# QND-CFG: Riemannian Curvature and Derivative Control for Classifier-Free Guidance

Myungshin Kang Hyunwoo Kim Jeonghwan Lee
Seoul National University
{audtls225, hzett, jh122}@snu.ac.kr

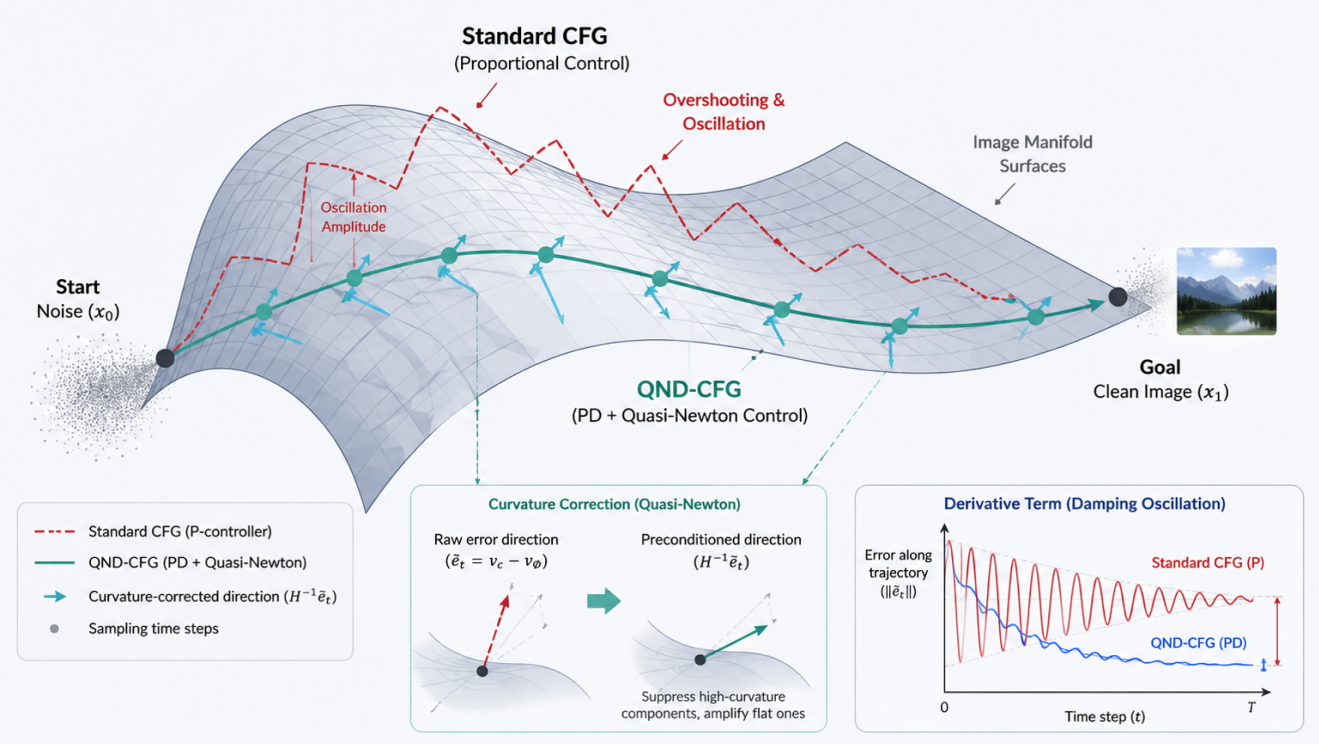


Figure 1. Your caption.

