P-GUIDE: PARAMETER-EFFICIENT PRIOR STEERING FOR SINGLE-PASS CFG INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Classifier-Free Guidance (CFG) is widely used for conditional generation in flow matching, yet its standard dual-pass form requires two backbone evaluations at each sampling step. We introduce **P-Guide**, a framework that applies guidance to the mean of a learned initial latent distribution and then uses one backbone evaluation per step. The resulting prior shift induces a local first-order trajectory response without explicit velocity-field extrapolation during sampling. We consider both homoscedastic and heteroscedastic priors; their comparable ImageNet performance indicates that learning the variance is secondary to mean steering in this setting. Experiments on class-conditional generation show that **P-Guide** reduces guided sampling GFLOPs by approximately 50% at the same step count while adding only a 1.31M-parameter prior module; on ImageNet, prior guidance reduces FID from 27.78 to 22.33 (U-Net) and from 32.84 to 24.61 (DiT-B/2) while remaining single-pass, offering a cost-effective balance between sampling efficiency and generation quality at moderate guidance scales.

1 INTRODUCTION

Continuous-time generative models, including Flow Matching (FM) (Lipman et al., 2023; Albergo & Vanden-Eijnden, 2023) and Rectified Flow (Liu et al., 2023), use deterministic Ordinary Differential Equations (ODEs) to transport simple source distributions into complex data distributions (Song et al., 2021b; Song & Ermon, 2019). Closely related to score-based diffusion models (Ho et al., 2020; Song et al., 2021a; Nichol & Dhariwal, 2021), these frameworks support scalable transformer-based generators (Peebles & Xie, 2023; Esser et al., 2024).

Semantic control is essential to modern conditional generators, including Stable Diffusion (Rombach et al., 2022), SDXL (Podell et al., 2023), Imagen (Saharia et al., 2022), Stable Diffusion 3 (Esser et al., 2024), and FLUX (Black Forest Labs, 2024). Classifier-Free Guidance (CFG) (Ho & Salimans, 2022) strengthens conditioning by extrapolating between conditional and unconditional predictions, including velocity fields in flow-based models (Dhariwal & Nichol, 2021; Esser et al., 2024). We examine the computational consequence of this mechanism in class-conditional generation.

Standard CFG evaluates conditional and unconditional branches at every sampling step, making guidance costly for large generators such as SDXL and SD3 (Podell et al., 2023; Esser et al., 2024). Distillation and few-step sampling reduce solver steps (Salimans & Ho, 2022; Meng et al., 2023; Song et al., 2023; Luo et al., 2023); we instead reduce guidance cost within each step through prior learning and backbone adaptation.

Learned-source methods show that the initial distribution can encode conditions, align source and target, or shorten transport (Lee et al., 2022; Scholz & Turner, 2025; Chen et al., 2025; Liu et al., 2025; Kim et al., 2026). These advances establish the source as a learnable component. Crucially, learned conditional and null means expose a source-space contrast analogous to the contrast CFG repeatedly forms from velocity predictions. P-Guide turns it into an inference-time control before integration.

P-Guide therefore moves guidance from the sampling trajectory to its origin (Figure 1). It fits and freezes a lightweight prior, then adapts the backbone to the learned source. At inference, it extrapolates once from the null to the conditional mean, then follows the conditional ODE with **one backbone evaluation per step**.

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

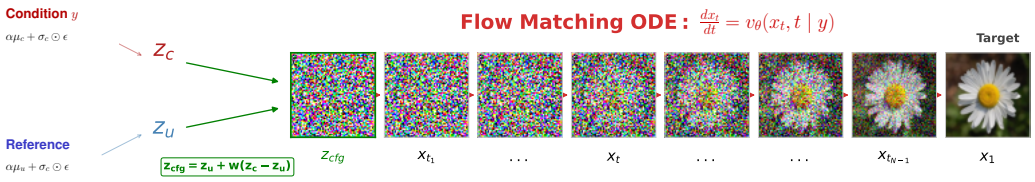


Figure 1: **Conceptual overview of the P-Guide framework.** Standard dual-pass CFG evaluates both v_{cond} and v_{uncond} at each integration step. P-Guide steers the mean of a learned prior and then follows the conditional ODE with **one backbone evaluation per step**. Here $\mu_c = \mu_\phi(y)$, $\mu_u = \mu_\phi(\emptyset)$, and $\sigma_c = \sigma_\phi(y)$; both seeds use the same Stage-2 mean coefficient α, share ϵ, and retain the conditional scale.

We study both homoscedastic and heteroscedastic priors. In either case, guidance acts only on the mean, while the variance determines source spread. Learning the variance improves some low-resolution results, but the two variants perform comparably on ImageNet. At equal step counts, P-Guide approximately halves guided sampling GFLOPs; matched-backbone-call results and scale sweeps locate its strongest operating point at moderate guidance.

Our Contributions. Our contributions are threefold:

- **Trajectory correspondence theory:** We derive the source-shift response with an explicit quadratic remainder, a necessary-and-sufficient condition for pointwise first-order agreement with CFG, and a finite-scale bound separating alignment and nonlinear errors.
- **Single-pass prior guidance:** P-Guide applies the conditional–null extrapolation pattern of CFG to learned source means through a 1.31M-parameter prior module, reducing guided sampling cost by approximately 50% at equal step counts.
- **Source parameterization study:** We compare homoscedastic and heteroscedastic priors and identify mean steering as the primary inference-time control, while keeping the conditional variance fixed during guidance.

2 Related Work

Diffusion and Flow-Based Generative Models. Diffusion models iteratively estimate scores to reverse noising (Ho et al., 2020; Song et al., 2021a; Nichol & Dhariwal, 2021; Song et al., 2021b), whereas Flow Matching regresses a velocity field that transports a source to data (Lipman et al., 2023). Rectified Flow promotes straighter trajectories to reduce ODE discretization error (Liu et al., 2023).

Learned Sources and Efficient Guidance. PriorGrad and Warm-Start Diffusion learn conditional initial statistics to simplify denoising or shorten transport (Lee et al., 2022; Scholz & Turner, 2025); CAR-Flow aligns source and target, CrossFlow maps cross-modal sources, and CSFM regularizes conditional flow-matching sources (Chen et al., 2025; Liu et al., 2025; Kim et al., 2026). These methods establish learned sources as initialization and transport mechanisms. P-Guide develops the source into an adjustable guidance interface by extrapolating paired conditional and null means while preserving the conditional scale.

Guidance distillation amortizes guided predictions into a student (Meng et al., 2023), while NoiseRefine learns an initial correction from guided-output supervision with a frozen generator (Ahn et al., 2026). SGFM instead modifies the source under a frozen vector field with recovery guarantees (Wang et al., 2025), and Follow the Mean shifts the conditional endpoint mean (Curvo et al., 2026). P-Guide learns conditional and null means from the data likelihood, shares their difference across noise samples, and adapts the backbone to this source. Its sampling scale controls CFG-like source-mean extrapolation without guided-output supervision or a frozen downstream field.

Analysis and Control of Guidance Dynamics. Recent work analyzes or modifies guidance within the sampling trajectory. Inner-CFG (Sun et al., 2024) uses structured continuous conditions, CFG-Zero* (Fan et al., 2025) changes early-stage scaling, CFG++ and Rectified-CFG++ (Chung et al.,

2025; Saini et al., 2025) impose trajectory constraints, and CFG-Ctrl and C^{2}FG (Wang et al., 2026; Gao et al., 2026) cast guidance as a control problem. P-Guide instead places the control at the source and studies its propagation through an otherwise fixed conditional trajectory.

3 BACKGROUND

3.1 FLOW MATCHING

Flow Matching (FM) provides a simulation-free framework to train continuous normalizing flows by regressing a time-dependent velocity field $v_\theta(x_t, t)$ (Lipman et al., 2023). Given a source distribution p_0 (typically standard Gaussian) and a target data distribution p_1, FM constructs a probability path p_t that transports p_0 to p_1, and is closely related to diffusion models and score-based generative modeling (Ho et al., 2020; Song et al., 2021b;a).

The sample trajectory x_t is governed by the following ODE:

$$\frac{dx_t}{dt} = v_\theta(x_t, t), \quad x_0 \sim p_0. \tag{1}$$

For a pair of samples (x_0, x_1), a commonly used construction is the independent conditional probability path, which uses linear interpolation: $x_t = (1-t)x_0 + tx_1$ (Lipman et al., 2023). The corresponding target velocity is constant: $u_t(x_t|x_0, x_1) = x_1 - x_0$. The model is trained using the Flow Matching objective:

$$\mathcal{L}_{FM}(\theta) = \mathbb{E}_{t,x_0,x_1}[\|v_\theta(x_t, t) - (x_1 - x_0)\|^2]. \tag{2}$$

3.2 CLASSIFIER-FREE GUIDANCE (CFG)

Classifier-Free Guidance (CFG) is the dominant technique for controllable generative modeling in diffusion and flow-based systems (Ho & Salimans, 2022).

During training, the condition is randomly replaced with a null token $\emptyset$, allowing a single model to learn both conditional velocity $v(x_t, t|y)$ and unconditional velocity $v(x_t, t|\emptyset)$. At inference time, the guided velocity field is computed as:

$$v_{cfg}(x_t, t) = v_\theta(x_t, t|\emptyset) + w(v_\theta(x_t, t|y) - v_\theta(x_t, t|\emptyset)), \tag{3}$$

where $w \geq 1$ is the guidance scale (Ho & Salimans, 2022). At $w = 1$, the unconditional term cancels and sampling uses the conditional prediction alone. For $w > 1$, standard CFG evaluates both branches, approximately doubling backbone computation at the same step count.

4 METHOD

P-Guide starts from the conditional–null contrast underlying CFG. The learned source means provide this direction at $t = 0$: P-Guide extrapolates it once and then follows the conditional flow. We next define the source, two-stage training, and sampling rule.

4.1 PRIOR SPACE PARAMETERIZATION

Standard generative flows transport an uninformed Gaussian source $p_0(z) = \mathcal{N}(0, \mathbf{I})$ to the data distribution p_1 (Ho et al., 2020; Lipman et al., 2023). A learned, data-dependent source can instead encode conditional structure before transport begins (Lee et al., 2022; Liu et al., 2025; Chen et al., 2025; Scholz & Turner, 2025). We therefore parameterize $p_0(z|y)$ and assign distinct roles to two scales. The coefficient $\alpha \in [0, 1]$ sets the learned-mean magnitude for a model configuration and is shared by Stage-2 training and inference; w controls inference-time movement along the conditional–null mean direction. Decreasing α contracts both means and their separation without changing the conditional variance or sampled noise.

Homoscedastic P-Guide (Base). The base model lets the condition y modulate only the source mean and keeps unit variance:

$$z^{(\alpha)} = \alpha\mu_\phi(y) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}). \tag{4}$$

At a realizable population optimum (Appendix B), the mean prediction satisfies $\mu_\phi(y) = \mathbb{E}[x \mid y]$. We define $z_c^{(\alpha)} = \alpha\mu_\phi(y)+\epsilon$ and $z_{\text{ref}}^{(\alpha)} = \alpha\mu_\phi(\emptyset) + \epsilon$, whose difference is $\alpha[\mu_\phi(y) - \mu_\phi(\emptyset)]$.

Heteroscedastic P-Guide (Extension). The extension also lets y control the scale of the latent initialization. With source parameters $(\mu_\phi(y), \sigma_\phi(y))$, we define

$$z^{(\alpha)} = \alpha\mu_\phi(y) + \sigma_\phi(y) \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}). \tag{5}$$

The initial state therefore encodes class-dependent mean and scale statistics before transport begins. The paired reference $z_{\text{ref}}^{(\alpha)} = \alpha\mu_\phi(\emptyset) + \sigma_\phi(y) \odot \epsilon$ retains the conditional standard deviation, so $z_c^{(\alpha)} - z_{\text{ref}}^{(\alpha)} = \alpha[\mu_\phi(y) - \mu_\phi(\emptyset)]$ isolates a pure mean displacement. Both states share ϵ, so their difference cancels sampling noise and the scale term. Guidance therefore follows a deterministic mean direction while preserving the conditional covariance $\sigma_\phi^2(y)$.

Algorithm 1 P-Guide: Training with mean scale α; the fixed-σ variant sets $\sigma = 1$

```
# Stage 1: train F_phi
x1, y = sample_data()
y = drop_to_null(y)
mu, sigma = F_phi(y)
e, s2 = x1 -mu, sigma**2
loss_prior = mean(e**2 / (2*s2)
              +0.5*log(s2))
update(F_phi, loss_prior)
freeze(F_phi) # Stage 2
alpha = configured_mean_scale
x1, y = sample_data()
eps = randn_like(x1)
mu, sigma = F_phi(y)
z = alpha *mu +sigma *eps
t = sample_uniform()
x_t = (1 -t) *z +t *x1
v = v_theta(x_t,t,y)
loss = mean((v -(x1 -z))**2) # MSE
update(v_theta, loss)
```

Algorithm 2 P-Guide: Sampling

```
# mean-only prior guidance
# alpha=0 uses the separate
    standard-CFG sampler
alpha = trained_mean_scale # same
    value as Stage 2
eps = randn(x_shape)
mu_c, sigma_c = F_phi(y)
x = alpha *mu_c +sigma_c *eps
if w != 1:
   mu_u, _ = F_phi(None)
   x = x +alpha*(w-1)*(mu_c -mu_u)
# one backbone call per solver
    step
for t, dt in schedule:
   x = x +v_theta(x,t,y) *dt
return x
```

4.2 Sequential Two-Stage Training

Training separates source estimation from transport learning. We first fit the prior module and freeze it; the backbone then learns transport from samples of that source (Kendall & Gal, 2017; Lipman et al., 2023).

Stage 1: Prior Learning. Stage 1 minimizes the conditional Gaussian negative log-likelihood (NLL) of target data x_1 given y, fitting the conditional mean and coordinate-wise variance (Kendall & Gal, 2017):

$$\mathcal{L}_{prior} = \mathbb{E}_{x_1,y}\left[\sum_i \left(\frac{(x_{1,i} - \mu_{\phi,i}(y))^2}{2\sigma_{\phi,i}^2(y)} + \frac{1}{2}\log\sigma_{\phi,i}^2(y)\right)\right]. \tag{6}$$

For the fixed-variance option, setting $\sigma_\phi(y) = 1$ reduces this objective to mean squared error up to a constant factor. Learning the variance instead weights each squared residual by its inverse (Kendall & Gal, 2017); the effect on generation quality is evaluated through the prior comparisons.

