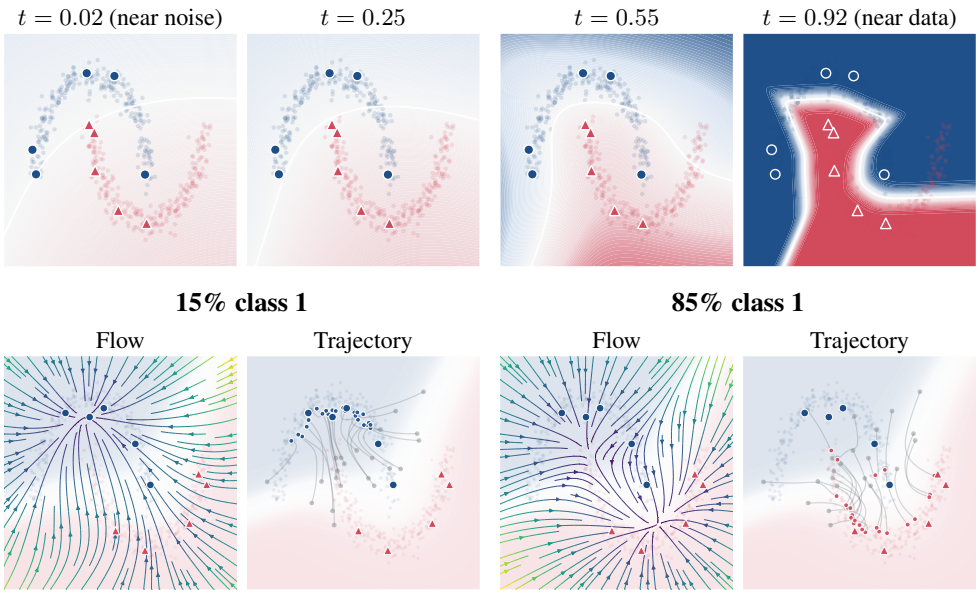


Figure 8: Two-moons control. **Top:** t changes with the reference set fixed. **Bottom:** the reference composition changes with the model fixed.

MNIST. We repeat the analysis on MNIST digits (0 and 1), where the reference set now operates on image-space representations. Fig. 10 shows that $M = 50$ soft-labeled references already produce reliable class steering, with steerability improving consistently as M grows. The same mechanism transfers from low-dimensional geometry to image space without modification.

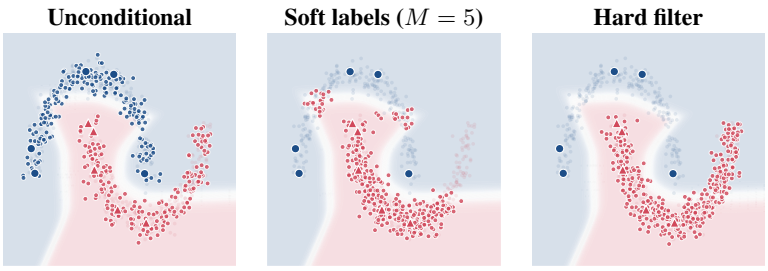


Figure 9: Inference-time condition changes; dataset and model are fixed.

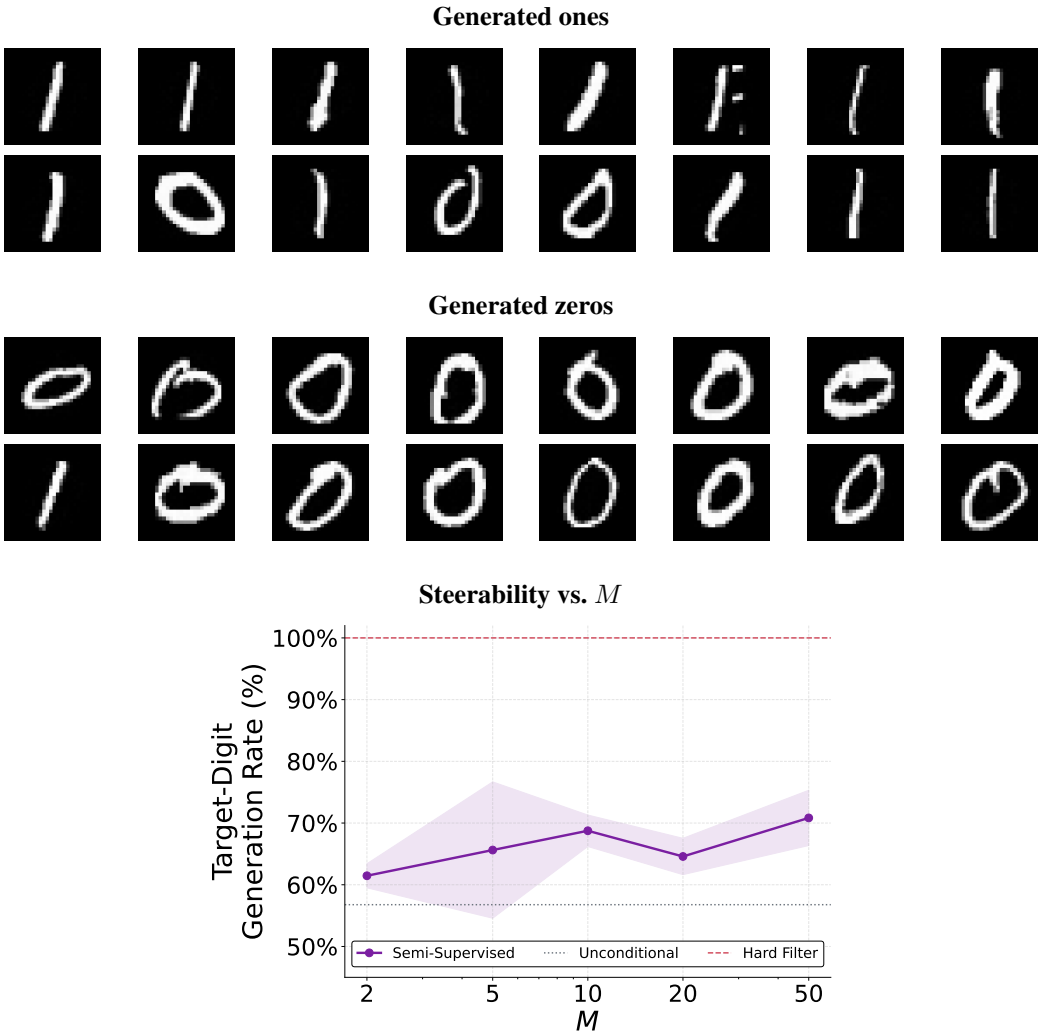


Figure 10: MNIST steering with soft-labeled references. The same model generates ones or zeros depending on the reference set, with model and noise fixed. Steerability as a function of reference-set size M shows that as few as $M = 50$ references already approach the hard-filter upper bound.