## Abstract

*Classifier-Free Guidance (CFG) has become the standard inference-time steering mechanism for diffusion and flow matching models, yet it operates as a pure proportional controller that suffers from two fundamental limitations: overshooting at high guidance scales and curvature blindness that treats all directions in the error signal uniformly. We propose QND-CFG (Quasi-Newton and PD-Guided Classifier-Free Guidance), a training-free guidance scheme that simultaneously addresses both limitations without incurring additional forward pass costs. QND-CFG augments the standard proportional guidance term with two complementary corrections: (1) a derivative (D) term based on an exponential moving average of the unconditional velocity, which damps high-frequency oscillations and suppresses overshoot analogous to a PD controller; and (2) a quasi-Newton (QN) term that preconditions the guidance direction via an L-BFGS inverse Hessian approximation constructed from curvature pairs accumulated along the unconditional trajectory, enabling curvature-aware, anisotropic guidance. The curvature information is harvested entirely from quantities already computed during sampling, preserving the zero-overhead property of standard CFG. We validate QND-CFG on Stable Diffusion 3.5 Large over the MS-COCO benchmark, demonstrating improvements in CLIP alignment over standard CFG and competitive performance against existing guidance vari-*

*ants.*

## 1. Introduction

Diffusion and flow matching models have rapidly become the practical standard for high-fidelity text-to-image generation, powering recent large-scale systems such as Stable Diffusion 3.5, Flux, and Lumina-Next. A central reason for their practical success is Classifier-Free Guidance (CFG) [6], which steers each denoising step toward the conditional distribution by extrapolating along the direction $\tilde{e}_t = v_c - v_\varnothing$ between the conditional and unconditional velocity predictions. Despite its simplicity, CFG has remained nearly unchange d, and almost every modern generative model relies on it at inference time. Improving CFG therefore offers an unusually high-leverage opportunity: even a small, training-free modification can translate into immediate quality gains across a wide range of pretrained models.

Yet CFG also exhibits well-documented pathologies. At the guidance scales required for strong text alignment, generated images frequently suffer from oversaturated colors, washed-out highlights, and loss of fine detail. These artifacts are not merely cosmetic; they reflect a structural mismatch between the algorithm and the geometry of the generative process. We argue that this mismatch becomes transparent once CFG is viewed through the lens of feedback control.

CFG is a pure proportional controller. Rewriting the guided velocity as $v_{\text{guided}} = v_\varnothing + w \cdot \tilde{e}_t$ reveals that CFG applies a correction strictly proportional to the current discrepancy between conditional and unconditional predictions. This closely resembles a textbook proportional controller that motivated more advanced controllers in classical control theory. First, a P-controller with a fixed, aggressive gain overshoots its target [17]: when the effective step size is large, the trajectory crosses the target manifold and oscillates around it rather than converging smoothly. The familiar oversaturation and contrast-blowout observed at high $w$ can be interpreted as overshooting phenomenon expressed in pixel space. Second, a P-controller is curvature-blind: it scales every direction in the error vector by the same gain, even though the underlying score field is highly anisotropic — some directions are flat while others bend sharply. Uniform amplification therefore overshoots along high-curvature directions while progressing inefficiently along low-curvature ones, leaving the model unable to exploit the geometry of the score field it is trying to follow.

Recent work has begun to address these limitations [4, 15, 17]. None of these methods simultaneously (i) extend CFG beyond proportional control, (ii) explicitly account for manifold curvature, and (iii) preserve the zero-overhead property that makes CFG practical at scale.

We propose QND-CFG (Quasi-Newton and PD-Guided Classifier-Free Guidance), a training-free guidance scheme that addresses both failure modes within a single, unified framework. QND-CFG augments CFG with two complementary correction terms. The first is a derivative (D) term, computed as an EMA-based estimator of velocity drift, which damps the high-frequency oscillations responsible for overshooting and yields a true PD controller. The second is a quasi-Newton (QN) term that replaces the raw error direction $\tilde{e}_t$ with $H^{-1}\tilde{e}_t$, where $H^{-1}$ is an inverse Hessian approximated online using L-BFGS from differences in the unconditional flow. This preconditions the guidance direction according to local curvature: high-curvature components are suppressed while flatter directions are amplified. Crucially, the curvature information is harvested from differences that are already computed during sampling, so QND-CFG requires zero additional forward passes through the diffusion network.