These outputs serve different roles in the downstream flow: the mean locates each class, whereas the variance describes its remaining source spread. Condition dropout produces the null mean $\mu_\phi(\emptyset)$ within the same prior model, making the conditional–null difference a shared steering direction across noise draws.

Stage 2: Flow Matching. After Stage 1 converges, we freeze the prior and train or fine-tune v_θ from $z^{(\alpha)} = \alpha\mu_\phi(y) + \sigma_\phi(y) \odot \epsilon$. The linear conditional path $x_t = (1-t)z^{(\alpha)} + tx_1$ yields

$$\mathcal{L}_{FM}^{(\alpha)}(\theta) = \mathbb{E}_{t,(x_1,y),z^{(\alpha)}\sim p_0^{(\alpha)}(z|y)}\left[\|v_\theta(x_t, t \mid y) - (x_1 - z^{(\alpha)})\|^2\right]. \tag{7}$$

Stage 2 fixes α and learns the source-to-data path induced by $z^{(\alpha)} \sim p_0^{(\alpha)}(z \mid y)$. Reusing this α at inference keeps the training and sampling sources aligned.

Each pair $(z^{(\alpha)}, x_1)$ determines both x_t and the regression target $x_1 - z^{(\alpha)}$. Changing α therefore changes the transport task, not just the sampler input. Each reported α defines a separate Stage-2 configuration with aligned training and inference sources. P-Guide then applies one conditional–null mean displacement and follows the fixed conditional velocity field, whereas CFG extrapolates that field at every step.

4.3 Trajectory-Level Approximation

For a model trained with mean scale α, let $\Delta\mu = \mu_\phi(y) - \mu_\phi(\emptyset)$ and $\Delta z_\alpha = \alpha\Delta\mu$. Guidance moves the trained conditional source by $h_{\alpha,w} = (w-1)\Delta z_\alpha = \alpha(w-1)\Delta\mu$ before ODE integration.

Theorem 1 (Mean-scale control of local error)**.** *Let $z_{cfg}^{(\alpha)} = z_c^{(\alpha)} + h_{\alpha,w}$ and $J_{t,c}^{(\alpha)} = D_z\Phi_t^y(z_c^{(\alpha)})$. If the conditional flow map Φ_t^y is C^2 and its Hessian norm is bounded by M_t along this displacement, then*

$$\Phi_t^y(z_{cfg}^{(\alpha)}) = \Phi_t^y(z_c^{(\alpha)}) + J_{t,c}^{(\alpha)} h_{\alpha,w} + R_t, \qquad \|R_t\| \le \frac{M_t}{2}\alpha^2(w-1)^2\|\Delta\mu\|^2. \tag{8}$$

The effective perturbation is $\alpha(w-1)\Delta\mu$. For comparable local curvature, halving α quarters the error bound, so a larger guidance scale can remain within the same local regime. Appendix B relates this result to standard CFG: it establishes the necessary-and-sufficient condition for pointwise first-order endpoint agreement and a finite-scale bound that separates alignment error from nonlinear response.

4.4 Prior Steering with One Backbone Evaluation per Step

P-Guide extrapolates the prior mean before ODE integration begins. Theorem 1 describes how this initial displacement propagates through the conditional flow. Sampling then follows that flow with one backbone evaluation per step.

The P-Guide Core Formula: Using the same α as Stage-2 training, we guide the scaled prior mean and retain the conditional standard deviation:

$$z_{cfg}^{(\alpha)} = \alpha[\mu_\phi(\emptyset) + w(\mu_\phi(y) - \mu_\phi(\emptyset))] + \sigma_\phi(y) \odot \epsilon. \tag{9}$$

The primary model uses $\alpha = 1$. Each mean-scale ablation uses its stated α in both Stage-2 training and sampling. At $w > 1$, Eq. (9) adds $(w-1)\Delta z_\alpha$ to the trained conditional source. The noise term is unchanged, so α and w act only on the mean. At $\alpha = 0$, $\Delta z_0 = 0$ and P-Guide follows its unsteered conditional path. The sampler draws $z_c^{(\alpha)}$, adds $(w-1)\Delta z_\alpha$, and then integrates the conditional ODE using only $v_\theta(\cdot, \cdot \mid y)$. Standard CFG instead combines conditional and null backbone predictions at every step. P-Guide places guidance at the trajectory origin and retains one backbone evaluation per step.

4.5 Toy Experiment: Visualizing Trajectory Steering

We conduct a 2D toy experiment that illustrates how an initial mean shift changes the trajectories of a fixed conditional flow.

Experimental Setup. The target distribution π_1 consists of two clusters at radius 5.0 (angles 0 and 4 radians, $\sigma = 0.5$). We use $\alpha = 1$, fit the prior module $F_\phi(y)$ to the cluster means, and train a Rectified Flow model (Liu et al., 2023) for $v_\theta(x_t, t|y)$.

Trajectory Anchoring through Prior Steering. P-Guide instead injects guidance at $t = 0$ by shifting the initial state to $z_{cfg} = z_{\text{ref}} + w(z_c - z_{\text{ref}})$, with one model evaluation per step. Figure 2 shows the resulting deterministic paths under the same adapted conditional flow.

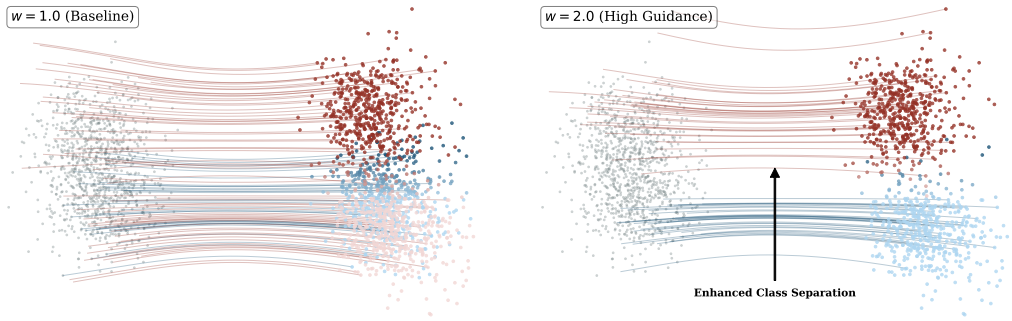


Figure 2: **Visualizing trajectory steering in a 2D toy setup.** Standard CFG mixes two velocity predictions per step. P-Guide applies the mean shift $z_c - z_{\text{ref}}$ at $t = 0$. These runs use $\alpha = 1$ in training and sampling. Increasing w from 1.0 to 2.0 increases class separation while retaining **one backbone evaluation per step**.

5 Experiments

Our experiments ask whether source-mean steering retains conditional control with one backbone evaluation per step, and how that constraint changes the quality–compute trade-off on MNIST (Lecun et al., 1998), CIFAR-10 (Krizhevsky & Hinton, 2009), and ImageNet-1k (Deng et al., 2009).

5.1 Experimental Setup

Hardware and Implementation. Experiments use one **NVIDIA GeForce RTX 4090 GPU** (24GB), the Rectified Flow objective (Liu et al., 2023), and the linear path $x_t = (1 - t)z + tx_1$. ImageNet-1k uses a frozen pretrained **AutoKL** (Esser et al., 2024); the prior is fitted for 40K steps, frozen, and followed by 400K-step backbone adaptation. Unless stated otherwise, all P-Guide results in the main text use $\alpha = 1$ in both Stage-2 training and sampling; Appendix C.1 is the explicit mean-scale exception. Sampling GFLOPs include backbone and prior inference but exclude AutoKL decoding.

Prior-Module Footprint. The PG module supplies class-conditioned source parameters using approximately **1.31 million parameters**. Stage 1 fits this module; Stage 2 adapts the backbone to the learned source, keeping its architecture unchanged. Appendix A describes the module and training setup; Appendix C.4 evaluates its capacity.

Metrics. We report **Fréchet Inception Distance (FID)** (Heusel et al., 2017), **Kernel Inception Distance (KID)** (Bińkowski et al., 2018), and **Inception Score (IS)** (Salimans et al., 2016) to quantify sample quality. The 50-step main-table quality metrics are 50K-sample point estimates; KID is reported without rescaling. Efficiency is measured via hardware-agnostic **GFLOPs** per generated image (50 sampling steps for the main tables), and we report **parameter counts in millions (M)**.

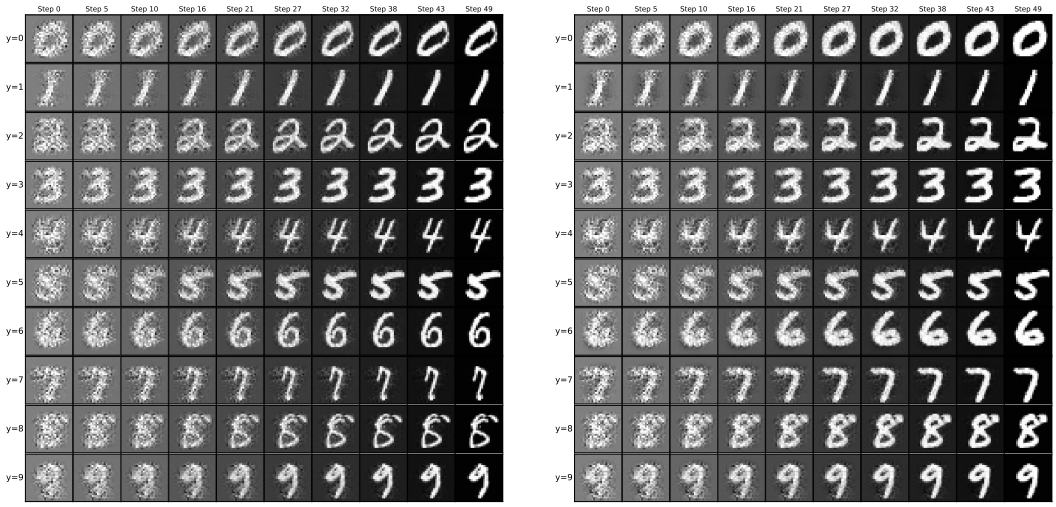
We also measure **generation accuracy** to assess label consistency under different guidance strengths. MNIST uses a fixed pretrained CNN with 99.1% accuracy (An et al., 2020), and CIFAR-10 uses a fixed pretrained ResNet-56 with 92.0% accuracy (He et al., 2016).

Additional Evidence in the Appendix. The supplementary experiments probe the zero and near-zero mean-scale regimes (Appendix C.1), source-scale behavior, prior capacity, and initialization (Appendices C.2–C.5). They also test distributed steering (Appendix D.1), interaction with standard CFG (Appendix D.2), and small-scale text-conditioned compatibility (Appendix E). These studies separate evidence for mean steering from effects tied to a particular prior or training configuration.

5.2 Results on MNIST and CIFAR-10

For CIFAR-10 we use a standard **U-Net** (Ronneberger et al., 2015; Ho et al., 2020); for MNIST we use **LightningDiT** (Yao et al., 2025). We compare single-pass P-Guide with standard dual-pass CFG.

Both CFM and P-Guide backbones receive 400K training steps under the same backbone settings, while P-Guide additionally fits its prior module.



(a) Trajectory from z_{cfg} to x_1 with $w = 1.0$

(b) Trajectory from z_{cfg} to x_1 with $w = 1.5$

Figure 3: **Visual comparison of P-Guide generation trajectories on MNIST.** Each row shows the evolution from prior ($t = 0$) to data ($t = 1$). The examples compare $w = 1.0$ and $w = 1.5$ under the same random seed, illustrating how the initial mean shift affects digit structure along the trajectory.

Quality and Computation. Table 1 separates three effects of guidance: sample quality, label consistency, and computation. For $w > 1$, standard CFM invokes dual-pass CFG and doubles its backbone cost, whereas P-Guide remains single-pass with nearly constant sampling cost. The quality metrics need not move together: on MNIST, learned-variance P-Guide reaches 99.96% accuracy at $w = 1.5$, but its FID rises to 13.80.

Learning the variance improves FID at $w = 1$ from 9.35 to 6.51 on CIFAR-10 and from 3.31 to 1.40 on MNIST. The advantage depends on the setting: at $w = 1.1$ on CIFAR-10, learned variance gives FID 10.48 versus 8.72 for fixed variance, despite higher classification accuracy. At $w = 1$, P-Guide samples its conditional prior without additional mean extrapolation, so the best low-resolution FIDs reflect the learned source and its trained backbone.

The trade-off differs by dataset. On CIFAR-10, P-Guide uses approximately half the guided sampling cost and reaches FID 6.51, whereas standard CFG reaches 2.54. On MNIST, the learned P-Guide source already gives its lowest FID at $w = 1$. Stronger prior steering can improve label accuracy while degrading FID, making the useful scale range metric- and dataset-dependent. Appendix D.2 examines prior steering combined with standard CFG.

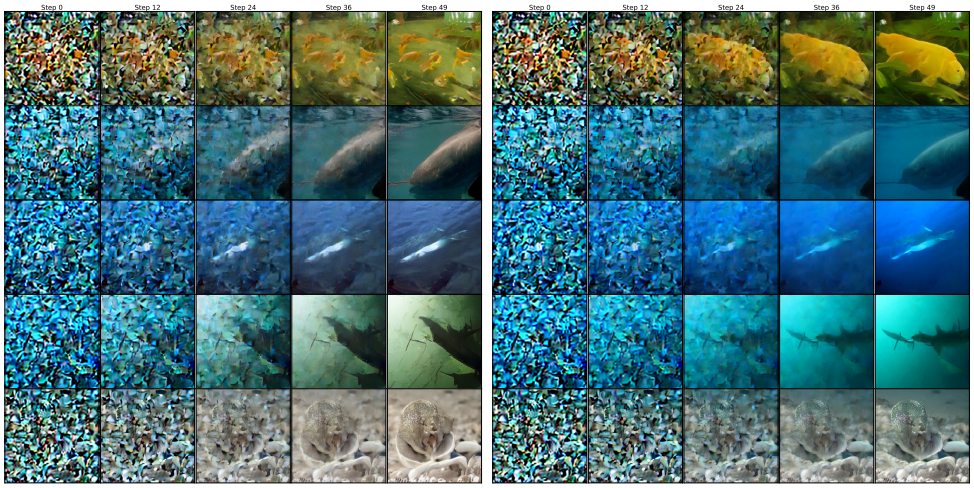
5.3 Scalability on ImageNet-1K

To assess scalability at 256×256, we evaluate U-Net (Ronneberger et al., 2015; Ho et al., 2020) and DiT-B/2 (Peebles & Xie, 2023). ImageNet models inherit pretrained CFM backbones (Dao et al., 2023) and then receive 400K adaptation steps. The ImageNet results therefore measure adaptation from pretrained initialization. Under the same 400K Stage-2 budget, the DiT-B/2 comparison gives FID 33.68 from pretrained initialization versus 49.03 from scratch (Appendix C.5). P-Guide additionally fits its prior for 40K steps. Figure 4 illustrates the resulting trajectories.