E Ablations

E.1 Guidance Schedule

The guidance correction in Eq. (9) is scaled by β_t, which controls both the strength and the timing of the intervention. We ablate two axes: the functional form of the schedule and the peak strength β_0. In all experiments in this appendix, we apply a late-time cutoff and set $\beta_t = 0$ for $t \geq 0.85$. This clipping prevents the correction from being applied in the final part of the trajectory, where the $(1-t)^{-1}$ factor can make high-strength schedules, especially the constant schedule, numerically unstable.

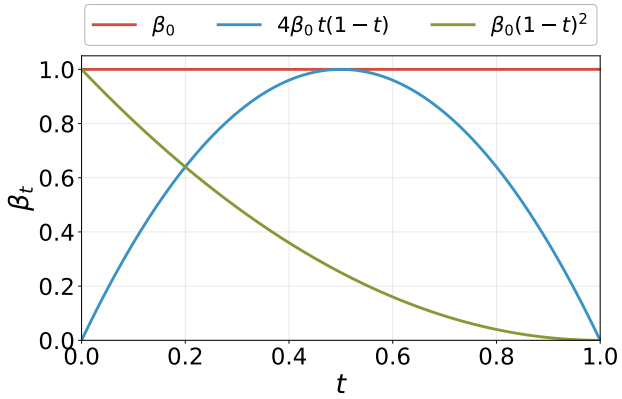


Figure 11: The three β_t schedules evaluated in this ablation, shown for $\beta_0 = 1$ before the shared late-time cutoff at $t = 0.85$: constant ($\beta_t = \beta_0$), quadratic decay ($\beta_t = \beta_0(1-t)^2$), and bell-shaped ($\beta_t = 4\beta_0 t(1-t)$). Constant applies uniform guidance until the cutoff; quadratic decay front-loads guidance and vanishes at $t = 1$, avoiding the $(1-t)^{-1}$ instability; bell-shaped guidance peaks at $t = 0.5$ and suppresses both early and late corrections.

All ablations in this appendix use the reference-set swap setting with the prompt *"an elephant in a jungle"* and the pink elephant reference set, with all other hyperparameters fixed to the values in Table 2.

E.2 Schedule Form

We compare three functional forms at fixed $\beta_0 = 0.5$.

- **Constant** ($\beta_t = \beta_0$): applies uniform guidance until the late-time cutoff. Without clipping, this schedule becomes unstable near $t = 1$ due to the $(1-t)^{-1}$ scaling of the correction, producing artifacts in the generated output.
- **Quadratic decay** ($\beta_t = \beta_0(1-t)^2$): front-loads the guidance signal and decays to zero as $t \to 1$, cancelling the divergence in the correction term. This is the schedule used in all main experiments.
- **Bell** ($\beta_t = \beta_0 \cdot 4t(1-t)$): suppresses guidance at both endpoints and concentrates the intervention around the midpoint of the trajectory. This avoids early-timestep interference when the overall structure is still being determined.

E.3 Guidance Strength β_0 and Schedule Shape

We fix the schedule and sweep $\beta_0 \in \{0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 1.0, 1.5, 2.0\}$ for each schedule form. Although the geometric-mixture derivation in Section 3.2 is stated for $\beta_t \in [0, 1]$, we relax this constraint in the ablation to treat β_0 as an extrapolated guidance strength. Values above one are therefore used as a stress test of stability and control strength, not as a normalized mixture coefficient. This sweep is separate from the task-specific settings in Table 2; the savanna bank-swap experiment uses $\beta_0 = 1.0$.

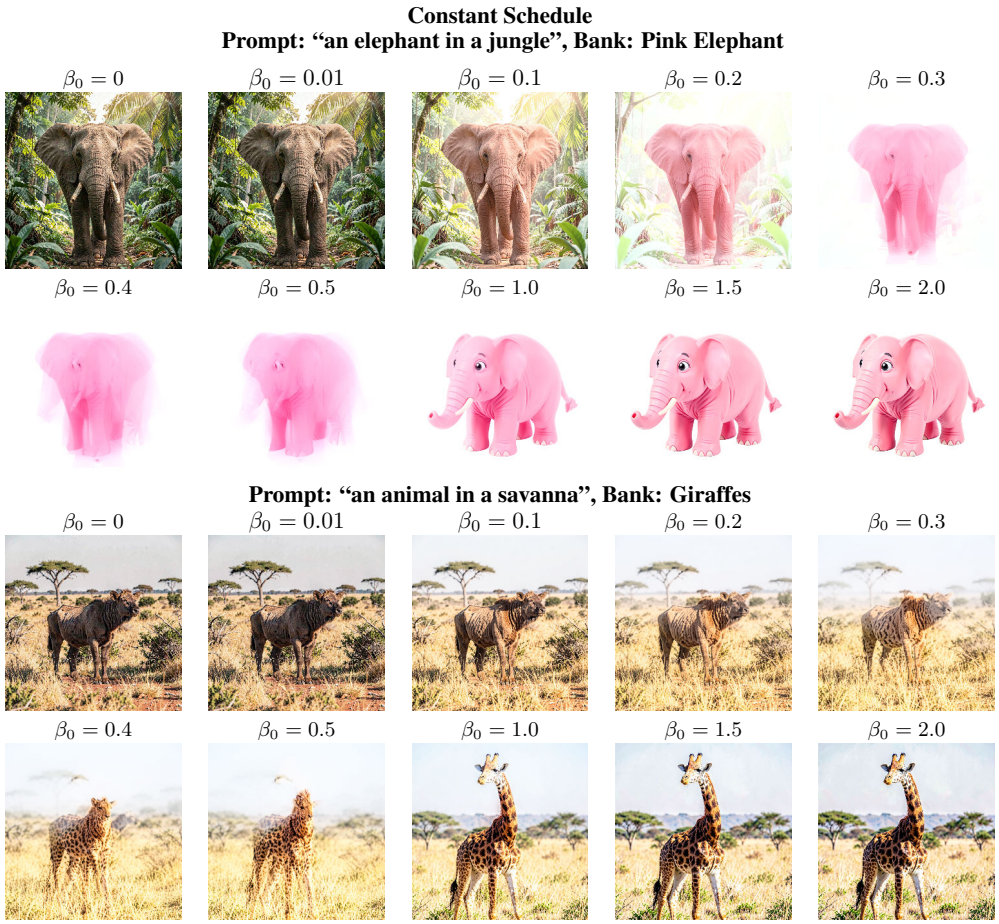


Figure 12: Ablation of guidance strength β_0 for the constant schedule. The constant schedule applies uniform guidance at every step. At moderate β_0 the target attribute transfers cleanly, but above $\beta_0 \approx 1$ artifacts appear near $t = 1$, visible as oversaturated colors and structural distortion.