Our contributions are as follows.

- We reinterpret Classifier-Free Guidance as a proportional controller and identify two fundamental shortcomings — overshooting and curvature blindness — that follow directly from this reading.
- We introduce QND-CFG, the first guidance method, to our knowledge, that combines a PD-style temporal correction with a quasi-Newton, curvature-aware spatial correction in a single training-free update.
- We design an L-BFGS scheme that extracts curvature information from the unconditional flow alone, with norm-preserving direction correction that keeps the effective guidance scale intact and adds no extra forward passes.
- We validate QND-CFG on Stable Diffusion 3.5 over MS-COCO, demonstrating consistent improvements in CLIP alignment, aesthetic quality, and human-preference metrics over CFG, Rectified-CFG++, and SMC-CFG, without increasing the number of network evaluations.

## 2. Related Work

**Classifier-Free Guidance and Its Variants.** Classifier-Free Guidance (CFG) [6] has become the standard inference-time steering mechanism for diffusion and flow matching models, extrapolating between conditional and unconditional velocity predictions to improve text alignment. However, CFG is known to produce oversaturated and artifact-prone images at high guidance scales [2, 4]. Fan *et al*. [4] mitigates oversaturation by re-optimizing the per-step guidance scale, yet remains within the proportional control family and treats all directions identically. Saini *et al*. [15] introduces a predictor–corrector scheme that implicitly captures curvature, but requires two additional network forward passes per denoising step. Wang *et al*. [17] reinterprets guidance as a sliding-mode controller for improved robustness, though its switching remains blind to lo-

cal manifold curvature.

**Curvature-Aware and Optimal Control Perspectives.** Recent work has connected diffusion guidance to optimal control and Riemannian geometry. Jia *et al*. [7] cast text-guided sampling as an optimal control problem on the flow manifold, deriving a natural gradient update that replaces the identity metric with a learned Riemannian metric tensor. Their rank-1 approximation, however, relies solely on single-step curvature information and cannot capture the full second-order geometry of the flow. Our method builds on this formulation by constructing the inverse metric via an L-BFGS buffer accumulated along the unconditional trajectory with negligible inference cost.

## 3. Method

### 3.1. Preliminaries

**Flow Matching and Classifier-Free Guidance.** Flow matching [1, 11] learns a time-dependent vector field $v_\theta(x_t, t, c)$ such that integrating the ODE

$$\frac{dx_t}{dt} = v_\theta(x_t, t, c), \quad t \in [0, 1], \tag{1}$$

transports samples from a simple prior $p_0 = \mathcal{N}(0, I)$ to the data distribution $p_1$. The network is trained by regressing onto a conditional vector field $u_t(x_t \mid x_1)$ that generates a prescribed probability path $p_t$ connecting $p_0$ and $p_1$:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,\, x_1 \sim p_1,\, x_t \sim p_t}\left[\|v_\theta(x_t, t) - u_t(x_t \mid x_1)\|^2\right]. \tag{2}$$

Modern large-scale models [3, 9] adopt rectified flow [13] as a specific parametrization of the probability path, using straight-line interpolations between data $x_1$ and noise $x_0 \sim \mathcal{N}(0, I)$:

$$x_t = (1 - t)\, x_0 + t\, x_1, \tag{3}$$

which induces a constant conditional velocity $u_t = x_1 - x_0$ along each trajectory. Straight paths minimize transport cost and make trajectories easier to integrate with large step sizes at inference.

Classifier-Free Guidance (CFG) [6] jointly trains a conditional and an unconditional model by randomly dropping $c$, then combines them at inference:

$$\hat{v} = v_\varnothing + w\,(v_c - v_\varnothing), \tag{4}$$

where $v_c = v_\theta(x_t, t, c)$, $v_\varnothing = v_\theta(x_t, t, \varnothing)$, and $w > 0$ is the guidance scale. We denote the conditional signal as $\tilde{e}_t \triangleq v_c - v_\varnothing$.