Architecture Comparison. Table 2 shows that P-Guide maintains one backbone evaluation per step across both architectures, with a small additional sampling cost. The default prior has **1.31 million parameters**; enlarging it to 13.1 million yields no FID improvement in the existing capacity comparison (Appendix C.4). The two variance formulations have comparable ImageNet FIDs, with the lower point estimate depending on the backbone: 22.33 versus 22.65 for U-Net and 25.07 versus

Table 1: CIFAR-10 and MNIST results from separately trained source–backbone pairs. Pass denotes backbone evaluations per step: P-Guide uses one, and CFM uses standard dual-pass CFG for $w > 1$. Here w is method-specific: a prior-steering scale for P-Guide and a velocity-guidance scale for CFM. GFLOPs are rounded to two decimals. Acc denotes classification accuracy (%). Shading marks the best P-Guide point estimate per metric (lowest for ↓, highest for ↑).

			CIFAR-10					MNIST				
Method	Pass	w	FID ↓	KID ↓	IS ↑	Acc (%) ↑	GFLOPs ↓	FID ↓	KID ↓	IS ↑	Acc (%) ↑	GFLOPs ↓
CFM	1	1.0	9.44	0.0063	8.61	70.53	710.43	5.87	0.0061	2.10	95.47	417.60
CFM	2	1.5	4.00	0.0017	9.52	86.07	1420.86	5.94	0.0061	2.10	99.61	835.20
CFM	2	2.0	2.54	0.0004	10.10	92.34	1420.86	6.12	0.0066	2.09	99.94	835.20
CFM	2	3.0	3.72	0.0009	10.42	96.63	1420.86	11.50	0.0120	2.06	99.98	835.20
PG($\sigma_\phi(y)$)	1	0.9	9.22	0.0049	8.82	80.78	713.38	2.39	0.0015	2.11	97.27	417.61
PG($\sigma_\phi(y)$)	1	1.0	**6.51**	**0.0026**	**9.07**	83.99	711.91	**1.40**	**0.0007**	2.11	98.97	417.60
PG($\sigma_\phi(y)$)	1	1.1	10.48	0.0041	9.03	**86.00**	713.38	2.18	0.0011	2.09	99.52	417.61
PG($\sigma_\phi(y)$)	1	1.5	38.46	0.0164	7.65	76.75	713.38	13.80	0.0098	2.02	**99.96**	417.61
PG($\sigma = 1$)	1	0.9	12.59	0.0081	8.32	68.53	713.38	3.21	0.0029	**2.13**	97.48	417.61
PG($\sigma = 1$)	1	1.0	9.35	0.0057	8.61	72.49	711.91	3.31	0.0030	**2.13**	98.22	417.60
PG($\sigma = 1$)	1	1.1	8.72	0.0044	8.67	76.69	713.38	3.61	0.0032	2.12	98.61	417.61
PG($\sigma = 1$)	1	1.5	18.23	0.0060	8.69	81.12	713.38	6.73	0.0055	2.08	99.51	417.61



(a) Trajectory from z_{cfg} to x_1 with $w = 1.0$ (b) Trajectory from z_{cfg} to x_1 with $w = 1.1$

Figure 4: **Visual comparison of P-Guide trajectories on ImageNet.** Each row shows the evolution from prior ($t = 0$) to data ($t = 1$). Increasing the guidance scale ($w = 1.0 \to 1.1$) sharpens semantic structure from early steps, illustrating trajectory-level control from the origin with one backbone evaluation per step.

24.61 for DiT-B/2. Thus, mean extrapolation is the principal inference-time control in the evaluated design; the conditional scale sets source spread but is not extrapolated by the main method.

Effect of Mean Steering. P-Guide controls conditional generation through the prior-steering scale w (Wang et al., 2026). Table 2 shows that moderate w improves both prior parameterizations, whereas both degrade at $w = 1.5$, marking a limited range for one-shot mean steering. Theorem 1 characterizes the local nonlinear response; Appendix D.1 evaluates distributed shifts, and Appendix C.1 isolates α: smaller values shrink the displacement, while $\alpha = 0$ reports a separate standard-CFG reference.

This pattern ties the useful range to the magnitude and timing of the source displacement, rather than to variance extrapolation. The mean-scale sweep changes $\alpha(w - 1)\Delta\mu$, while distributed steering injects smaller shifts along the trajectory. Their contrasting behavior helps explain why a moderate one-shot shift helps but a larger shift can leave the regime learned during adaptation.

Table 2: ImageNet-1k (256×256), 50 steps. P-Guide is single-pass and mean-only; standard CFG is reported for both backbones. The method-specific w controls prior means for P-Guide and velocities for CFG. Params (M): backbone + PG, excluding AutoKL (83.65 M). GFLOPs include prior inference and are rounded to two decimals.

	U-Net						DiT-B/2					
	P-Guide				Standard CFG		**P-Guide**				Standard CFG	
Scale w	FID ↓		GFLOPs ↓	Params	FID ↓	GFLOPs ↓	FID ↓		GFLOPs ↓	Params	FID ↓	GFLOPs ↓
	$\sigma = 1$	$\sigma_\phi(y)$					$\sigma = 1$	$\sigma_\phi(y)$				
1.0	28.40	27.78	3654.40		28.72	3652.92	32.84	33.68	2301.60		35.04	2300.12
1.1	22.65	**22.33**	3655.87	407.42	21.36	7305.84	26.41	26.80	2303.08	130.50	27.92	4600.24
1.2	22.67	23.01	3655.87	+1.31	15.80	7305.84	**24.61**	25.07	2303.08	+1.31	22.05	4600.24
1.5	43.38	44.53	3655.87		**8.53**	7305.84	38.72	38.43	2303.08		**11.03**	4600.24

Comparison with Standard CFG. At 50 solver steps, Table 2 shows that guided P-Guide cuts GFLOPs by approximately half, while standard CFG reaches the lower FID of 11.03 at $w = 1.5$. Table 3 instead fixes ten DiT-B/2 backbone evaluations, allocating them to five dual-pass CFG steps or ten single-pass P-Guide steps. Because backbone calls dominate sampling cost, this design approximately matches total FLOPs apart from P-Guide's small prior-inference overhead. Reallocating saved unconditional evaluations to conditional solver steps yields best tested FIDs of 26.90 for P-Guide and 30.01 for CFG.

Table 3: ImageNet with DiT-B/2 matched on ten backbone evaluations, approximately aligning total FLOPs because backbone computation dominates. CFG uses five dual-pass steps; P-Guide uses ten single-pass steps plus prior inference.

Method	Steps	Backbone evaluations	$w = 1.1$	$w = 1.2$	$w = 1.5$	$w = 2.0$
Standard CFG	5	10	55.76	47.82	**30.01**	32.83
P-Guide	10	10	30.21	**26.90**	34.89	90.49

Interpreting the Two Compute Regimes. These protocols answer complementary questions. At 50 steps, standard CFG spends twice as many backbone evaluations and attains the best FID, while P-Guide quantifies the savings from removing the unconditional branch. At ten backbone calls, P-Guide reallocates those calls to twice as many conditional solver updates and achieves the lower FID. The matched-call result therefore reflects both single-pass guidance and the additional integration steps it permits; it should not be read as an equal-step quality advantage. Because the scale sweep degrades under stronger extrapolation, the operating point depends jointly on backbone-call budget, solver discretization, and guidance scale.

Limitations. Our main evaluation covers class-conditional generation after source fitting and backbone adaptation. The sampling savings therefore apply when the adapted model is reused across generated samples and do not remove the preceding adaptation cost. One-shot mean steering remains effective over a limited scale range, while standard CFG attains better 50-step quality with two backbone evaluations per step. Appendix E tests text-to-image compatibility at small scale; single-pass text guidance at scale remains unevaluated.

6 Conclusion

P-Guide turns the learned source into a guidance interface by applying conditional–null mean extrapolation before transport. It fits source statistics, adapts the backbone, preserves conditional scale at inference, and uses one backbone call per step. This approximately halves guided GFLOPs at equal step counts and gives the lower FID point estimate at ten matched calls; standard CFG retains the better 50-step FID. Source-mean steering is therefore a compute-efficient operating point for reusable adapted models.

AI USE STATEMENT

In this work, we used large language models (LLMs) solely for language editing to improve the clarity, grammar, and academic style of author-written text. Research conception, methodological development, experimental design, analysis and interpretation of results, and formulation of scientific conclusions were carried out independently by the authors without assistance from generative AI tools. We reviewed all AI-assisted edits to ensure that they preserved the intended meaning and accurately reflected our work. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses public image benchmarks (MNIST, CIFAR-10, and ImageNet-1k) and involves no human subjects, personal data, or new data collection. The method inherits the bias and misuse risks of its generative backbones, including the production of biased or fabricated imagery; its intervention is confined to the initial latent distribution. The released code and pretrained components follow the licenses of the underlying datasets and backbones, and the authors declare no conflicts of interest.

REPRODUCIBILITY STATEMENT

Section 5.1 and Appendix A describe the hardware, training and evaluation settings, and anonymous implementation. Appendix B gives the theoretical assumptions and proofs. Appendices C, D, and F contain ablations, extensions, and qualitative results.

REFERENCES

Donghoon Ahn, Jiwon Kang, Sanghyun Lee, Jaewon Min, Minjae Kim, Wooseok Jang, Hyoungwon Cho, Sayak Paul, SeonHwa Kim, Eunju Cha, Kyong Jin, and Seungryong Kim. A noise is worth diffusion guidance. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/985786d06c1e45e9e8c65f7aca3547e4-Abstract-Conference.html.

Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *International Conference on Learning Representations (ICLR)*, 2023.

Sanghyeon An, Minjun Lee, Sanglee Park, Heerin Yang, and Jungmin So. An ensemble of simple convolutional neural network models for mnist digit recognition, 2020. URL https://arxiv.org/abs/2008.10400.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://arxiv.org/abs/1801.01401.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Chen Chen, Pengsheng Guo, Liangchen Song, Jiasen Lu, Rui Qian, Tsu-Jui Fu, Xinze Wang, Wei Liu, Yinfei Yang, and Alex Schwing. CAR-flow: Condition-aware reparameterization aligns source and target for better flow matching. In *Neural Information Processing Systems (NeurIPS)*, 2025. URL https://openreview.net/forum?id=idnW3BiZcV.

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. *arXiv preprint arXiv:2310.18235*, 2023. URL https://arxiv.org/abs/2310.18235.

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

Pedro M. P. Curvo, Maksim Zhdanov, Floor Eijkelboom, and Jan-Willem van de Meent. Follow the mean: Reference-guided flow matching. *arXiv preprint arXiv:2605.10302*, 2026.

Quan Dao, Hao Phung, Binh Nguyen, and Anh Tran. Flow matching in latent space. *arXiv*, 2023. URL https://arxiv.org/abs/2307.08698.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Neural Information Processing Systems (NeurIPS)*, 34, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, 2024.

Weichen Fan, Amber Yijia Zheng, Raymond A. Yeh, and Ziwei Liu. Cfg-zero*: Improved classifier-free guidance for flow matching models. 2025.

Jiayang Gao, Tianyi Zheng, Jiayang Zou, Fengxiang Yang, Shice Liu, Luyao Fan, Zheyu Zhang, Hao Zhang, Jinwei Chen, Peng-Tao Jiang, Bo Li, and Jia Wang. C^2fg: Control classifier-free guidance via score discrepancy analysis. 2026. URL https://arxiv.org/abs/2603.08155.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *arXiv preprint arXiv:2310.11513*, 2023. URL https://arxiv.org/abs/2310.11513.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Neural Information Processing Systems (NeurIPS)*, NIPS'17, pp. 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Neural Information Processing Systems (NeurIPS)*, 2020.

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. *arXiv preprint arXiv:2106.09685*, 2021. URL https://arxiv.org/abs/2106.09685.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *arXiv preprint arXiv:2303.11897*, 2023. URL https://arxiv.org/abs/2303.11897.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *arXiv preprint arXiv:2307.06350*, 2023. URL https://arxiv.org/abs/2307.06350.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Neural Information Processing Systems (NeurIPS)*, pp. 5574–5584, 2017.

Junwan Kim, Jiho Park, Seonghu Jeon, and Seungryong Kim. Better source, better flow: Learning condition-dependent source distribution for flow matching, 2026. URL https://arxiv.org/abs/2602.05951.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 12268–12290, 2024. doi: 10.18653/v1/2024.acl-long.663. URL https://aclanthology.org/2024.acl-long.663/.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Sang-gil Lee, Heeseung Kim, Chaehun Shin, Xu Tan, Chang Liu, Qi Meng, Tao Qin, Wei Chen, Sungroh Yoon, and Tie-Yan Liu. Priorgrad: Improving conditional denoising diffusion models with data-dependent adaptive prior. In *International Conference on Learning Representations (ICLR)*, 2022.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024. URL https://arxiv.org/abs/2406.13743.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900, 2022. URL https://arxiv.org/abs/2201.12086.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://arxiv.org/abs/1405.0312.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. *arXiv preprint arXiv:2404.01291*, 2024. URL https://arxiv.org/abs/2404.01291.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Qihao Liu, Xi Yin, Alan Yuille, Andrew Brown, and Mannat Singh. Flowing from words to pixels: A noise-free framework for cross-modality evolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2755–2765, 2025.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference, 2023.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik P. Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14297–14306, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Meng_On_Distillation_of_Guided_Diffusion_Models_CVPR_2023_paper.html.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning (ICML)*. PMLR, 2021.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 36479–36494, 2022.

Shreshth Saini, Shashank Gupta, and Alan Bovik. Rectified CFG++ for flow based models. In *Neural Information Processing Systems (NeurIPS)*, 2025. URL https://openreview.net/forum?id=NosdT1FHPv.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations (ICLR)*, 2022.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Neural Information Processing Systems (NeurIPS)*, NIPS'16, pp. 2234–2242, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Jonas Scholz and Richard E. Turner. Warm starts accelerate conditional diffusion. 2025. URL https://arxiv.org/abs/2507.09212.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021a.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Neural Information Processing Systems (NeurIPS)*, 2019.

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 (ICLR)*, 2021b.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning (ICML)*, 2023.

Shikun Sun, Longhui Wei, Zhicai Wang, Zixuan Wang, Junliang Xing, Jia Jia, and Qi Tian. Inner classifier-free guidance and its taylor expansion for diffusion models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/1269a9f5a4839c0fb80ed80abd973b84-Abstract-Conference.html.