E.4 Reference-Set Size

We study the effect of the reference-set size on the diversity of generated outputs. We fix the prompt, model, and guidance schedule, and vary only the number of reference examples used at inference time. All reference sets are constructed from a single attribute-aligned bank, with subsets sampled uniformly at random.

We evaluate reference-set sizes

$$M \in \{1, 2, 4, 8, 16, 32, 64, 128\}.$$

For each size, we sample multiple random subsets from the full reference bank and generate images using a fixed set of seeds.

We measure diversity using average pairwise LPIPS between generated samples. This metric captures perceptual variation and provides an estimate of the support of the generated distribution. Error bars indicate variability across randomly sampled reference subsets.

We observe that diversity increases consistently with the size of the reference set. Small reference sets produce more concentrated outputs, while larger sets yield a broader range of samples. This indicates that reference-mean-guided dynamics do not simply constrain generation toward a fixed target, but instead redistribute probability mass across the reference distribution.

Importantly, this increase in diversity occurs even when the reference set corresponds to a single semantic attribute, suggesting that the method captures intra-class variation rather than collapsing to

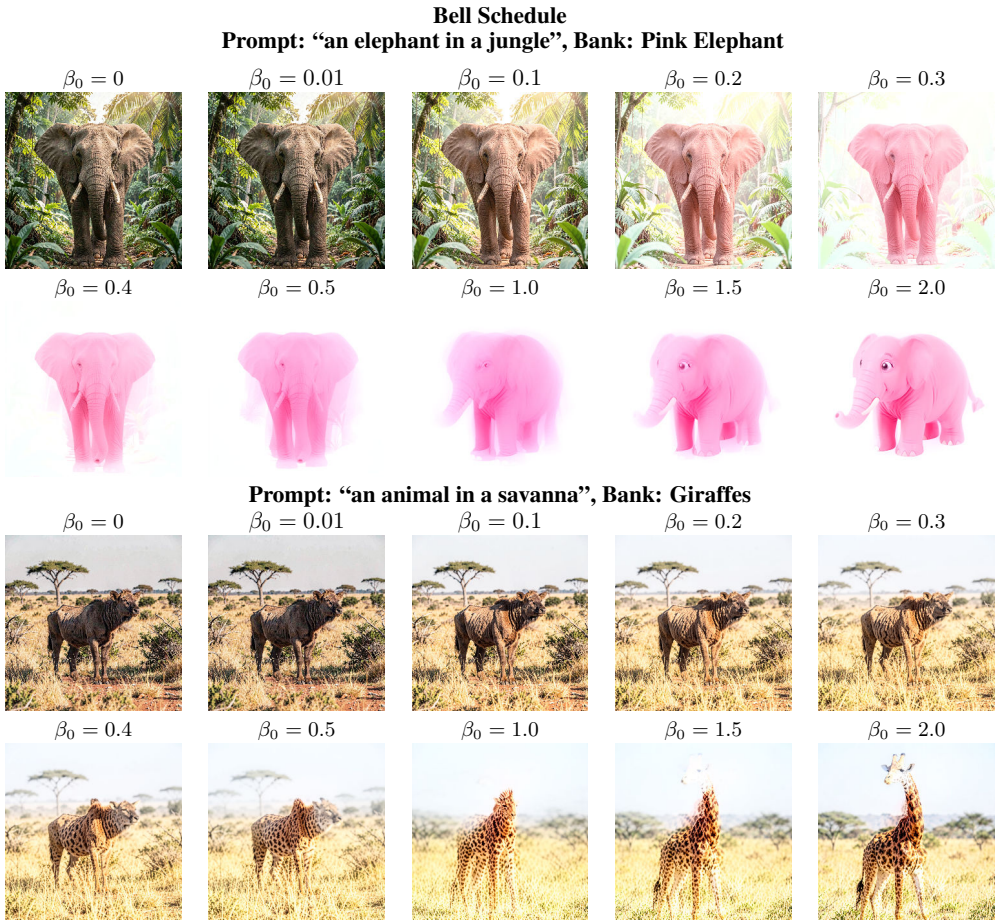


Figure 13: Ablation of guidance strength β_0 for the bell-shaped schedule. The bell-shaped schedule concentrates guidance around the midpoint of the trajectory and suppresses both early and late corrections. Relative to the constant schedule, it delays attribute transfer slightly but remains stable at larger β_0 values.

a single mode. This behavior contrasts with guidance methods that often reduce diversity as control strength increases.

While this experiment focuses on diversity, controllability is established separately in the main text.