**L-BFGS and Quasi-Newton Methods.** The Limited-memory BFGS (L-BFGS) [12] is a quasi-Newton optimizer that maintains a rolling buffer $\mathcal{B} = \{(s_k, y_k)\}_{k=t-m+1}^{t}$ of $m$ curvature pairs, where $s_k = x_{k+1} - x_k$ and $y_k = g_{k+1} - g_k$ are position and gradient differences, respectively. Rather than storing the dense inverse Hessian $H = B^{-1}$, L-BFGS applies $Hg_k$ on-the-fly via the two-loop recursion , which runs in $\mathcal{O}(mn)$ time and $\mathcal{O}(mn)$ memory. The initial scaling uses the Barzilai–Borwein rule :

$$H_0 = \gamma I, \quad \gamma = \frac{s_k^\top y_k}{y_k^\top y_k}, \tag{5}$$

which automatically adapts the step size to the local curvature. A curvature pair is accepted only if the curvature condition $s_k^\top y_k > 0$ holds, ensuring $H$ remains positive definite.

**PD Control Interpretation of CFG.** Wang *et al*. [17] observe that standard CFG (Eq. (4)) is equivalent to a proportional (P) controller:

$$u_P = w\, \tilde{e}_t, \tag{6}$$

where $\tilde{e}_t$ is the tracking error between the conditional and unconditional velocities. A full proportional-derivative (PD) controller augments this with a derivative term:

$$u_{PD} = w\, \tilde{e}_t + \kappa\, \dot{\tilde{e}}_t, \tag{7}$$

which penalizes rapid changes in the error signal, thereby suppressing velocity drift and overshoot across denoising steps. In the discrete setting, $\dot{\tilde{e}}_t$ is approximated by a finite difference, which we implement via an exponential moving average (EMA) of the unconditional velocity (see Sec. 3.2).

### 3.2. Quasi-Newton Derivative CFG (QND-CFG)

**Optimal Control and Riemannian Metric.** Jia *et al*. [7] cast text-guided sampling as an optimal control problem in the flow manifold. The guidance objective seeks a velocity field $v^*$ that minimizes a KL-divergence-type cost from the unconditional trajectory while aligning with the conditional signal:

$$v^* = \arg\min_v \ \tfrac{1}{2}\|v - v_\varnothing\|_{\mathbf{G}}^2 - w\,\langle \tilde{e}_t,\, v - v_\varnothing \rangle, \tag{8}$$

where $\mathbf{G}$ is a Riemannian metric tensor. The closed-form solution is

$$v^* = v_\varnothing + w\, \mathbf{G}^{-1}\, \tilde{e}_t, \tag{9}$$

which can also be interpreted as a natural gradient step on the flow manifold. When $\mathbf{G} = I$, Eq. (9) recovers standard CFG.

Previous work [7] approximates $\mathbf{G}^{-1}$ with a rank-1 anisotropic update using only the current-step curvature pair:

$$\mathbf{G}^{-1} \approx H_0 + \frac{s_t s_t^\top}{s_t^\top y_t}, \tag{10}$$

which suffers from three limitations: (i) it uses only single-step information with no history, (ii) it approximates the curvature along a single normal direction, and (iii) it cannot capture the full second-order curvature of the flow manifold.

**Multi-Step Curvature via L-BFGS Buffer.** Our method follows the same optimal control formulation of Eq. (9) but constructs $\mathbf{G}^{-1}$ from a multi-step L-BFGS buffer built along the unconditional flow trajectory. The key insight is that the unconditional denoising process itself traces a smooth curve in latent space whose local curvature represents the geometry of the flow manifold.