Hanyang Wang, Yiyang Liu, Jiawei Chi, Fangfu Liu, Ran Xue, and Yueqi Duan. Cfg-ctrl: Control-based classifier-free diffusion guidance. 2026. URL https://arxiv.org/abs/2603.03281.

Zifan Wang, Alice Harting, Matthieu Barreau, Michael M. Zavlanos, and Karl H. Johansson. Source-guided flow matching. *arXiv preprint arXiv:2508.14807*, 2025.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024. URL https://arxiv.org/abs/2410.10629.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15703–15712, 2025.

A Implementation and Evaluation Details

A.1 Prior Module

The default prior module is a compact conditional ResNet with FiLM modulation (Figure 5). It maps the class or null condition to source parameters, with approximately 1.31 million parameters in the ImageNet configuration. The fixed-variance version predicts the mean; the learned-variance version additionally predicts the coordinate-wise source scale.

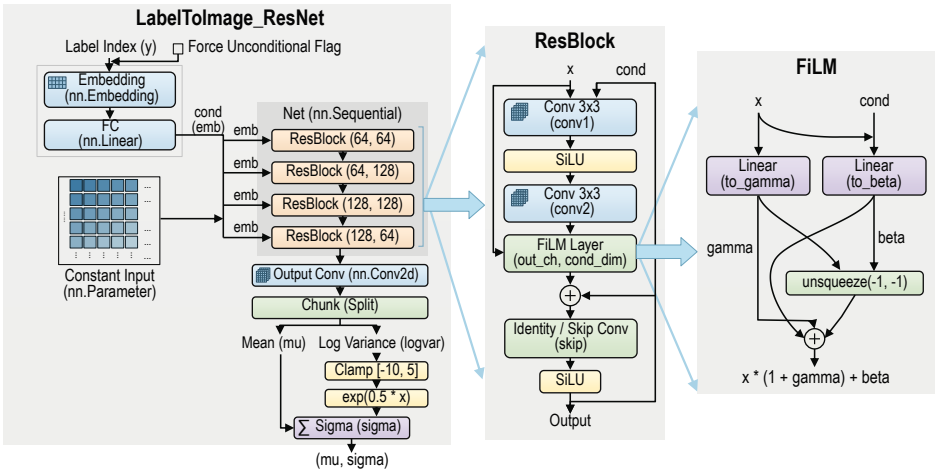


Figure 5: Architecture of the default P-Guide prior steering module (approximately 1.31 million parameters for ImageNet). The model consists of a lightweight conditional ResNet with FiLM modulation.

The ImageNet module uses a 256-dimensional embedding for each of the 1000 classes and the null condition, followed by a 256×256 linear layer. A learned $64 \times 32 \times 32$ tensor is processed by four FiLM-modulated residual blocks with output widths $64, 128, 128, 64$. A final convolution predicts mean and log-variance maps for the four latent channels. The log variance ℓ is clamped to $[-10, 5]$, and the standard deviation is $\exp(\ell/2)$.

Shared parameterization of source statistics. Stage 1 estimates class-dependent moments through shared convolutional parameters and class embeddings. A direct table of diagonal Gaussian moments would store $2(K+1)d$ scalar values for K classes and a d-dimensional source, including the null condition. For ImageNet, $K = 1000$ and $d = 4 \times 32 \times 32$, giving approximately 8.20 million values. The default network uses approximately 1.31 million parameters, about 16% of this count. The network couples moment estimates across classes through shared weights, whereas a direct moment table estimates each class separately. The population optimum in Appendix B characterizes realizable moment fitting under either parameterization.

A.2 Training and Sampling Budgets

Algorithm 1 specifies the Gaussian-NLL prior fitting and subsequent flow training. Condition dropout supplies the class-conditioned and null-conditioned prior parameters.

For MNIST and CIFAR-10, the standard Gaussian-source CFM and P-Guide backbones are each trained for 400K steps under the same backbone training settings. For ImageNet-1k, both the Gaussian-source baseline and P-Guide continue training pretrained CFM backbones for 400K steps. P-Guide first fits its prior module for 40K steps. The ImageNet experiments operate in the AutoKL latent space; AutoKL has approximately 83.65 million parameters. The main flow-matching experiments use one NVIDIA GeForce RTX 4090 GPU with 24 GB of memory.

The 50-step comparisons in Tables 1 and 2 use 50K generated samples. At $w = 1$, both methods use one conditional backbone evaluation per step. At $w > 1$, standard CFG uses two evaluations per step, while P-Guide evaluates its prior module at initialization and follows the conditional flow with one

evaluation per step. Thus, equal solver steps and equal backbone evaluations correspond to different comparisons: Table 2 fixes 50 steps, whereas Table 3 fixes ten DiT-B/2 backbone evaluations and allocates them to five CFG steps or ten P-Guide steps.

The ten-evaluation comparison measures quality at a fixed number of backbone calls, with additional prior-module evaluations for P-Guide. Standard CFG's FID decreases across this sweep up to its largest tested scale, $w = 1.5$.

Computation and parameter accounting. Sampling GFLOPs count the backbone and prior-module forward computations, excluding prior fitting, backbone training, and AutoKL decoding. In the ImageNet DiT-B/2 profiling run, PyTorch's `FlopCounterMode` counts an Euler trajectory with 50 updates and batch size ten; the total is divided by ten and 10^9 to obtain GFLOPs per image. At $w = 1$, P-Guide evaluates the conditional prior only. Nontrivial mean guidance also evaluates the null-condition prior once, producing the small difference in sampling GFLOPs. Parameter counts in millions (M) are computed as $\sum_{p \in \text{parameters}} \text{numel}(p)/10^6$. Table 2 separates backbone and prior-module counts and excludes AutoKL. Appendix C.4 compares module capacities.

A.3 Image Quality Evaluation

For CIFAR-10 and MNIST, FID, KID, and IS are computed with `torch-fidelity`. FID and KID use the 2048-dimensional features of `inception-v3-compat`; IS uses its classification outputs. The evaluation converts image tensors to unsigned 8-bit values. MNIST grayscale images are repeated across three channels for the Inception network. The CIFAR-10 reference comprises 50K training images. KID is the mean kernel distance returned by the evaluator, reported on its original scale. Classification accuracy uses the separate fixed classifiers described in Section 5.1.

A.4 Anonymous Implementation

The anonymous P-Guide implementation is available at:

`https://anonymous.4open.science/r/P-Guide-3DC6`

The anonymous release includes the prior-steering module and the ImageNet-1k training pipeline. Algorithms 1 and 2 specify the two-stage training procedure and mean-only sampling rule.

B Analysis of Prior Steering

This section derives the local trajectory response to prior-space steering and relates it to velocity-space guidance. We use the following notation:

- $z \in \mathbb{R}^d$ denotes the initial latent/prior state,
- $\Phi_t^y(z)$ denotes the flow map of the conditional ODE at time t,
- $V_t(z) = v_\theta(\Phi_t^y(z), t \mid y)$ denotes the velocity along the trajectory initialized at z, with condition y fixed,
- $\epsilon \sim \mathcal{N}(0, I)$ denotes the shared base noise used to couple the conditional and reference seeds.

B.1 Population Interpretation of Prior Fitting

Population optimum. For the population interpretation of prior fitting, assume finite second moments, a model class that can represent the conditional and marginal means, and attainment of the Gaussian-NLL population minimum. For learned variance, additionally assume positive coordinate variances that the model can represent jointly with these means. Assume condition dropout is independent of the data and label, with probability strictly between zero and one. Then

$$\mu_\phi(y) = \mathbb{E}[x \mid y], \qquad \mu_\phi(\emptyset) = \mathbb{E}[x]. \tag{10}$$

For one coordinate, let $m_y = \mathbb{E}[x \mid y]$ and $q_y = \mathrm{Var}(x \mid y)$. At a positive scale s, its conditional population risk is

$$\mathbb{E}[\ell \mid y] = \frac{q_y + (\mu - m_y)^2}{2s^2} + \frac{1}{2}\log s^2,$$

which is minimized at $\mu = m_y$ for fixed s. When $q_y > 0$ and the variance is learned, the joint minimum is attained at $\mu = m_y$, $s^2 = q_y$. Independent dropout gives the marginal moments at the null input. This characterizes the ideal population optimum of Stage 1. The endpoint results below additionally require local ODE regularity; their pointwise alignment condition is $\Delta z_\alpha = \alpha \Delta\mu = d_\alpha^\star$.

B.1.1 Inverse-Variance Weighting

For the learned-variance option, the prior module outputs a mean and a positive variance:

$$\mu_\phi(y), \qquad \sigma_\phi^2(y) > 0. \tag{11}$$

Write $\ell_i(x, y)$ for the Gaussian NLL of coordinate i. The diagonal model sums the coordinatewise expressions:

$$\mathcal{L} = \mathbb{E}_{x,y}\left[\sum_i \left(\frac{(x_i - \mu_{\phi,i}(y))^2}{2\sigma_{\phi,i}^2(y)} + \frac{1}{2}\log \sigma_{\phi,i}^2(y)\right)\right]. \tag{12}$$

Proposition B.1 (Mean-output gradient)**.** *Under the NLL objective in Eq.* (12)*, the per-sample coordinate loss satisfies:*

$$\nabla_{\mu_\phi} \ell = -\frac{x - \mu_\phi(y)}{\sigma_\phi^2(y)}. \tag{13}$$

For a fixed residual, the magnitude of the mean-output gradient scales as $1/\sigma_\phi^2(y)$*.*

Proof. The terms depending on $\mu_\phi(y)$ are

$$\frac{1}{2\sigma_\phi^2(y)}\|x - \mu_\phi(y)\|^2. \tag{14}$$

Differentiating with respect to the mean output while holding the scale output fixed gives

$$\nabla_{\mu_\phi} \ell = \frac{1}{2\sigma_\phi^2(y)} \cdot 2(\mu_\phi(y) - x) = -\frac{x - \mu_\phi(y)}{\sigma_\phi^2(y)}. \tag{15}$$

Thus, holding the residual fixed, the mean-output gradient is weighted by $1/\sigma_\phi^2(y)$. □

Inverse-variance weighting couples mean fitting to source-scale learning in Stage 1. The resulting source then determines the initial distribution used to train the downstream flow. Tables 1 and 2 compare the resulting generation quality: learned variance improves some low-resolution settings, while fixed and learned variance give comparable ImageNet FIDs.

B.2 Scores and Posterior Means

Let $x \sim p(x)$ be the data distribution and let y denote the condition. The unconditional score is

$$s_t(x) = \nabla_x \log p_t(x), \tag{16}$$

and the conditional score is

$$s_t(x \mid y) = \nabla_x \log p_t(x \mid y). \tag{17}$$

By Bayes' rule,

$$s_t(x \mid y) - s_t(x) = \nabla_x \log p_t(y \mid x). \tag{18}$$

Hence, standard score-space CFG can be written as

$$s_{\text{cfg}}(x) = s_t(x) + w\big(s_t(x \mid y) - s_t(x)\big) = s_t(x) + w\nabla_x \log p_t(y \mid x). \tag{19}$$

B.3 Formal Assumptions

Throughout, write $v_c(x,t) = v_\theta(x,t \mid y)$ and $v_u(x,t) = v_\theta(x,t \mid \emptyset)$. We use the following local assumptions.

Assumption 1 (Regular flow map). For each fixed t, the flow map Φ_t^y of the conditional ODE $\dot{x}_t = v_c(x_t,t)$ is twice continuously differentiable in its initial state on a neighborhood of the source segment, with bounded Hessian there. All derivative bounds refer to the specified time and source region.

Convention (Shared-noise coupling). The conditional seed and paired reference share the same base noise ϵ, conditional scale, and mean coefficient $\alpha \in [0,1]$. The coefficient is fixed during Stage-2 training and reused at inference:

$$z_c^{(\alpha)} = \alpha\mu_\phi(y) + \sigma_\phi(y) \odot \epsilon, \qquad z_{\text{ref}}^{(\alpha)} = \alpha\mu_\phi(\emptyset) + \sigma_\phi(y) \odot \epsilon. \tag{20}$$

Their difference is the pure mean displacement $\Delta z_\alpha := z_c^{(\alpha)} - z_{\text{ref}}^{(\alpha)} = \alpha\Delta\mu$, where $\Delta\mu = \mu_\phi(y) - \mu_\phi(\emptyset)$. The primary setting uses $\alpha = 1$. At $\alpha = 0$, $\Delta z_0 = 0$ and P-Guide follows its unsteered conditional path.

Assumption 2 (Local perturbation regime). The guided seed is

$$z_{cfg}^{(\alpha)} = z_{\text{ref}}^{(\alpha)} + w\Delta z_\alpha = z_c^{(\alpha)} + (w-1)\Delta z_\alpha. \tag{21}$$

The segments from $z_{\text{ref}}^{(\alpha)}$ and $z_c^{(\alpha)}$ to $z_{cfg}^{(\alpha)}$ remain inside the smooth neighborhood of Assumption 1. The expansions below are local in $w\Delta z_\alpha$ and $(w-1)\Delta z_\alpha$, respectively.

B.4 Trajectory-Level Approximation

We now prove the main-text trajectory approximation in both reference-point and conditional-point forms.

Lemma B.1 (Mean-scale control of local error)**.** *Let $J_{t,\text{ref}}^{(\alpha)} = D_z\Phi_t^y(z_{\text{ref}}^{(\alpha)})$ and $J_{t,c}^{(\alpha)} = D_z\Phi_t^y(z_c^{(\alpha)})$. Taylor's formula gives*

$$\begin{aligned} \Phi_t^y(z_{cfg}^{(\alpha)}) &= \Phi_t^y(z_{\text{ref}}^{(\alpha)}) + wJ_{t,\text{ref}}^{(\alpha)}\Delta z_\alpha + R_{t,\text{ref}}, \\ \Phi_t^y(z_{cfg}^{(\alpha)}) &= \Phi_t^y(z_c^{(\alpha)}) + (w-1)J_{t,c}^{(\alpha)}\Delta z_\alpha + R_{t,c}, \end{aligned} \tag{22}$$

with integral remainders along the corresponding source segments. If $M_{t,\text{ref}}$ and $M_{t,c}$ bound the Hessian operator norm on these segments, then

$$\|R_{t,\text{ref}}\| \le \frac{M_{t,\text{ref}}}{2}\alpha^2 w^2\|\Delta\mu\|^2, \qquad \|R_{t,c}\| \le \frac{M_{t,c}}{2}\alpha^2(w-1)^2\|\Delta\mu\|^2. \tag{23}$$

Proof. Apply Taylor's formula with integral remainder first at $z_{\mathrm{ref}}^{(\alpha)}$ with displacement $w\Delta z_\alpha$, and then at $z_c^{(\alpha)}$ with displacement $(w-1)\Delta z_\alpha$. Taking norms, using $\int_0^1 (1-s)\,ds = \frac{1}{2}$, and substituting $\Delta z_\alpha = \alpha \Delta\mu$ gives both bounds. □

The reference-point form gives an $O(\|w\Delta z\|^2)$ error, whereas the conditional-point form isolates the additional effect beyond $w = 1$. If local curvature and neighborhood size remain comparable as α changes, reducing α contracts the perturbation linearly and the Taylor remainder quadratically; the corresponding local guidance range therefore scales approximately as $1/\alpha$.