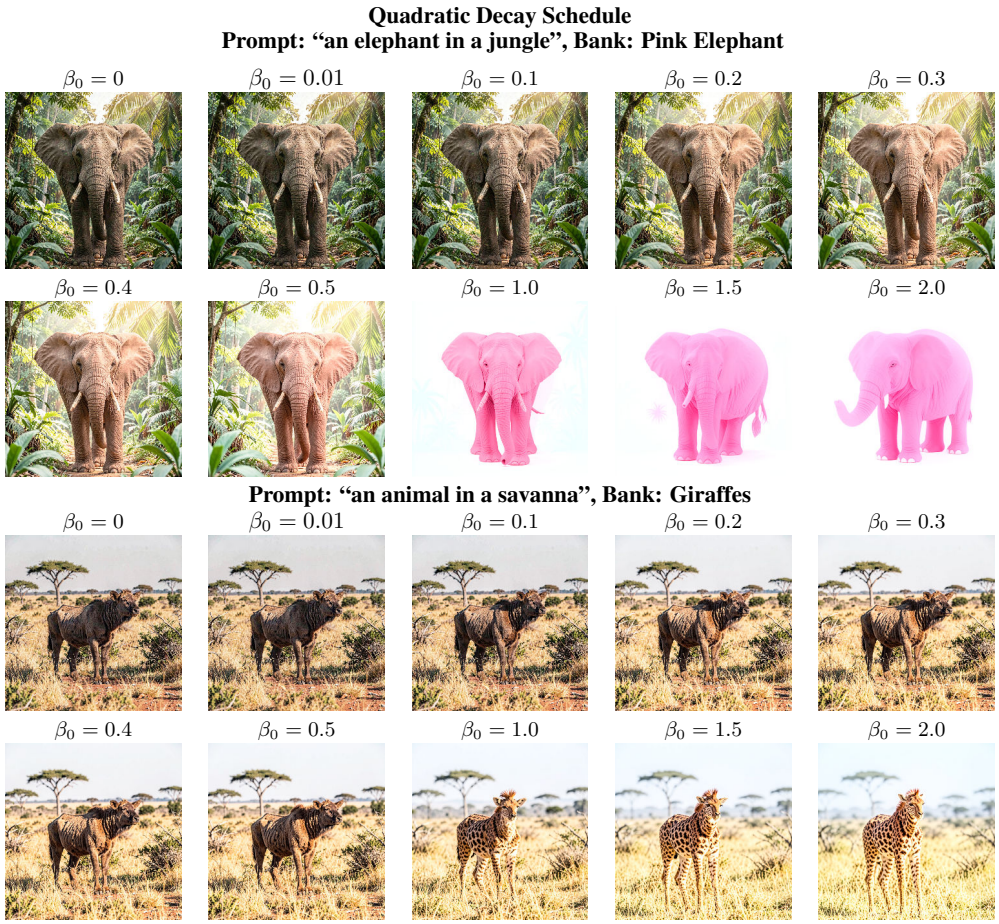


Figure 14: Ablation of guidance strength β_0 for the quadratic decay schedule. The quadratic schedule front-loads guidance and decays to zero at $t = 1$, cancelling the late-time divergence. Across the full β_0 range it provides the cleanest attribute transfer with the fewest late-time artifacts, which is why this schedule is used in the main experiments.

E.5 Number of Function Evaluations (NFE)

We study the effect of the number of function evaluations (NFE) on RMG control using a challenging ring-leap control task. We use the prompt

> “a gymnast performing a ring leap, full body visible, airborne, one leg extended forward, the back leg bent high behind the head, arched back, pointed toes, arms extended, dynamic sports photograph”

together with a fixed reference set of ring-leap images. All experiments use the quadratic decay guidance schedule described in Appendix E.1.

We sweep both the number of function evaluations and the guidance strength. Specifically, we evaluate

$$\text{NFE} \in \{10, 20, 30, 50, 100, 200\}$$

and

$$\beta_0 \in \{0.1, 0.2, 0.4, 0.5, 1.0\}.$$

All other hyperparameters, including prompt, reference set, and random seeds, are held fixed.

We also observe that runtime scales approximately linearly with NFE, highlighting a practical trade-off between control quality and computational cost.

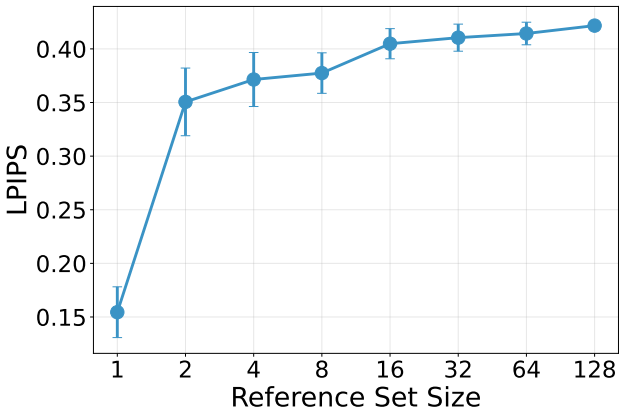


Figure 15: Reference-set size ablation. LPIPS diversity increases with the number of reference examples M. Error bars indicate variability across randomly sampled reference subsets.