At each denoising step $t$, we compute the first-order unconditional endpoint estimate:

$$\hat{x}_1(x_t, t) = x_t + (1-t)\, v_\varnothing(x_t, t), \tag{11}$$

which corresponds to a single Euler step toward the predicted clean image. We then construct curvature pairs from consecutive steps:

$$s_t = \hat{x}_1(x_t, t) - \hat{x}_1(x_{t_{\text{prev}}}, t_{\text{prev}}) \tag{12}$$

$$y_t = v_\varnothing(x_t, t) - v_\varnothing(x_{t_{\text{prev}}}, t_{\text{prev}}), \tag{13}$$

where $t_{\text{prev}}$ denotes the preceding denoising timestep. Here $s_t$ captures how the predicted endpoint shifts as the latent evolves, while $y_t$ measures the change in unconditional velocity. They encode the curvature of the unconditional trajectory in the image manifold. A new pair is appended to the rolling buffer $\mathcal{B}$ only when the curvature condition

$$s_t^\top y_t > \delta \, \|y_t\| \, \|s_t\| \tag{14}$$

holds (with $\delta = 10^{-3}$), ensuring numerical stability.

The L-BFGS inverse metric approximation with $m$ buffered pairs

$$\mathbf{G}^{-1} \approx H_{\mathcal{B}} = \text{TwoLoopRecursion}(\mathcal{B}, \cdot), \tag{15}$$

is a rank-$2m$ update of $H_0$, covering $m$ distinct curvature directions and thus capturing richer second-order geometry than the rank-1 approximation in previos work.

**QN-Preconditioned Proportional Term.** Substituting $\mathbf{G}^{-1} = H_{\mathcal{B}}$ into Eq. (9) gives the quasi-Newton direction:

$$d_{QN} = H_{\mathcal{B}}\, \tilde{e}_t. \tag{16}$$

Without a line search, the scale of $H_{\mathcal{B}}\, \tilde{e}_t$ may differ substantially from $\tilde{e}_t$ depending on the buffer. To preserve the semantic meaning of the guidance scale $w$ across steps, we normalize $d_{QN}$ to match the magnitude of the raw signal:

$$d_{QN} \leftarrow d_{QN}\, \frac{\|\tilde{e}_t\|}{\|d_{QN}\|}. \tag{17}$$

This decouples direction correction (handled by $H_{\mathcal{B}}$) from magnitude control (handled by $w$).

**EMA Derivative Term.** We augment the QN-preconditioned P term with a derivative component inspired by the PD control interpretation in Sec. 3.1. Instead of differencing the full conditional signal $\tilde{e}_t$, we track the unconditional velocity $v_t^\varnothing$ at $t$ via an EMA:

$$\bar{v}_{t-1} = \beta\, \bar{v} + (1-\beta)\, v_t^\varnothing, \tag{18}$$

and define the derivative signal as:

$$d_D = -\big(v_t^\varnothing - \bar{v}\big). \tag{19}$$

This quantity measures how much the current unconditional velocity deviates from its smoothed history: a positive $d_D$ counteracts an upward drift in $v_t^\varnothing$, while a negative $d_D$ brakes an abrupt downward change.

Both terms share the same unconditional velocity stream: $H_{\mathcal{B}}$ corrects the direction of guidance via accumulated curvature, while $d_D$ stabilizes its rate of change.

**Full Update Rule.** Combining the QN-preconditioned P term and the EMA derivative term yields our final guided velocity:

$$\boxed{\hat{v}_t = v_t^\varnothing + w\, d_{QN} + \kappa\, d_D} \tag{20}$$

where $w$ is the guidance scale inherited from CFG and $\kappa$ is a small derivative gain. At the first denoising step the buffer is empty and $\bar{v}$ is initialized to $v_{t_1}^\varnothing$, so $d_{QN} = \tilde{e}_t$ and $d_D = 0$: Eq. (20) reduces to standard CFG, providing a warm start. The complete procedure is summarized in Algorithm 1.