B.5 Velocity Perturbation Induced by Prior Shift

Let $V_t(z) = v_c(\Phi_t^y(z), t)$ denote the velocity along the trajectory initialized at z.

Lemma B.2 (First-order velocity difference)**.** *Under Assumptions 1 and 2, if V_t is twice continuously differentiable in z on the segment from $z_{\mathrm{ref}}^{(\alpha)}$ to $z_c^{(\alpha)}$ and time and source derivatives commute, then*

$$V_t(z_c^{(\alpha)}) - V_t(z_{\mathrm{ref}}^{(\alpha)}) = \frac{d}{dt}\big(\Phi_t^y(z_c^{(\alpha)}) - \Phi_t^y(z_{\mathrm{ref}}^{(\alpha)})\big) = \dot{J}_{t,\mathrm{ref}}^{(\alpha)} \Delta z_\alpha + \tilde{R}_t, \tag{24}$$

where $\|\tilde{R}_t\| = O(\|\Delta z_\alpha\|^2) = O(\alpha^2 \|\Delta\mu\|^2)$.

Proof. By definition, $V_t(z) = \frac{d}{dt}\Phi_t^y(z)$. Taking the difference between the two trajectories,

$$V_t(z_c^{(\alpha)}) - V_t(z_{\mathrm{ref}}^{(\alpha)}) = \frac{d}{dt}\Big(\Phi_t^y(z_c^{(\alpha)}) - \Phi_t^y(z_{\mathrm{ref}}^{(\alpha)})\Big). \tag{25}$$

Applying the same integral Taylor formula to $z \mapsto V_t(z)$ gives

$$V_t(z_c^{(\alpha)}) - V_t(z_{\mathrm{ref}}^{(\alpha)}) = D_z V_t(z_{\mathrm{ref}}^{(\alpha)}) \Delta z_\alpha + \hat{R}_t, \tag{26}$$

with $\|\hat{R}_t\| = O(\|\Delta z_\alpha\|^2)$. Since $D_z V_t(z_{\mathrm{ref}}^{(\alpha)}) = \partial_t J_{t,\mathrm{ref}}^{(\alpha)}$, this yields

$$V_t(z_c^{(\alpha)}) - V_t(z_{\mathrm{ref}}^{(\alpha)}) = \dot{J}_{t,\mathrm{ref}}^{(\alpha)} \Delta z_\alpha + \tilde{R}_t, \tag{27}$$

where $\|\tilde{R}_t\| \le \frac{1}{2} \sup_{s\in[0,1]} \|D_z^2 V_t(z_{\mathrm{ref}}^{(\alpha)} + s\Delta z_\alpha)\|_{\mathrm{op}} \|\Delta z_\alpha\|^2$. □

B.6 First-Order Effect of Standard CFG

The following analysis characterizes source and velocity guidance on a shared source–backbone system: v_c and v_u come from the same checkpoint, and both endpoint families start from the learned conditional seed. This setting isolates when the two guidance operations agree. Appendix D.2 instantiates it with a checkpoint that provides both velocity branches. The main tables provide the complementary system-level comparison between separately trained Gaussian-source CFG and P-Guide source–backbone pairs.

Define $g(x,t) = v_c(x,t) - v_u(x,t)$. Standard CFG with scale w follows the modified field $v_u + w(v_c - v_u) = v_c + (w-1)g$. For fixed mean scale α, let $x_t^{(\alpha)} = \Phi_t^y(z_c^{(\alpha)})$, $J_t^{(\alpha)} = D_z\Phi_t^y(z_c^{(\alpha)})$, and $g_t^{(\alpha)} = g(x_t^{(\alpha)}, t)$. Let $x_t^{\mathrm{cfg},\alpha}(\eta)$ solve the modified ODE with $\eta = w - 1$ and $x_0^{\mathrm{cfg},\alpha}(\eta) = z_c^{(\alpha)}$.

Proposition B.2 (CFG endpoint change)**.** *If v_c is C^2 in state, g is C^1 in state, the relevant derivatives are locally bounded on a common solution neighborhood, and $J_t^{(\alpha)}$ is invertible for $t \in [0,1]$, then*

$$x_1^{\mathrm{cfg},\alpha}(\eta) - x_1^{(\alpha)} = \eta J_1^{(\alpha)} d_\alpha^\star + O(\eta^2), \qquad d_\alpha^\star = \int_0^1 (J_t^{(\alpha)})^{-1} g_t^{(\alpha)}\, dt. \tag{28}$$

Proof. Write $A_t = \nabla_x v_c(x_t^{(\alpha)}, t)$. Standard smooth dependence of ODE solutions on a parameter gives the first variation $\xi_t = \partial_\eta x_t^{\mathrm{cfg},\alpha}(\eta)\Big|_{\eta=0}$, which satisfies

$$\dot{\xi}_t = A_t \xi_t + g_t^{(\alpha)}, \qquad \xi_0 = 0. \tag{29}$$

The conditional-flow Jacobian satisfies $\dot{J}_t^{(\alpha)} = A_t J_t^{(\alpha)}$ and $J_0^{(\alpha)} = I$. Therefore

$$\frac{d}{dt}\Big((J_t^{(\alpha)})^{-1}\xi_t\Big) = (J_t^{(\alpha)})^{-1} g_t^{(\alpha)}, \tag{30}$$

and integration from 0 to 1 yields $\xi_1 = J_1^{(\alpha)} d_\alpha^\star$.

For completeness, the second variation $\zeta_t = \partial_\eta^2 x_t^{\mathrm{cfg},\alpha}(\eta)\Big|_{\eta=0}$ obeys

$$\dot{\zeta}_t = A_t\zeta_t + D_x^2 v_c(x_t^{(\alpha)}, t)[\xi_t, \xi_t] + 2D_x g(x_t^{(\alpha)}, t)\xi_t, \qquad \zeta_0 = 0. \tag{31}$$

The assumed local derivative bounds and the variation-of-constants formula bound ζ_t on $[0, 1]$. Hence the endpoint is twice differentiable in η with a locally bounded second derivative, and Taylor's theorem gives the stated $O(\eta^2)$ remainder. □

The vector $d_\alpha^\star$ pulls each per-step CFG velocity difference back to the scaled conditional source and integrates it over time.

B.7 Alignment Condition

Corollary B.1 (First-order alignment)**.** *For fixed condition y, shared noise realization ϵ, and mean scale α, P-Guide and standard CFG change the endpoint to first order in the same way if and only if*

$$\Delta z_\alpha = d_\alpha^\star = \int_0^1 (J_t^{(\alpha)})^{-1} g_t^{(\alpha)}\, dt. \tag{32}$$

Proof. By Lemma B.1 and Proposition B.2, both endpoint changes are $\eta J_1^{(\alpha)}$ applied to Δz_α and $d_\alpha^\star$, respectively. Invertibility of the ODE flow Jacobian gives the equivalence. □

At a realizable population optimum, moment fitting identifies $\Delta\mu$ with $\mathbb{E}[X \mid y] - \mathbb{E}[X]$, and mean scaling gives $\Delta z_\alpha = \alpha\Delta\mu$. Corollary B.1 adds the dynamical requirement $\Delta z_\alpha = d_\alpha^\star$ for first-order endpoint alignment. Decreasing α controls the perturbation size and quadratic remainder; the alignment residual remains the first-order term.

For a fixed condition, Δz_α is shared across noise draws, whereas $d_\alpha^\star$ varies with the reference trajectory through ϵ. The corollary therefore characterizes each coupled trajectory. The next proposition lifts the calculation to endpoint distributions and distinguishes almost-sure alignment from cancellation in expectation. The dependence of $J_1^{(\alpha)}$ and $d_\alpha^\star$ on ϵ is suppressed below.

Proposition B.3 (Weak first-order distributional response)**.** *Fix y and let ϵ be random under the shared-noise coupling. Write $X_P(\eta, \epsilon) = \Phi_1^y(z_c^{(\alpha)} + \eta\Delta z_\alpha)$, $X_G(\eta, \epsilon) = x_1^{\mathrm{cfg},\alpha}(\eta)$, and $X_0(\epsilon) = X_P(0, \epsilon) = X_G(0, \epsilon)$. Suppose the pointwise first variations above hold almost surely and, for some $\eta_0 > 0$, the norms of $\partial_\eta X_P(\eta, \epsilon)$ and $\partial_\eta X_G(\eta, \epsilon)$ are dominated on $|\eta| \le \eta_0$ by an integrable random variable. Then, for every $\psi \in C_b^1(\mathbb{R}^d)$,*

$$\frac{d}{d\eta}\left(\mathbb{E}_\epsilon[\psi(X_P(\eta, \epsilon))] - \mathbb{E}_\epsilon[\psi(X_G(\eta, \epsilon))]\right)\Big|_{\eta=0} = \mathbb{E}_\epsilon\Big[\nabla\psi(X_0)^\top J_1^{(\alpha)}\big(\Delta z_\alpha - d_\alpha^\star\big)\Big]. \tag{33}$$

Consequently, almost-sure pointwise alignment implies weak first-order agreement of the endpoint distributions. Weak agreement may also arise from cancellation of nonzero trajectory-level residuals in the displayed expectation.

Proof. For almost every ϵ, the chain rule and the first-variation formulas give

$$\partial_\eta\psi(X_P(\eta, \epsilon))|_{\eta=0} = \nabla\psi(X_0)^\top J_1^{(\alpha)}\Delta z_\alpha, \tag{34}$$

$$\partial_\eta\psi(X_G(\eta, \epsilon))|_{\eta=0} = \nabla\psi(X_0)^\top J_1^{(\alpha)} d_\alpha^\star. \tag{35}$$

Because $\nabla\psi$ is bounded and the endpoint derivatives have an integrable dominator, dominated differentiation permits interchanging the derivative and expectation. Subtracting the two identities proves the formula. Almost-sure pointwise alignment makes the integrand zero, while weak agreement only requires its expectation to vanish. □

B.8 Error Beyond First Order

Proposition B.4 (Finite-scale error)**.** *Define $F_P(\eta) = \Phi_1^y(z_c^{(\alpha)} + \eta\Delta z_\alpha)$ and $F_G(\eta) = x_1^{\mathrm{cfg},\alpha}(\eta)$. Suppose both are twice differentiable on the segment between 0 and η. If the endpoint Hessian of Φ_1^y is bounded there by $M_{1,c}$ and $\sup_s \|F_G''(s\eta)\| \le K_G$ for $s \in [0,1]$, then*

$$\begin{aligned} F_P(\eta) - F_G(\eta) &= \eta J_1^{(\alpha)}(\Delta z_\alpha - d_\alpha^\star) + R_\eta, \\ \|R_\eta\| &\le \frac{\eta^2}{2}\left(M_{1,c}\alpha^2\|\Delta\mu\|^2 + K_G\right). \end{aligned} \tag{36}$$

Consequently, the norm of the endpoint mismatch is bounded by the norm of the displayed first-order term plus the same quadratic bound.

Proof. Both endpoint families agree at zero: $F_P(0) = F_G(0) = x_1^{(\alpha)}$. Their first derivatives are $F_P'(0) = J_1^{(\alpha)}\Delta z_\alpha$ by the chain rule and $F_G'(0) = J_1^{(\alpha)} d_\alpha^\star$ by Proposition B.2. Taylor's formula with integral remainder, written on the fixed interval $s \in [0,1]$, gives

$$R_\eta = \eta^2 \int_0^1 (1-s)\big(F_P''(s\eta) - F_G''(s\eta)\big)\, ds. \tag{37}$$

Because $F_P''(s\eta) = D_z^2\Phi_1^y(z_c^{(\alpha)} + s\eta\Delta z_\alpha)[\Delta z_\alpha, \Delta z_\alpha]$, we have $\|F_P''(s\eta)\| \le M_{1,c}\|\Delta z_\alpha\|^2 = M_{1,c}\alpha^2\|\Delta\mu\|^2$. Taking norms, applying the bound on F_G'', and using $\int_0^1 (1-s)\, ds = \frac{1}{2}$ proves the result for either sign of η. □

The first term is the alignment residual, while smoothness controls the quadratic term. The P-Guide contribution scales explicitly as $\alpha^2(w-1)^2$; the CFG constant K_G and the alignment residual need not decrease with α because the base trajectory, $d_\alpha^\star$, and $J_t^{(\alpha)}$ also change with the scaled starting point.

B.9 Exact Special Cases

Affine fields with a common linear part. Let $v_c(x,t) = B_t x + b_c(t)$ and $v_u(x,t) = B_t x + b_u(t)$ share the linear part B_t. Then $g_t = b_c(t) - b_u(t)$ is state independent, both endpoint maps are affine in η, and the endpoint equality $x_1^P(\eta) = x_1^{\mathrm{cfg}}(\eta)$ holds for every finite η if and only if

$$\Delta z_\alpha = \int_0^1 (J_t^{(\alpha)})^{-1}\big(b_c(t) - b_u(t)\big)\, dt, \tag{38}$$

the same pullback identity with g_t state independent. Distinct initial states still give distinct trajectories whenever $(w-1)\Delta z_\alpha \neq 0$.

Shared-covariance Gaussians. If at some time t the conditional and null marginals are $\mathcal{N}(m_c, C_t)$ and $\mathcal{N}(m_u, C_t)$ with $C_t \succ 0$, and $s_r = \nabla_x \log p_t(x \mid r)$, then $s_c - s_u = C_t^{-1}(m_c - m_u)$ is constant in x, and the standard CFG score difference is a fixed linear image of the mean difference. This score-level relation supports the mean-difference construction; endpoint correspondence is governed by the pullback identity in Corollary B.1. A class-mixture null marginal generally falls outside this shared-covariance Gaussian case.

B.10 A Motivating Connection to the Conditional Score

The following posterior-mean identity under additive Gaussian corruption connects source conditioning to conditional scores and motivates noisy-state-conditioned source modules. The endpoint-alignment result above follows from the flow-map analysis.