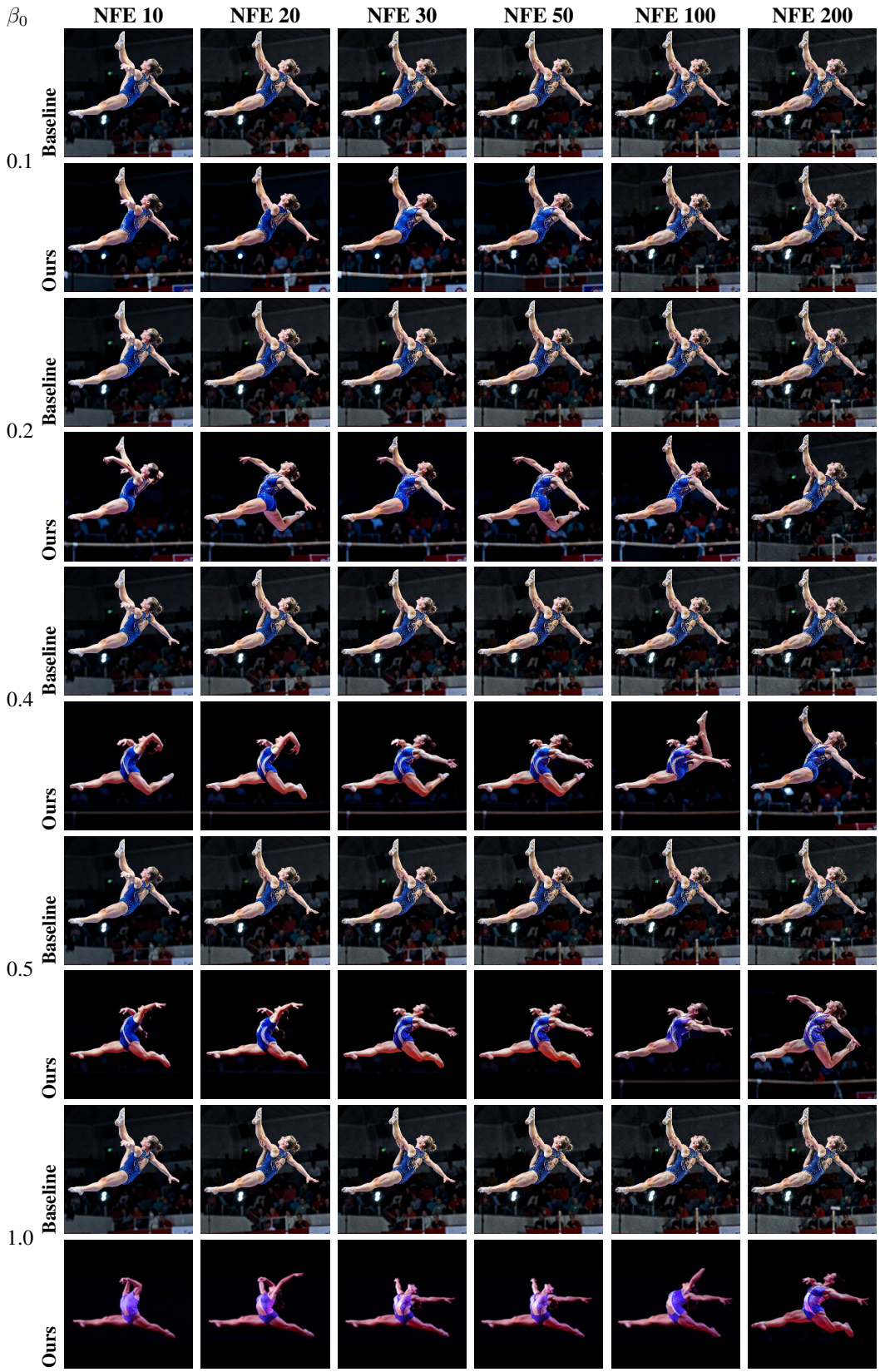


Figure 16: Ring-leap control task across guidance strengths and solver budgets. Columns vary NFE, rows vary β_0, and each β_0 is shown as a baseline/guided pair.

F Additional Experiments

F.1 Prompt–Reference Interaction

A natural question is whether RMG control simply replicates prompt engineering. We test this by varying prompt and reference set independently: prompts are neutral or attribute-specific, crossed with no reference set, a neutral elephant reference set, and a pink elephant reference set (Fig. 17).

The results show three regimes. With no reference set, the output follows the prompt. The neutral reference set suppresses the pink attribute even when the prompt specifies it (bottom centre). The pink reference set introduces the attribute against a neutral prompt and amplifies it when both agree. Prompt and reference set are independent, composable control axes.

To quantify this effect, we report a CLIP-based directional pinkness score for the guided samples. The score is defined as the similarity to the prompt *"a pink elephant in a jungle"* minus the similarity to *"a gray elephant in a jungle"*, so higher values indicate a stronger pink attribute. Full metric details are provided in Appendix C.5.

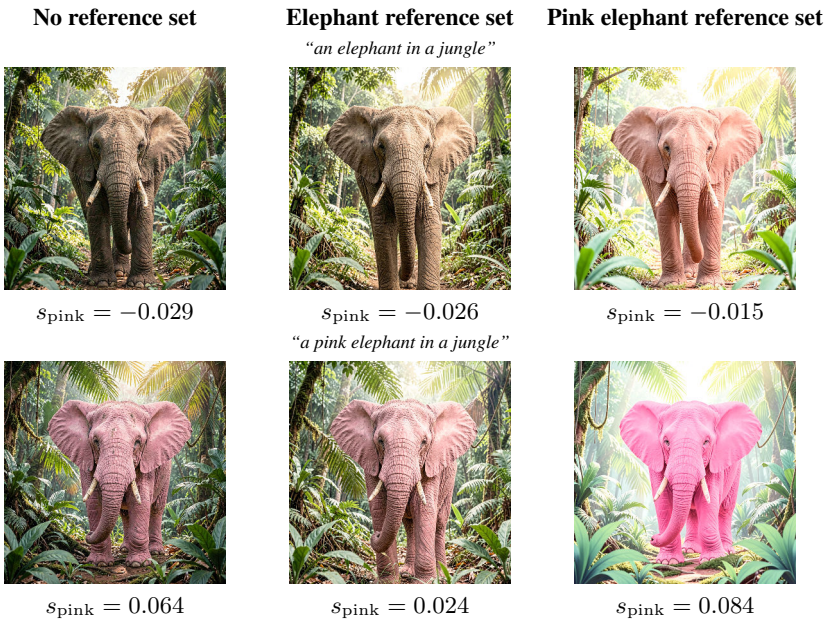


Figure 17: Prompt–reference interaction. Rows change the prompt; columns change the reference set; the noise seed is fixed.

F.2 Reference Composition

We provide additional evidence that reference-mean guidance enables continuous control through the composition of the reference distribution. In these experiments, the reference set is formed by mixing two banks that correspond to different attributes, while keeping the prompt and sampling procedure fixed. By varying the mixture proportion, we measure how the generated distribution changes in response.

Setup. For each experiment, we construct a reference set by combining two attribute-specific banks and varying the fraction of the target attribute in the bank from 0% to 100%. Unless otherwise stated, each bank contains 20 images, and the reference composition is varied over the set

$$\{0, 25, 50, 75, 100\}\%.$$

All hyperparameters used in these experiments are the same as those reported in Appendices C and C.3.

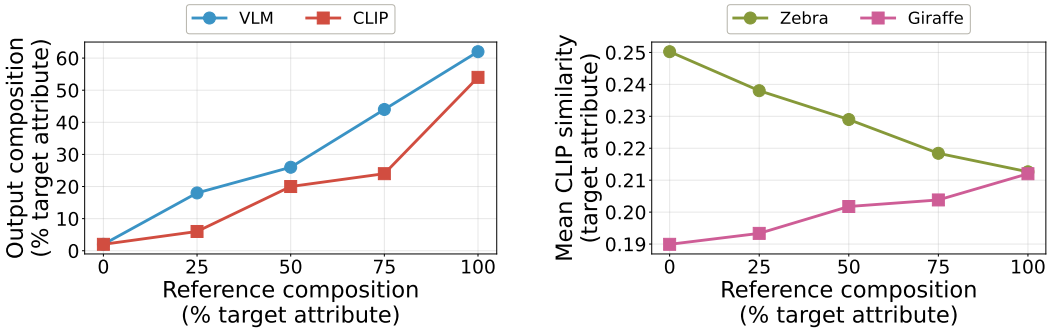


Figure 18: Quantitative controllability under reference composition for the prompt *"an animal in a savanna"*. **Left:** output composition measured as the fraction of generated images classified as giraffes by Qwen2-VL-7B or by CLIP text-image comparison. **Right:** mean CLIP similarity to the giraffe and zebra prompts. Increasing the proportion of giraffes in the reference distribution leads to a corresponding increase in giraffe frequency in the generated outputs, together with improved semantic alignment.