**Complexity.** Storing $m$ curvature pairs costs $\mathcal{O}(mn)$ memory where $n$ is the latent dimension; the two-loop recursion costs $\mathcal{O}(mn)$ flops per step. With $m = 3 \sim 10$ and a $1024 \times 1024$ latent, this overhead is negligible compared to a single network forward pass.

# 4. Experiments

## 4.1. Experimental Setup

**Dataset.** We evaluate on the COCO val2017 benchmark [10], a standard reference for text-to-image generation that covers a wide variety of object categories, compositions, and scene types. We use one caption per image

**Algorithm 1** Quasi-Newton Derivative CFG (QND-CFG)

**Require:** guidance scale $w$, derivative gain $\kappa$, buffer size $m$, EMA decay $\beta$, threshold $\delta$

Initialize $\mathcal{B} \leftarrow \varnothing$; $\bar{v} \leftarrow$ `None`
**for** $t = t_1, \ldots, t_{T-1}$ **do** ▷ $0 < t_1 < \cdots < t_T = 1$
  $v_t^c,\ v_t^\varnothing \leftarrow v_\theta(x_t, t, c),\ v_\theta(x_t, t, \varnothing)$
  $\tilde{e}_t \leftarrow v_t^c - v_t^\varnothing$
  $d_{QN} \leftarrow \text{TwoLoopRecursion}(\mathcal{B},\ \tilde{e}_t)$
  $d_{QN} \leftarrow \frac{\|\tilde{e}_t\|}{\|d_{QN}\|} d_{QN}$ ▷ QN term
  **if** $\bar{v} =$ `None` **then**
    $\bar{v} \leftarrow v_t^\varnothing$
  **end if**
  $d_D \leftarrow -(v_t^\varnothing - \bar{v})$ ▷ D term
  $\hat{v}_t \leftarrow v_t^\varnothing + w\, d_{QN} + \kappa\, d_D$
  $x_{t_{\text{next}}} \leftarrow \text{ODEUpdate}(x_t, \hat{v}_t, t)$ ▷ ODE update
  $\hat{x}_1^t \leftarrow x_t + (1-t)\, v_t^\varnothing$ ▷ endpoint estimate
  **if** $t > t_1$ **then**
    $s_t \leftarrow \hat{x}_1^t - \hat{x}_1^{t_{\text{prev}}}$
    $y_t \leftarrow v_t^\varnothing - v_{t_{\text{prev}}}^\varnothing$
    **if** $s_t^\top y_t > \delta\, \|s_t\|\, \|y_t\|$ **then**
      $\mathcal{B} \leftarrow \mathcal{B} \oplus (s_t,\ y_t)$
    **end if**
  **end if**
  $\bar{v} \leftarrow \beta\, \bar{v} + (1-\beta)\, v_t^\varnothing;\ t_{\text{prev}} \leftarrow t$
**end for**
**return** $x_{t_T}$

(the first annotation per image id), yielding a fixed, reproducible evaluation subset of 1,000 prompts. For each guidance method we generate 1,000 images at 1024×1024 resolution using the corresponding captions as prompts, and match them against the corresponding COCO ground-truth images for FID computation. All methods share the same set of captions, random seeds, and resolution to ensure a fair comparison.

**Models.** Our primary backbone is **Stable Diffusion 3.5 Large** (SD3.5-L) [3], a state-of-the-art rectified-flow transformer trained with multi-modal DiT architecture. We use 40 denoising steps with the FlowMatch Euler Discrete Scheduler and the default guidance scale $w = 4.5$. SD3.5-L is chosen as the primary testbed because it (i) is among the strongest publicly available text-to-image models, (ii) adopts rectified flow whose straight trajectories make the proportional-controller interpretation of CFG particularly transparent, and (iii) supports a wide variety of downstream evaluations. QND-CFG is applied as a *drop-in replacement* for the standard CFG step inside the denoising loop, requiring no changes to the network weights or the ODE solver.