Proposition B.5 (Posterior-mean difference under additive Gaussian corruption)**.** *Let x_{clean} be a clean random variable whose conditional and unconditional corrupted densities are positive and differentiable in x_t, and assume the relevant conditional first moments are finite. Let*

$$x_t = x_{\mathrm{clean}} + \sigma_t\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{39}$$

where ϵ is independent of (x_{clean}, y) and $\sigma_t > 0$. Then

$$\mathbb{E}[x_{\text{clean}} \mid x_t, y] - \mathbb{E}[x_{\text{clean}} \mid x_t] = \sigma_t^2(\nabla_{x_t} \log p_t(x_t \mid y) - \nabla_{x_t} \log p_t(x_t)) = \sigma_t^2 \nabla_{x_t} \log p(y \mid x_t). \tag{40}$$

Proof. Tweedie's identity for the stated additive corruption gives $\mathbb{E}[x_{\text{clean}} \mid x_t] = x_t + \sigma_t^2 \nabla_{x_t} \log p_t(x_t)$ and, conditionally, $\mathbb{E}[x_{\text{clean}} \mid x_t, y] = x_t + \sigma_t^2 \nabla_{x_t} \log p_t(x_t \mid y)$. Subtracting and applying Bayes' rule proves the claim. □

Proposition B.5 applies to additive Gaussian corruption. Its posterior means condition on x_t, whereas P-Guide conditions on y. Accordingly, the P-Guide result rests on the local trajectory response of Theorem 1 and the alignment condition above.

B.11 Stability of Velocity-Space Control

For comparison with source steering, velocity-space CFG directly changes the ODE vector field. Consider

$$v_{\text{cfg}}(x, t) = v_u(x, t) + w\big(v_c(x, t) - v_u(x, t)\big). \tag{41}$$

Assume v_u, v_c are jointly continuous on $[0, T] \times \mathbb{R}^d$ and globally Lipschitz in x, uniformly in t, with constants L_u, L_c. For fixed w, the guided field has Lipschitz constant at most $|1 - w|L_u + |w|L_c$, so the initial-value problem has a unique solution on $[0, T]$. Let x_t^{base} solve $\dot{x}_t = v_u(x_t, t)$. For guided and base trajectories from the same initial state, set $e(t) = \|x_t^{\text{cfg}} - x_t^{\text{base}}\|$ and $g(t) = \|w(v_c - v_u)(x_t^{\text{cfg}}, t)\|$. Subtracting the two ODEs gives $e(t) \le \int_0^t [L_u e(s) + g(s)]\, ds$. Grönwall's inequality then yields

$$\|x_t^{\text{cfg}} - x_t^{\text{base}}\| \le e^{Lt} \int_0^t \|w(v_c(x_s, s) - v_u(x_s, s))\|\, ds, \tag{42}$$

where $L = L_u$ and $x_s = x_s^{\text{cfg}}$. More precisely, $e(t) \le \int_0^t e^{L(t-s)} g(s)\, ds$, which gives the displayed bound. The same estimate holds under local Lipschitz conditions on a common region containing both trajectories throughout $[0, T]$, provided the solutions exist on that interval. We use this bound as a comparison baseline for the well-posedness and continuity of velocity-space guidance.

P-Guide instead changes the initial condition while leaving the downstream vector field unchanged. Its sensitivity is governed by the flow Jacobian J_t and differs from the accumulated velocity forcing above.

B.12 Unified Local Correspondence Theorem

Under the shared-noise construction, the guided prior is

$$z_{\text{cfg}}^{(\alpha)} = \alpha\big[\mu_\phi(\emptyset) + w\big(\mu_\phi(y) - \mu_\phi(\emptyset)\big)\big] + \sigma_\phi(y) \odot \epsilon = z_{\text{ref}}^{(\alpha)} + w\Delta z_\alpha = z_c^{(\alpha)} + (w - 1)\Delta z_\alpha \tag{43}$$

Theorem B.1 (Unified local correspondence)**.** *Fix y and ϵ. Let $\eta = w - 1$ and suppose the assumptions of Lemma B.1 and Proposition B.2 hold. On a local interval where the CFG endpoint has second derivative bounded by K_G,*

$$\begin{aligned} F_P(\eta) - F_G(\eta) &= \eta J_1^{(\alpha)}(\Delta z_\alpha - d_\alpha^\star) + R_\eta, \\ \|R_\eta\| &\le \frac{\eta^2}{2}\big(M_{1,c}\alpha^2\|\Delta\mu\|^2 + K_G\big). \end{aligned} \tag{44}$$

Consequently, P-Guide and standard CFG have the same pointwise first-order endpoint response for this trajectory if and only if

$$\Delta z_\alpha = d_\alpha^\star = \int_0^1 (J_t^{(\alpha)})^{-1} g_t^{(\alpha)}\, dt. \tag{45}$$

The P-Guide contribution to the finite-scale remainder contracts quadratically with α, while the alignment residual remains an explicit, separately identifiable term.

Proof. Lemma B.1 gives $F'_P(0) = J_1^{(\alpha)} \Delta z_\alpha$ and the explicit P-Guide remainder. Proposition B.2 gives $F'_G(0) = J_1^{(\alpha)} d_\alpha^\star$, and Proposition B.4 bounds their finite-scale difference. Since $J_1^{(\alpha)}$ is invertible, equality of the first derivatives is equivalent to $\Delta z_\alpha = d_\alpha^\star$. □

The reference-point Taylor remainder is $O(\alpha^2 w^2 \|\Delta\mu\|^2)$, whereas the conditional-point remainder in the theorem is $O(\alpha^2(w-1)^2\|\Delta\mu\|^2)$. The same α defines the Stage-2 training and inference sources, and neither α nor w changes the shared random term.

B.13 Closed-Form CFG in Distribution Space

Consider the unconditional and conditional source distributions:

$$p_u(x) = \mathcal{N}(\mu_u, \sigma_u^2 \mathbf{I}), \quad p_c(x) = \mathcal{N}(\mu_c, \sigma_c^2 \mathbf{I}). \tag{46}$$

A natural distribution-level analogue of CFG is:

$$p_{\text{cfg}}(x) \propto p_u(x)^{1-w} p_c(x)^w, \tag{47}$$

which corresponds to a linear interpolation in log-density space:

$$\log p_{\text{cfg}}(x) = (1-w)\log p_u(x) + w \log p_c(x) + C. \tag{48}$$

Substituting the Gaussian log-density

$$\log p(x) = -\frac{1}{2\sigma^2}\|x-\mu\|^2 - \frac{1}{2}\log(2\pi\sigma^2), \tag{49}$$

and ignoring constants independent of x, we obtain:

$$\log p_{\text{cfg}}(x) = -\frac{1-w}{2\sigma_u^2}\|x-\mu_u\|^2 - \frac{w}{2\sigma_c^2}\|x-\mu_c\|^2 + C. \tag{50}$$

We now expand both quadratic terms explicitly:

$$\|x-\mu_u\|^2 = \|x\|^2 - 2\mu_u^\top x + \|\mu_u\|^2, \tag{51}$$

$$\|x-\mu_c\|^2 = \|x\|^2 - 2\mu_c^\top x + \|\mu_c\|^2. \tag{52}$$

Substituting back:

$$\begin{aligned}\log p_{\text{cfg}}(x) = &-\frac{1-w}{2\sigma_u^2}(\|x\|^2 - 2\mu_u^\top x + \|\mu_u\|^2) \\ &- \frac{w}{2\sigma_c^2}(\|x\|^2 - 2\mu_c^\top x + \|\mu_c\|^2) + C.\end{aligned} \tag{53}$$

Grouping terms by powers of x:

$$\begin{aligned}\log p_{\text{cfg}}(x) = &-\frac{1}{2}\left(\frac{1-w}{\sigma_u^2} + \frac{w}{\sigma_c^2}\right)\|x\|^2 \\ &+ \left(\frac{1-w}{\sigma_u^2}\mu_u + \frac{w}{\sigma_c^2}\mu_c\right)^\top x + C'.\end{aligned} \tag{54}$$

Matching this with the canonical Gaussian form

$$-\frac{1}{2\sigma_{\text{cfg}}^2}\|x-\mu_{\text{cfg}}\|^2, \tag{55}$$

we obtain:

$$\frac{1}{\sigma_{\text{cfg}}^2} = \frac{1-w}{\sigma_u^2} + \frac{w}{\sigma_c^2}, \tag{56}$$

$$\mu_{\text{cfg}} = \sigma_{\text{cfg}}^2 \left(\frac{1-w}{\sigma_u^2}\mu_u + \frac{w}{\sigma_c^2}\mu_c\right). \tag{57}$$

This is a normalizable Gaussian only when $(1-w)/\sigma_u^2 + w/\sigma_c^2 > 0$; for $w > 1$ this condition can fail, and in the diagonal case it must hold in every coordinate.

In the special case where $\sigma_u = \sigma_c = \sigma$, the above expressions simplify to:

$$\frac{1}{\sigma_{\text{cfg}}^2} = \frac{1}{\sigma^2}, \quad \mu_{\text{cfg}} = (1-w)\mu_u + w\mu_c. \tag{58}$$

In the equal-covariance case, the formal analogue reduces to linear mean extrapolation with unchanged covariance, $\mu_{\text{cfg}} = \mu_c + (w-1)(\mu_c - \mu_u)$, matching the source marginal induced by the default $\alpha = 1$ mean-only rule in Eq. (9). Choosing $\alpha < 1$ scales the mean in both Stage-2 training and inference to control the perturbation magnitude rather than preserve exact density extrapolation.

With unequal scales, natural-parameter density extrapolation and mean-only P-Guide define different source marginals. P-Guide supplements the marginal specification with the shared-noise coupling

$$z_c^{(\alpha)} = \alpha\mu_c + \sigma_c \odot \epsilon, \qquad z_{\text{ref}}^{(\alpha)} = \alpha\mu_u + \sigma_c \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{59}$$

which fixes the displacement at $\alpha(\mu_c - \mu_u)$ and preserves the conditional covariance. Appendix C.3 evaluates an affine mean-plus-scale construction, while Appendix C.1 varies α with fixed variance. These source constructions act through the boundary condition, while standard CFG supplies distributed forcing along the trajectory.

B.14 Distribution-Level CFG as Source Reweighting

Distribution-level interpolation changes the *shape of the source density* but leaves the *trajectory dynamics* fixed.

Sampling directly from $p_{\text{cfg}}(x)$ effectively:

- defines a new initial distribution,
- while keeping the dynamics $v(x_t, t \mid z)$ unchanged.

The resulting samples follow the transport of the re-weighted source under the fixed flow. A marginal density such as p_{cfg} leaves the pairing of conditional and null samples unspecified. P-Guide supplies that pairing through the shared ϵ, giving every guided source sample the deterministic displacement $\Delta z_\alpha = z_c^{(\alpha)} - z_{\text{ref}}^{(\alpha)}$. This coupling supports the local initial-condition analysis above, while Corollary B.1 characterizes correspondence with per-step CFG.

B.15 Empirical Comparison on MNIST

We compare the closed-form distribution-level Gaussian $p_{\text{cfg}}(x)$ with P-Guide's shared-noise source on MNIST. Both sources use the same initial noise and guidance scale ($w = 1.5$) and are propagated through the same adapted flow.

Figure 8 in Appendix F presents this qualitative comparison. At $w = 1.5$, the log-density source produces more distorted digits, while the shared-noise construction preserves clearer digit structure. The behavior is consistent across digit classes: increasing the guidance scale does not lead to systematic improvements under the distribution-level formulation.

The comparison shows how a fixed-flow source reweighting depends on its extrapolation geometry, especially when conditional and null source variances differ.

Table 4: Train–test-consistent mean-scale ablation on ImageNet-1k with DiT-B/2 and fixed variance $\sigma = 1$. For $\alpha > 0$, the same α is used in Stage-2 training and sampling, and the reported scales are P-Guide scales w_{PG} with $w_{\mathrm{CFG}} = 1$. The gray $\alpha = 0$ row is a separate standard-CFG reference and reports the matched w_{CFG} values from the main ImageNet comparison. Pass and GFLOPs are listed as $(w = 1.0 \,/\, w > 1)$. Bold marks the lowest reported FID in each row. N/A denotes a scale outside the meaningful P-Guide range for that configuration.

α	$w = 1.0$	1.1	1.2	1.5	2.0	Pass $(1.0 \,/\, > 1)$	GFLOPs $(1.0 \,/\, > 1)$
1.0	32.84	26.41	**24.61**	38.72	N/A	$1 \,/\, 1$	$2301.60 \,/\, 2303.08$
0.5	33.83	29.77	**27.30**	29.66	N/A	$1 \,/\, 1$	$2301.60 \,/\, 2303.08$
0.1	31.24	30.46	29.65	27.70	**25.35**	$1 \,/\, 1$	$2301.60 \,/\, 2303.08$
0 (CFG)	35.04	27.92	22.05	**11.03**	15.28	$1 \,/\, 2$	$2300.12 \,/\, 4600.24$

C Ablation Studies

The ablations proceed from the core mechanism to its implementation: mean scale, learned source scale, scale extrapolation, prior capacity, and backbone initialization.

C.1 Effect of Mean Scale

Equation (9) uses a mean-scale coefficient α in both Stage-2 training and sampling. The primary setting uses $\alpha = 1$. The ablations use $\alpha = 0.5$ and $\alpha = 0.1$ to reduce the class-dependent mean displacement while preserving the noise distribution. At $\alpha = 0$, the source becomes standard Gaussian noise and $\Delta z_0 = 0$, so P-Guide follows its unsteered conditional path. The gray row reports standard CFG as a separate reference. All rows use ImageNet-1k, DiT-B/2, fixed variance $\sigma = 1$, and 50 solver steps.

At a fixed w, α scales the class-dependent mean displacement while leaving the noise term and conditional standard deviation unchanged. A smaller α supports a wider tested guidance range. With $\alpha = 1$, FID improves through $w = 1.2$ but rises to 38.72 at $w = 1.5$, placing $w = 2.0$ outside its meaningful range. With $\alpha = 0.5$, the $w = 1.5$ result remains better than its $w = 1$ baseline, while $w = 2.0$ is likewise outside the selected range. With $\alpha = 0.1$, FID continues to improve through $w = 2.0$, reaching 25.35. This trend agrees with Theorem 1: the effective perturbation is $\alpha(w-1)\Delta\mu$, and the local quadratic remainder scales as $\alpha^2(w-1)^2$. Reducing α therefore permits a larger w for a comparable displacement and local error bound. The $\alpha = 0$ row provides the separate standard-CFG reference.