Evaluation metrics. We quantify controllability using both discrete and continuous semantic measures.

Attribute frequency. To estimate the fraction of generated images exhibiting the target attribute, we use the vision-language model `Qwen/Qwen2-VL-7B` [51] as a zero-shot attribute classifier. For each generated image, we ask a binary question tailored to the task, for example:

> "Which animal does the main animal in this image most resemble: zebra or giraffe? Answer with exactly one word: zebra or giraffe."

The model's answer is then mapped to one of the candidate attributes. This yields an estimate of the output composition as a function of the reference composition.

For the CLIP-labeled curve in the output-composition plot, we use an analogous zero-shot classifier based on text-image similarity: an image is assigned to the target attribute when its CLIP similarity to the target prompt exceeds its similarity to the distractor prompt. We then report the percentage of images assigned to the target class. This produces a discrete CLIP-derived composition estimate, distinct from the continuous CLIP similarity scores shown in the right plot.

CLIP similarity. We also compute CLIP similarity scores between generated images and text prompts corresponding to the target and distractor attributes. This provides a continuous measure of semantic alignment, complementing the discrete attribute-frequency estimate above.

Summary. The quantitative curves below show that increasing the proportion of a target attribute in the reference distribution produces a corresponding increase in its prevalence in the generated outputs. The qualitative grids further illustrate how the generated samples change across the same progression.

F.3 SPG Diversity as a Function of Reference-Set Size

We evaluate how the diversity of SPG samples changes with the number of reference examples available at inference time. For each reference-set size M, we generate samples with the same trained model and measure diversity using average pairwise LPIPS between generated images. This isolates whether increasing the reference set broadens the generated distribution rather than collapsing samples toward a small number of retrieved examples.

Fig. 21 shows that LPIPS increases with reference-set size, indicating that larger reference sets support more diverse generations while preserving the reference-conditioned control behavior reported in the main text.

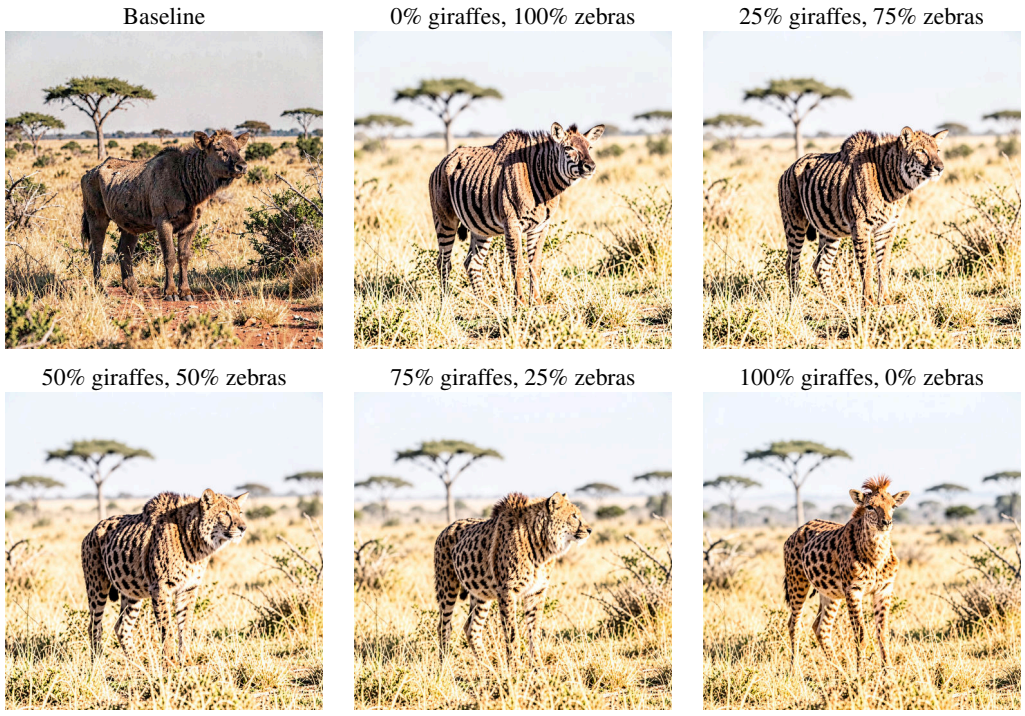


Figure 19: Qualitative controllability for the prompt *“an animal in a savanna”*. The reference distribution is constructed by mixing zebra and giraffe banks while holding the prompt fixed. As the proportion of giraffes in the reference set increases, the generated outputs shift correspondingly toward giraffe-like samples.

F.4 Suppressing Reference-Set Nuisance Artifacts

We use a controlled reference set in which all examples share a white background. This setting tests whether guidance transfers only useful semantic structure or also copies nuisance properties of the reference bank. Because RMG uses the closed-form reference mean directly, it transfers the shared white background along with the object appearance. SPG, by contrast, uses the reference mean as an anchor and refines it through a learned residual, which allows the model to preserve object-level guidance without reproducing the background artifact. Fig. 22 shows this comparison.