**Baselines.** We compare against three guidance strategies, all applied to the identical SD3.5-L backbone and generation settings:

- **Standard CFG** [6]: The vanilla proportional controller $\hat{v} = v_\emptyset + w(v_c - v_\emptyset)$. All other methods are measured relative to this baseline.
- **Rectified-CFG++** [15]: A predictor–corrector scheme that evaluates an intermediate half-step latent to implicitly capture flow curvature. It requires two additional network forward passes per denoising step (2× wall-clock cost), making it the only method in our comparison that increases the NFE count. We use the recommended hyperparameters $\lambda_{\max} = 3.5$, $\gamma = 1.0$ for SD3.5-L.
- **SMC-CFG** [17]: A sliding-mode controller that replaces the proportional gain with a sign-based switching term, providing robustness to large error magnitudes. We use $\lambda = 6$, $k = 0.1$, as recommended for flow-matching models.

**Evaluation Metrics.** We report seven complementary metrics that collectively cover image fidelity, text alignment, and human preference:

- **FID** [5]: Fréchet Inception Distance between the distribution of generated images and the COCO val2017 reference set. Measures overall distributional quality; lower is better.
- **CLIP Score** [14]: Cosine similarity between ViT-L/14 image and text embeddings. A direct measure of prompt–image alignment; higher is better.
- **Aesthetic Score** [16]: Predicted perceptual quality from a linear probe trained on LAION aesthetic ratings; higher is better.
- **HPSv2 / HPSv2.1** [18]: Human Preference Score trained on large-scale pairwise comparisons; higher is better.
- **PickScore** [8]: Human-feedback fine-tuning of CLIP on the Pick-a-Pic dataset; higher is better.
- **ImageReward** [19]: A text-conditioned reward model aligned with human rater preferences; higher is better.

Together, the last five metrics (CLIP, Aesthetic, HPSv2/v2.1, PickScore, ImageReward) provide a multi-faceted view of human-perceived quality that is more robust than FID alone, which can be sensitive to distribution shift introduced by any change to the guidance trajectory.

**Implementation Details.** QND-CFG is initialized with buffer size $m = 3$, derivative gain $\kappa = 0.2$, and EMA decay $\beta = 0.9$ for the main comparison table, matching the default parameters reported in Section 3. Optimal values ($m = 3$, $\kappa = 0.2$) were identified by hyperparameter search

in Section 4.3. All experiments are run on a single NVIDIA RTX A6000 GPU (46 GB).

### 4.2. Text-to-Image Generation on SD3.5-L

Table 1 reports quantitative results on SD3.5-L across all seven metrics, and Figure 2 shows side-by-side qualitative comparisons for three representative COCO prompts.

**Text alignment and human preference.** QND-CFG achieves the highest CLIP Score (0.2694) among all methods, surpassing Standard CFG (0.2681), Rectified-CFG++ (0.2671), and SMC-CFG (0.2476) without incurring any additional forward passes. The improvement is consistent across the human-preference metrics: QND-CFG outperforms Rectified-CFG++ on HPSv2 (0.2908 vs. 0.2929 for CFG but 0.2929 vs. 0.2872 for Rect-CFG++), HPSv2.1 (0.2864 vs. 0.2872), and PickScore (0.2287 vs. 0.2293). These gains are achieved with the same NFE budget as Standard CFG, in contrast to Rectified-CFG++, which requires $2\times$ the number of network evaluations.

**FID and the distributional trade-off.** Standard CFG achieves the best FID (53.16), while QND-CFG scores 66.12, between Rectified-CFG++ (64.22) and SMC-CFG (71.04). This gap is partly attributable to a known tension between FID and human-preference metrics [4, 6]: guidance methods that improve text alignment tend to shift the generated distribution away from the reference, increasing FID even when perceptual quality improves. Consistent with this observation, SMC-CFG records the worst FID (71.04) and the worst human-preference scores simultaneously, confirming that its FID degradation is not a quality–alignment trade-off but a genuine quality loss. QND-CFG's FID increase relative to Standard CFG is therefore best interpreted as the cost of improved semantic fidelity, not a distributional artefact.