We further evaluate $\alpha = 0.01$ with standard CFG fixed at $w_{\mathrm{CFG}} = 1$ and sweep the P-Guide scale w_{PG} from 1 to 40 (Figure 6). The FID decreases from 34.07 at $w_{\mathrm{PG}} = 1$ to 27.21 at $w_{\mathrm{PG}} = 20$, then increases to 33.43 at $w_{\mathrm{PG}} = 40$. Thus, the near-zero mean scale shifts the best operating point to a larger P-Guide scale, while stronger guidance beyond $w_{\mathrm{PG}} = 20$ degrades FID.

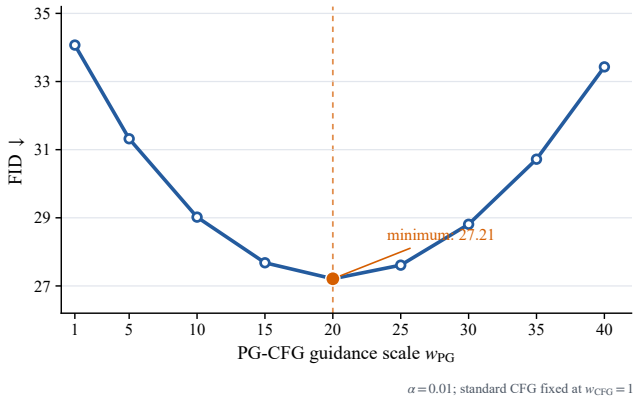


Figure 6: Near-zero mean-scale ablation on ImageNet-1k with $\alpha = 0.01$. The standard CFG scale is fixed at $w_{\mathrm{CFG}} = 1$; all horizontal-axis values are P-Guide scales w_{PG}.

C.2 Learned Prior Scale Diagnostics

The learned standard deviation σ_ϕ has mean/std/min/max $0.820/0.127/0.446/1.634$ over the 1000 ImageNet classes. Here, std describes variation in predicted source scales across classes. The mean scales are 0.472 in the CIFAR-10 setting and 0.164 on MNIST.

ImageNet's average scale is closer to the fixed value 1 than the low-resolution scales, consistent with the smaller differences between learned- and fixed-variance FIDs. The direction of those FID differences varies across backbones and guidance scales (Table 2). The smaller learned scales in the low-resolution settings change source spread more markedly, with the FID effect depending on guidance strength.

Variance learning changes both the residual weighting during prior fitting and the noise scale used to train and sample the flow. These effects are coupled in the source-learning comparisons. During inference, the conditional variance sets the source spread, and the mean shift performs guidance. Table 5 isolates additional scale extrapolation at inference: the FID changes are small and directionally mixed at moderate guidance, with a larger degradation on U-Net at $w = 1.5$.

C.3 Effect of Additional Scale Guidance

We next test whether guidance should also modify the source scale. With $\alpha = 1$ in training and sampling, the main P-Guide rule steers only the prior mean and retains the conditional standard deviation:

$$z_{cfg}^{(\alpha)} = \alpha[\mu_\phi(\emptyset) + w(\mu_\phi(y) - \mu_\phi(\emptyset))] + \sigma_\phi(y) \odot \epsilon. \tag{60}$$

To assess whether additional scale guidance helps, we compare it with an ablation that also extrapolates the standard deviation:

$$z_{cfg}^{\text{mean+scale},\alpha} = \alpha[\mu_\phi(\emptyset) + w(\mu_\phi(y) - \mu_\phi(\emptyset))] + \big[\sigma_\phi(\emptyset) + w(\sigma_\phi(y) - \sigma_\phi(\emptyset))\big] \odot \epsilon. \tag{61}$$

Table 5 compares both rules on U-Net and DiT-B/2 under single-pass inference.

Table 5: Mean-only CFG (main method) vs. additional scale guidance (Mean+Std, which extrapolates the standard deviation). Results are reported in FID (↓) across different guidance scales w.

	U-Net		DiT-B/2	
w	Mean-only	Mean+Std	Mean-only	Mean+Std
1.1	22.33	22.09	26.80	26.71
1.2	23.01	23.14	25.07	24.75
1.5	44.53	47.62	38.43	38.31

Mean-only and mean-plus-scale guidance yield similar ImageNet FIDs at $w = 1.1$ and 1.2, with absolute differences of 0.09–0.32. The direction of the change varies across settings. At $w = 1.5$, adding scale guidance worsens U-Net FID from 44.53 to 47.62, while both variants have already degraded relative to moderate guidance.

Together with the learned-versus-fixed source-scale comparison in Table 2, these results separate two roles: the conditional scale determines source spread, whereas the mean displacement provides the effective guidance operation.

C.4 Influence of Prior Steering Module Capacity

We compare the default 1.31M-parameter prior module described in Appendix A with PG-Large, an enlarged 13.1M-parameter variant.

PG-Large increases the number of residual blocks and the hidden width while retaining the same overall design.

Table 6 shows that the 1.31M-parameter default module attains FID 33.68 at 2301.60 GFLOPs, compared with 34.34 at 2303.65 GFLOPs for the approximately $10\times$ larger variant.

Table 6: Effect of prior module capacity on ImageNet-1k (256×256). All results use DiT-B/2 backbone with a single inference pass.

PG Module	Params (M)	FID ↓	GFLOPs ↓
P-Guide (Default)	1.31	33.68	2301.60
P-Guide (Large)	13.1	34.34	2303.65

The default module therefore provides the better FID with one tenth of the prior-module parameters.

C.5 Effect of Initialization: Pretrained vs. From-Scratch Training

To isolate the effect of backbone initialization, we compare training from scratch with initialization from a pretrained flow model (Dao et al., 2023). Both DiT-B/2 models use single-pass inference and receive the same 400K-step Stage-2 training budget.

Table 7 reports the resulting FID at identical sampling cost.

Table 7: Comparison between pretrained initialization and training from scratch for Stage-2 P-Guide. Both models are trained for 400K steps.

Initialization	FID ↓	GFLOPs ↓
From Scratch (400K)	49.03	2301.60
Pretrained + 400K	33.68	2301.60

Under the same 400K-step Stage-2 budget, pretrained initialization improves FID from 49.03 to 33.68 at identical sampling cost. The pretrained model additionally uses the backbone training completed before Stage 2.

Both settings use the learned-source training procedure, so this comparison measures the effect of backbone initialization on final generation quality.

D EXTENSIONS OF PRIOR STEERING

This section tests three extensions of source-mean control: distributing the shift across solver updates, combining it with velocity guidance, and applying it to a scale-conditioned generator.

D.1 DISTRIBUTED PRIOR STEERING

To examine how injection timing affects the effective guidance range, we distribute the mean displacement across several solver updates using the same checkpoint. This changes the sampling rule while retaining one backbone evaluation per solver update. This experiment uses the default mean scale $\alpha = 1$. In the fixed-variance formulation, let $d = \mu_c - \mu_u$, $\beta = w_{\mathrm{PG}} - 1$, and $z_c = \mu_c + \epsilon$. One-shot P-Guide starts from $z_c + \beta d$. The distributed variant starts from z_c and uses

$$x_{k+1} = \mathcal{S}_k(x_k) + \beta a_k d, \qquad a_k \geq 0, \quad \sum_k a_k = 1, \tag{62}$$

where $\mathcal{S}_k$ denotes a solver update and weights are nonzero only at selected injection steps. The total injected vector is βd, and the injection times determine the final trajectory. For S solver updates, the sampling convention uses $K_{\mathrm{eff}} = \min(K, S-1)$ and applies the injections after the first K_{eff} updates; $K = 0$ denotes the original one-shot initialization.

Table 8: Distributed P-Guide diagnostic at 20 solver steps using the same checkpoint across rows. Entries are FID. K is the requested number of injection updates; the schedule caps it at $S - 1$.

K	$w_{\mathrm{PG}} = 1.0$	$w_{\mathrm{PG}} = 1.5$	$w_{\mathrm{PG}} = 2.0$
0 (one-shot)	33.38	37.94	82.68
5	33.38	28.61	54.90
10	33.38	26.27	37.71
15	33.38	26.36	30.41
20	33.38	26.72	28.41

Table 8 shows that injection timing materially expands the useful guidance range. Distributing the displacement improves FID from 37.94 to 26.27 at $w = 1.5$ and from 82.68 to 28.41 at $w = 2.0$; both results are below the $w = 1$ conditional-source baseline of 33.38.

For the discrete sampler, fix S, d, and weights a_k independent of β. Let $\bar{x}_0 = z_c$, $\bar{x}_{k+1} = \mathcal{S}_k(\bar{x}_k)$ be the reference trajectory at $\beta = 0$. Assume each update map is C^2 near that trajectory. Differentiating the sampler with respect to β at zero gives

$$\eta_{k+1} = B_k \eta_k + a_k d, \qquad B_k = D\mathcal{S}_k(\bar{x}_k), \qquad \eta_0 = 0. \tag{63}$$

Here $\eta_k = \partial_\beta x_k(\beta)|_{\beta=0}$. Define the discrete propagator $P_{j,i} = B_{j-1} \cdots B_i$ for $i < j$, and $P_{i,i} = I$. Unrolling the recurrence yields

$$x_S(\beta) - \bar{x}_S = \beta \sum_{k=0}^{S-1} a_k P_{S,k+1} d + O(\beta^2). \tag{64}$$

The index $k + 1$ corresponds to injection after update k. The remainder is local in β for this fixed schedule. One-shot initialization instead gives $\beta P_{S,0} d + O(\beta^2)$. If each injection used the transported direction $P_{k+1,0} d$, then

$$\beta \sum_{k=0}^{S-1} a_k P_{S,k+1} P_{k+1,0} d = \beta P_{S,0} d. \tag{65}$$

The implemented sampler uses the fixed direction d, so injection times change its endpoint response through the factors $P_{S,k+1}$. Cross-step interactions enter the remainder alongside the individual quadratic terms. For exact ODE updates, these products reduce to the continuous variational propagators in Appendix B.

D.2 Compatibility with Standard CFG

Joint source and velocity control combines P-Guide at initialization with standard classifier-free guidance (CFG) during ODE integration. This experiment therefore uses a checkpoint and protocol that provide both conditional and null velocity evaluations, unlike the main single-pass P-Guide sampler.

We evaluate combinations of guidance scales for P-Guide (w_{PG}) and standard CFG (w_{CFG}). Specifically, $w_{\text{PG}} \in \{1.0, 1.1, 1.2\}$ and $w_{\text{CFG}} \in \{1.0, 1.1, 1.2, 1.5, 2.0\}$, yielding 15 configurations. Table 9 reports results on ImageNet-1k (256×256) with a U-Net backbone.

Table 9: Joint evaluation of P-Guide and standard CFG on ImageNet-1k (256×256). Rows correspond to P-Guide scale w_{PG}, columns correspond to CFG scale w_{CFG}. All results use a U-Net backbone with the learned P-Guide source; $w_{\text{PG}} = 1$ retains the conditional prior. For $w_{\text{CFG}} > 1$, sampling uses two backbone evaluations per step.

$w_{\text{PG}} \backslash w_{\text{CFG}}$	1.0	1.1	1.2	1.5	2.0
1.0	27.78	23.37	20.01	15.44	25.29
1.1	22.33	18.74	16.16	13.26	21.00
1.2	23.01	20.43	18.77	18.54	29.35

The two controls are complementary within a bounded range. FID reaches 13.26 at $(w_{\text{PG}}, w_{\text{CFG}}) = (1.1, 1.5)$ but rises to 29.35 at $(1.2, 2.0)$. Because standard CFG remains active, this combined sampler uses two backbone evaluations per step.

D.3 Compatibility with a Scale-Conditioned Few-Step Generator

We also evaluate P-Guide with a MNIST drifting-model generator trained with a continuous generator-scale input and class-label dropout. We denote this separate control by γ_{model} to distinguish it from P-Guide's mean scale α. Table 10 reports the standard-CFG baseline. Tables 11 and 12 vary one control at a time, holding the other fixed.

Table 10: MNIST drifting-model baseline evaluated with standard CFG.

γ_{model}	FID ↓	Accuracy (%) ↑
1.0	25.21	95.19
1.1	26.95	95.32
1.2	29.47	95.11
1.5	37.00	94.90
2.0	49.01	95.00
3.0	80.17	93.85

Table 11: P-Guide prior-scale sweep for the MNIST drifting model with $\gamma_{\text{model}} = 1.0$.

w_{PG}	FID ↓	Accuracy (%) ↑
1.0	43.40	78.59
1.1	40.24	82.78
1.2	37.10	86.63
1.5	30.07	93.12
2.0	24.40	97.68
3.0	23.14	99.49
4.0	27.15	99.53
5.0	32.83	99.48

Table 12: Generator-scale sweep for the MNIST drifting model with $w_{\text{PG}} = 3.0$.

γ_{model}	FID ↓	Accuracy (%) ↑
1.0	22.94	99.44
1.1	22.43	99.47
1.2	21.87	99.52
1.5	19.45	99.58
2.0	18.03	99.76
3.0	33.50	94.46

Prior steering and generator conditioning jointly control the trained MNIST model. The best FID is 18.03 at $(w_{\text{PG}}, \gamma_{\text{model}}) = (3.0, 2.0)$, with 99.76% accuracy, while larger control values degrade FID. Independent evaluations of the shared point $(3.0, 1.0)$ yield FIDs 23.14 and 22.94 in the two sweeps.

E SMALL-SCALE TEXT-TO-IMAGE PILOT

The main experiments study a finite class-conditioning space. This pilot extends P-Guide to continuous text representations and asks two questions: whether a learned source mean remains compatible with a pretrained text-conditioned generator, and how source steering changes prompt following under controlled sampling.

E.1 EXPERIMENTAL SETUP

Data and backbone. We use 20,000 image–caption pairs cached from the COCO train2017 split (Lin et al., 2014) and the 600M-parameter, 512-pixel SANA diffusion-transformer checkpoint (Xie et al., 2024). Images are encoded into $32 \times 16 \times 16$ latents. We compare (i) standard LoRA fine-tuning (Hu et al., 2021) from an isotropic Gaussian source, (ii) PG-LoRA evaluated with P-Guide disabled, and (iii) the same PG-LoRA with the learned source mean enabled. Both LoRA variants update the query, key, and value projections with rank 16 and contain 6.19M trainable LoRA parameters. Each variant is optimized for 10,000 steps in bfloat16 with an effective batch size of four, a peak LoRA learning rate of 10^{-4}, 500 warmup steps, and cosine decay.