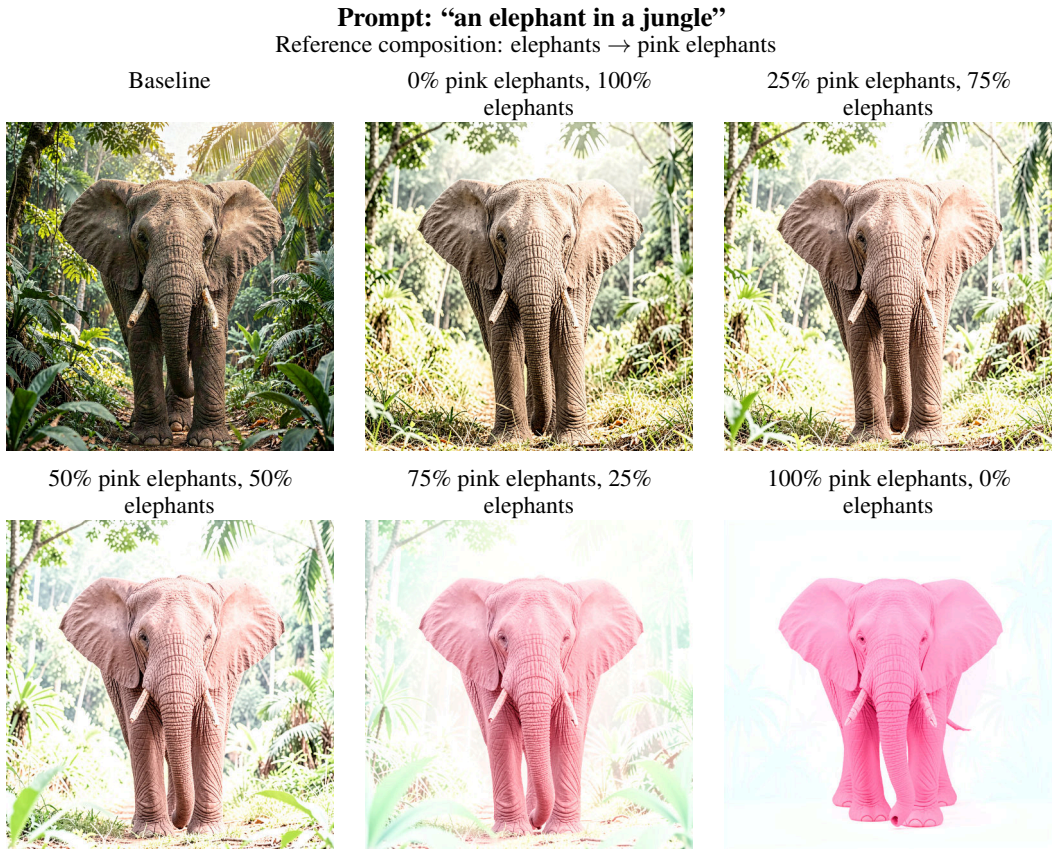


Figure 20: Qualitative controllability for the prompt *“an elephant in a jungle”*. The reference distribution is constructed by mixing elephant and pink-elephant banks while keeping the prompt fixed. Increasing the fraction of pink elephants in the reference set progressively shifts the outputs toward the target color attribute.

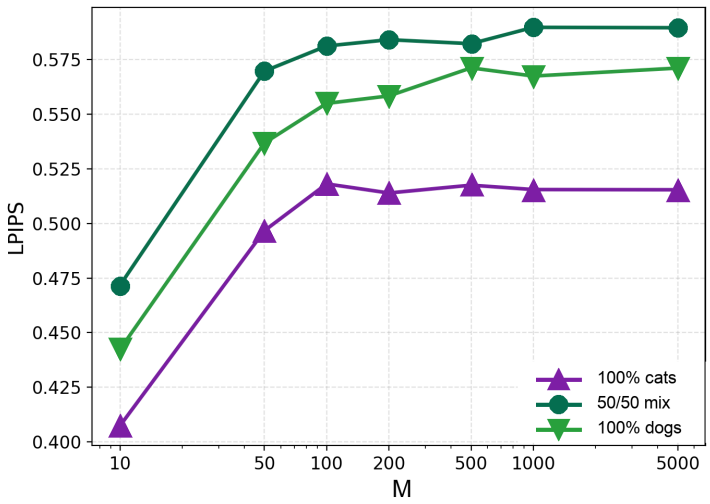


Figure 21: SPG diversity as a function of reference-set size. Average pairwise LPIPS increases with the number of reference examples M, indicating that larger reference sets broaden the generated distribution rather than inducing retrieval-like collapse.

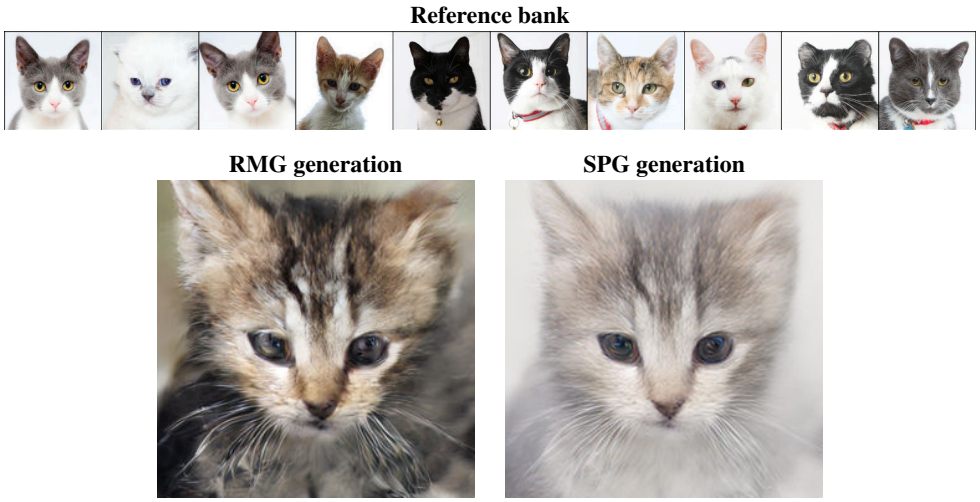


Figure 22: White-background reference-bank comparison. The reference bank consists of examples with a shared white background. RMG transfers this nuisance property to the generated sample, whereas SPG preserves the object-level guidance without copying the white background.

G Reference Banks



Figure 23: Reference bank of 20 images of pink elephants.



Figure 24: Reference bank of 20 images of blue elephants.



Figure 25: Reference bank of 20 images of giraffes.