Text-conditioned source. The pilot uses a token-aware prior that maps frozen SANA text-encoder features to a latent mean. It contains 2.24M trainable parameters, two attention layers, eight heads, and hidden width 192. Its variance is fixed; only the conditional mean is learned. The prior and LoRA backbone use learning rates 2×10^{-5} and 10^{-4}, respectively, and text conditions are replaced by the null condition with probability 0.1. This parameterization follows the broader principle of learning condition-dependent source distributions for text-to-image flow models (Kim et al., 2026).

Sampling protocol. We evaluate 30 held-out compositional prompts covering object presence, attributes, counting, spatial or semantic relations, and unusual object combinations. Every method uses the same prompt, seed, 20-step Euler schedule, and conventional CFG scale of 4.5. Guided PG-LoRA additionally applies a one-shot source-mean extrapolation with scale 1.5 at initialization. Because all variants retain conventional CFG, this experiment evaluates compatibility and prompt adherence rather than single-pass efficiency.

E.2 AUTOMATIC PROMPT-FOLLOWING EVALUATION

We decompose the 30 prompts into 129 atomic checks: 51 object, 44 attribute, 3 counting, and 31 relation questions. A local BLIP VQA model (Li et al., 2022) answers each question. We report mean atomic accuracy (Atom), the fraction of prompts satisfying every check (All), category-level accuracy, a global yes/no prompt-match score (Global), and a semantic-consistency aggregate (Semantic). The protocol draws on question-answering and scene-graph faithfulness evaluation (Hu et al., 2023; Cho et al., 2023), object-level alignment (Ghosh et al., 2023), compositional evaluation (Huang et al., 2023), VQA-based global alignment (Lin et al., 2024; Li et al., 2024), and separate semantic-quality assessment (Ku et al., 2024). Because our questions and aggregation differ from the official pipelines, we report these values as local proxy metrics.

Table 13: Automatic prompt-following results on 30 held-out compositional prompts (%). Standard LoRA is the quantitative reference. "PG off" evaluates PG-LoRA without source steering, while "PG guidance" applies the learned source-mean shift.

Method	Atom	All	Object	Attr.	Count	Relation	Global	Semantic
Standard LoRA	82.17	40.00	90.20	81.82	66.67	70.97	90.00	86.09
PG-LoRA (PG off)	**82.95**	40.00	90.20	81.82	**100.00**	70.97	86.67	84.81
PG-LoRA + PG guidance	82.17	36.67	88.24	79.55	**100.00**	**74.19**	**93.33**	**87.75**
Δ PG off vs. standard	+0.78	+0.00	+0.00	+0.00	+33.33	+0.00	−3.33	−1.28
Δ PG guidance vs. standard	+0.00	−3.33	−1.96	−2.27	+33.33	+3.23	+3.33	+1.66

E.3 RESULTS

With source steering disabled, PG-LoRA raises atomic accuracy from 82.17% to 82.95% while leaving the all-correct, object, attribute, and relation scores unchanged. Enabling the source shift redistributes the satisfied constraints: relative to standard LoRA, relation accuracy increases by 3.23 points, global prompt match by 3.33 points, and semantic consistency by 1.66 points. Counting accuracy increases from 66.67% to 100%, although the category contains only three questions and the change corresponds to one additional correct answer.

Thus, source steering changes the balance among constraint types rather than uniformly improving every metric. Relation accuracy, global prompt match, and semantic consistency increase, while all-correct, object, and attribute accuracy decrease and aggregate atomic accuracy remains unchanged. Under this protocol, the result shows compatibility with LoRA adaptation and identifies a prompt-level trade-off.

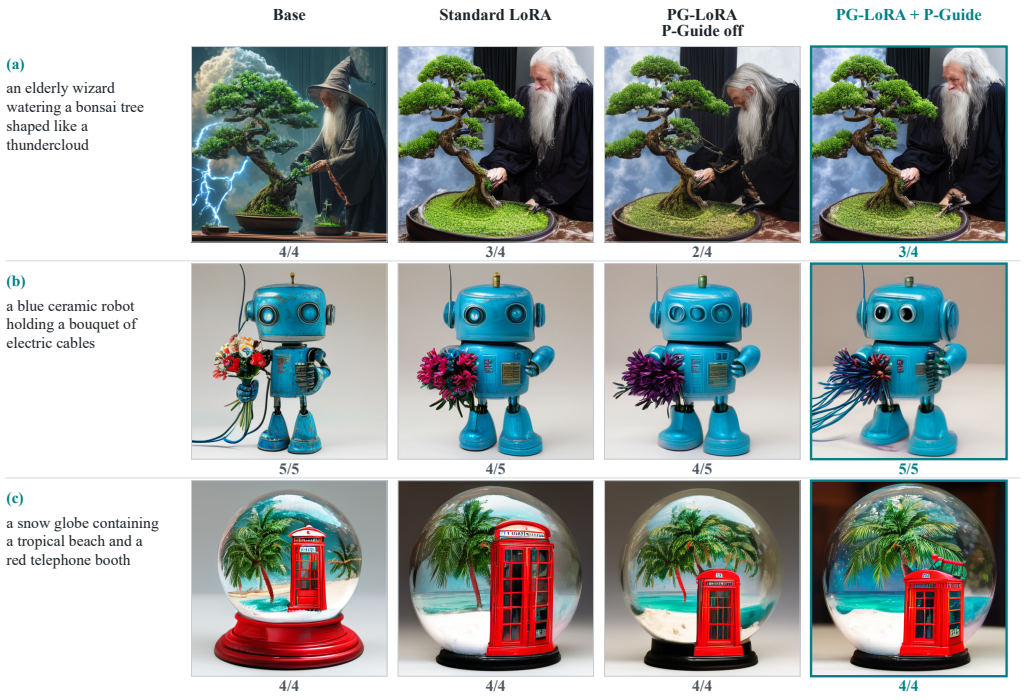


Figure 7: Qualitative comparison on three held-out compositional prompts. All methods use CFG 4.5 and 20 Euler steps; samples within each row share the prompt and seed, and P-Guide uses source scale 1.5. The labels below the images give satisfied atomic checks, such as 4/4. The pretrained Base column provides visual context; Table 13 uses Standard LoRA as the quantitative reference.

E.4 SCOPE AND LIMITATIONS

This pilot evaluates feasibility and prompt-level trade-offs using 20,000 COCO pairs, one 600M-parameter SANA backbone, one checkpoint per method, 30 prompts, one seed per prompt, and one BLIP-based evaluator. Multiple seeds, official benchmark pipelines, distributional quality and diversity, human preference, and source-only text guidance remain unevaluated.

F Additional Qualitative Results

These examples visualize how source construction and guidance strength affect trajectories on MNIST, CIFAR-10, and ImageNet-1k.

F.1 Gaussian Source Extrapolations on MNIST

We sample from the normalizable Gaussian $p_{\text{cfg}}(x)$ and the P-Guide source, then propagate both through the same adapted MNIST flow. Figure 8 compares the resulting samples under the same initial noise and guidance scale ($w = 1.5$).

At $w = 1.5$, the log-density source produces more distorted digits, while P-Guide preserves clearer digit structure. The different source distributions thus produce different sample structures under the same learned flow.

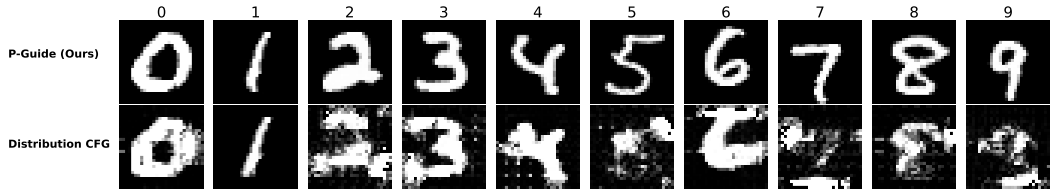
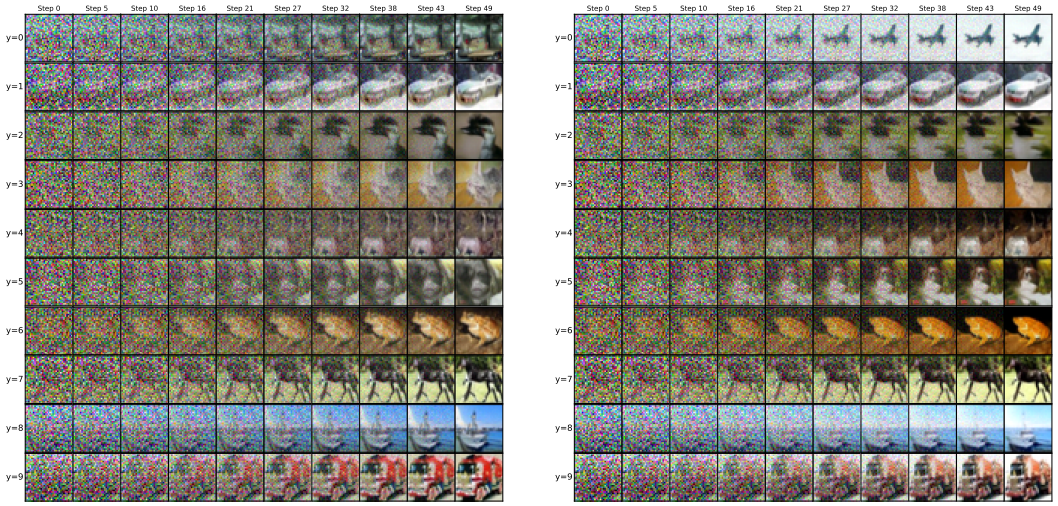


Figure 8: Gaussian source extrapolations on MNIST at $w = 1.5$. P-Guide preserves clearer digit structure than log-density extrapolation under the same adapted flow.

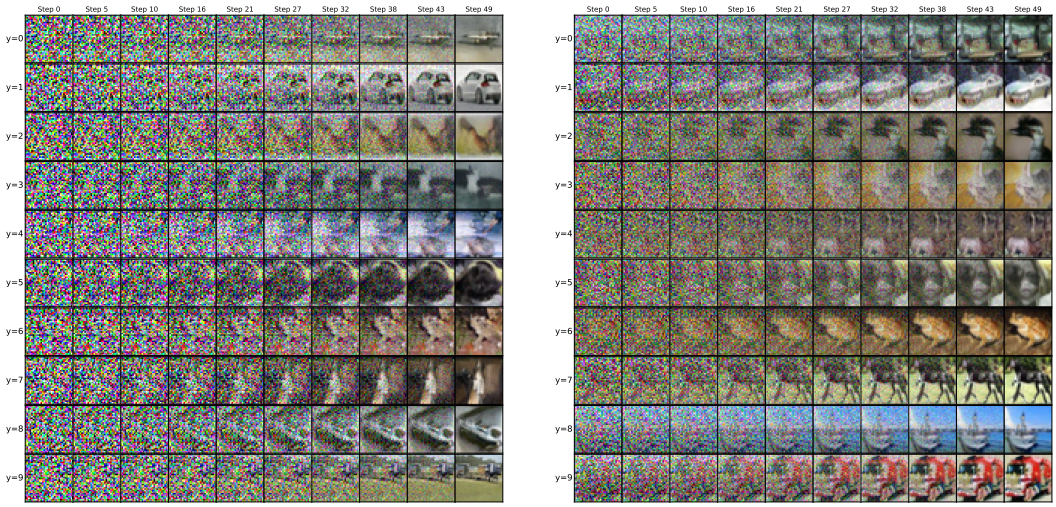
F.2 CIFAR-10 Trajectories



(a) Trajectory from z_{cfg} to x_1 with $w = 1.0$

(b) Trajectory from z_{cfg} to x_1 with $w = 1.2$

Figure 9: **Visual comparison of P-Guide generation trajectories on CIFAR-10.** Each row shows the evolution of a sample from its initial latent state ($t = 0$) to the final generated image ($t = 1$). Comparing $w = 1.0$ and $w = 1.2$ illustrates how changing the initial mean affects intermediate and final images under the conditional flow. Table 1 separately quantifies quality and classification accuracy across guidance scales.

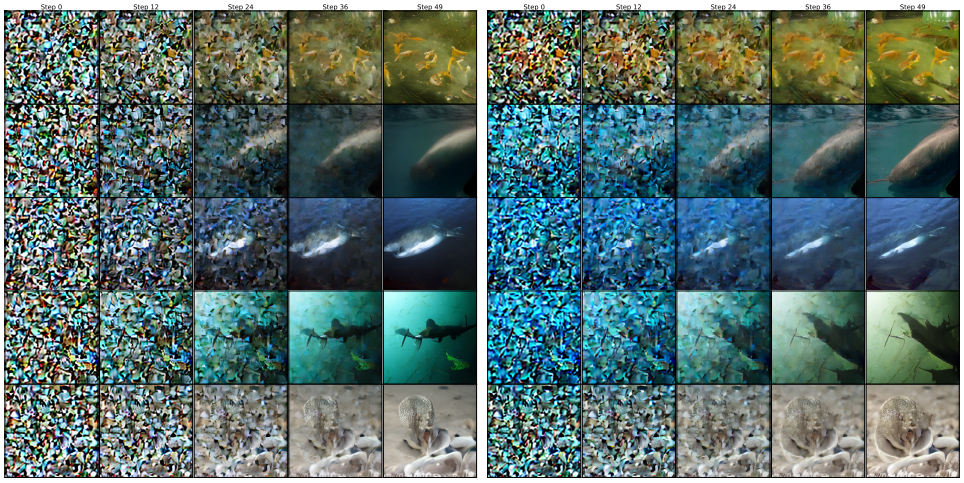


(a) Standard CFM with Gaussian prior ($w = 1.0$)

(b) P-Guide with class-conditioned prior ($w = 1.0$)

Figure 10: **Trajectory comparison under different initial latent distributions on CIFAR-10.** Each row visualizes the evolution of a sample from its initial state ($t = 0$) to the final generated image ($t = 1$). Standard CFM starts from a class-independent Gaussian source; P-Guide starts from a learned class-conditioned source. Each conditional backbone is trained for its corresponding source distribution. At $w = 1$, both models use their conditional predictions directly, showing the trajectories produced by each source–backbone pair.

F.3 IMAGENET TRAJECTORIES



(a) Standard CFM with Gaussian prior ($w = 1.0$) (b) P-Guide with class-conditioned prior ($w = 1.0$)

Figure 11: **Trajectory comparison under different initial latent distributions on ImageNet-1k.** Each row shows the evolution of a sample from its initial state ($t = 0$) to the final generated image ($t = 1$). Standard CFM uses a class-independent Gaussian source, while P-Guide uses a learned class-conditioned source and an adapted backbone. At $w = 1$, P-Guide samples that conditional source without additional mean extrapolation.