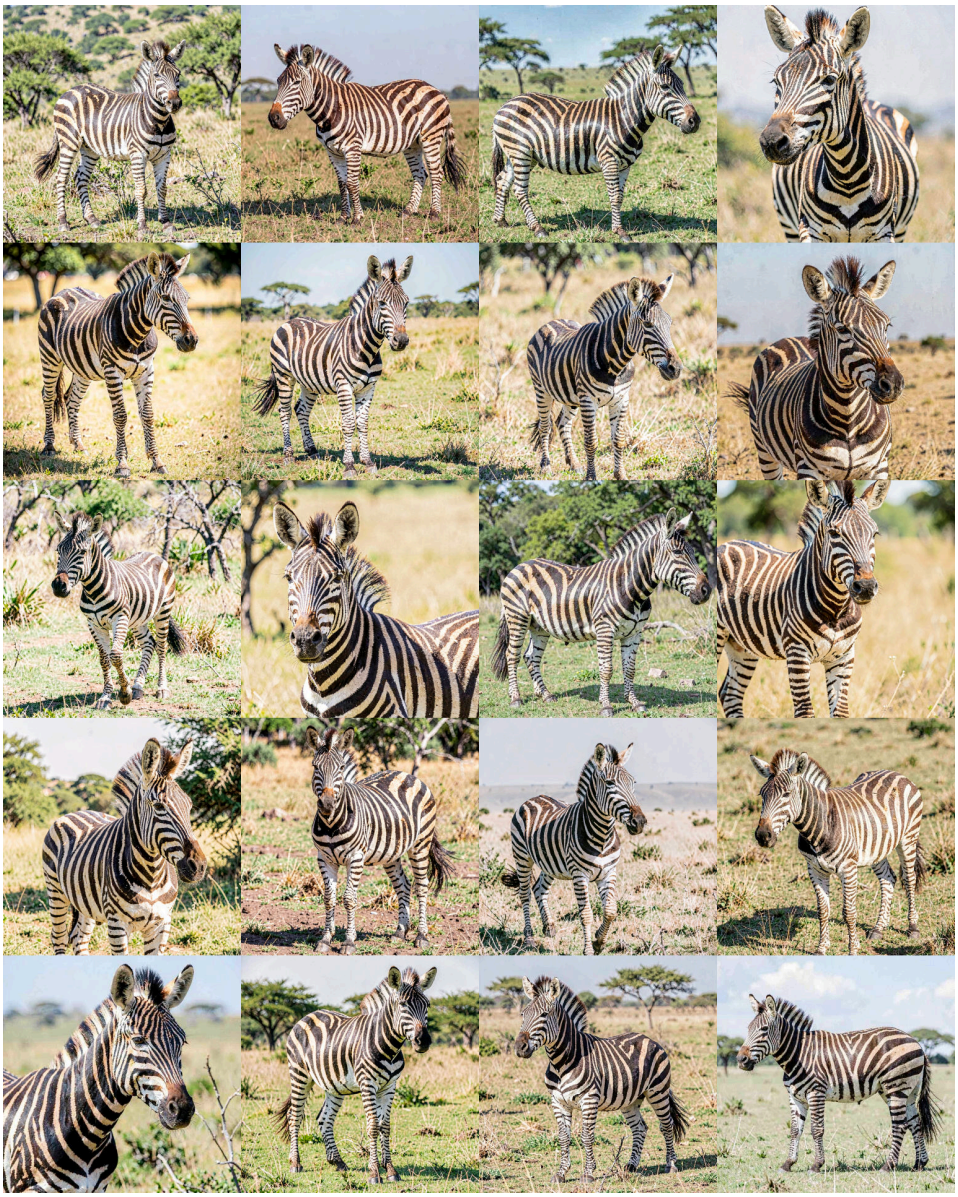


Figure 26: Reference bank of 20 images of zebras.



Figure 27: Reference bank of 20 images of elephants.



Figure 28: Reference bank of 20 images of keyholes.



Figure 29: Reference bank of 20 images of Van Gogh style images.



Figure 30: Reference bank of 20 pencil-sketch house images.



Figure 31: Reference bank of 20 cinematic house images.

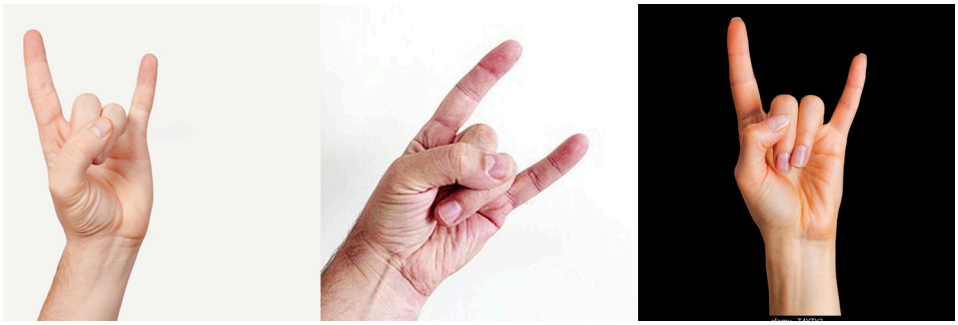


Figure 32: Reference bank of three hand-pose images used for the sign-of-the-horns experiment.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the proposed methods, RMG and SPG, and the claims are supported by the theory in Section 3 and empirical results in Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Section 6, including dependence on reference-set quality, scaling to large reference banks, modality-specific design choices, and misuse risks.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The assumptions are stated in Sections 2 and 3, and complete derivations are provided in Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Experimental setup, prompts, schedules, reference-bank construction, metrics, and ablations are described in Section 4 and appendices C, E and F.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Code is not included in the anonymous submission. The paper provides implementation details and hyperparameters needed to reproduce the methods, and uses cited public models and datasets where applicable.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Training details for SPG and sampling details for RMG, including schedules, cutoffs, softmax temperature, reference-set sizes, prompts, and evaluation metrics, are given in Appendices C, E and F.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Quantitative experiments report sample counts or variability where appropriate, including 10,000-sample SPG class-composition estimates and error bars in reference-set size ablations.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The appendix reports the main compute setup, including FLUX.2 inference on A100 hardware and SPG training on 4 A100 GPUs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work studies controllable generative modeling using existing models and datasets, and we are not aware of deviations from the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 6 discusses beneficial uses such as personalization and low-data adaptation, as well as risks from improved controllability of image generators.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a new pretrained image generator, dataset, or high-risk asset; it evaluates methods on existing models and curated reference sets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Existing datasets, models, and evaluation tools are cited in the main text and appendix, including FLUX.2, AFHQv2-related evaluation, GenEval, CLIP, Qwen-VL, and REPA-E.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce a new dataset, benchmark, or released model asset.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or human-subject studies.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human-subject research.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The core methods do not rely on LLMs, but the appendix discloses the use of Qwen2-VL-7B as a vision-language evaluator for attribute-frequency estimates.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.