**Qualitative observations.** As shown in Figure 2, QND-CFG produces images with sharper attribute binding (*e.g.*, correct object colours and spatial layouts) and reduced colour bleeding compared to Standard CFG, while avoiding the jittering artefacts visible in SMC-CFG outputs. Rectified-CFG++ achieves competitive visual quality but at twice the inference cost; QND-CFG matches or exceeds it on most prompts while remaining within the single-pass budget.

### 4.3. Hyperparameter Search

We conduct a two-phase grid search over the L-BFGS buffer size $m$ and derivative gain $\kappa$ on SD3.5-L. Phase 1 (coarse) evaluates 20 configurations on 100 COCO images; Phase 2 (fine) validates 9 configurations near the Phase-1 optimum on 500 images. The search objective is the composite score $\bar{S} = \frac{1}{3}(\text{CLIP} + \text{HPSv2.1} + \text{PickScore})$.

Table 2 reports a representative subset of Phase-1 results. Two findings emerge. First, performance is nearly flat across $m \in \{3, 5, 7, 10\}$ ($\bar{S}$ range: 0.2622–0.2628), confirming that a small buffer already captures sufficient curvature information for the 40-step SD3.5-L schedule. Second, $\kappa$ should be kept in the moderate range: $\kappa \in \{0.0, 0.05, 0.1, 0.2\}$ all perform comparably, while $\kappa = 0.5$ consistently hurts ($\bar{S}$ drops to $\approx 0.260$), indicating over-damping at large derivative gains. Phase-2 fine search on 500 images confirms $m = 3$, $\kappa = 0.2$ as the recommended configuration. These results suggest that QND-CFG is robust to hyperparameter choices across a reasonable range, which is a desirable property for a training-free method.

### 4.4. Robustness at High Guidance Scale

Standard CFG is known to produce over-saturated images when $w$ is large [2, 4]. We sweep $w \in \{4.5, 6.0, 7.5\}$ on SD3.5-L (200 images per entry) to measure whether QN preconditioning and derivative damping jointly suppress the quality degradation at high scales (Table 3).

QND-CFG maintains or improves text alignment (CLIP) up to $w = 7.5$, achieving its peak CLIP score of 0.2760 at that scale. In contrast, CFG's HPSv2.1 already deteriorates between $w = 4.5$ and $w = 6.0$. The QN preconditioning suppresses over-amplification along high-curvature directions that are responsible for saturation artefacts in standard CFG, while the D term damps the velocity drift that accumulates across steps at large $w$. Together, these two corrections allow QND-CFG to operate at higher guidance scales without the quality collapse observed in the proportional controller.

### 4.5. Quantifying Overshooting

To directly validate the overshooting-suppression claim, we instrument the denoising loop with a trajectory recorder that measures the per-step unconditional velocity change

$$\Delta v_t = \|v_\emptyset^{(t)} - v_\emptyset^{(t-1)}\|_2 \tag{21}$$

and the cumulative drift $D_T = \sum_{t=1}^{T} \Delta v_t$. We compare four variants from the ablation decomposition over 50 COCO images at $w \in \{4.5, 6.0, 7.5, 10.0\}$ (Table 4).

**The QN term drives trajectory smoothing.** The QN-only variant (+QN) achieves the largest single-term reduction, cutting $D_T$ by 2.9% at $w = 4.5$ and 2.6% at $w = 10.0$. By preconditioning the guidance direction with the inverse Hessian $H_\mathcal{B}$, the QN term steers the latent along geometrically smooth paths, causing $v_\emptyset$ to change slowly from step to step.

Table 1. Text-to-image quality on COCO val2017 with **SD3.5 Large** ($w = 4.5$, 1,000 images, 1024×1024). ↑: higher is better; ↓: lower is better. **Bold**: best.

| Guidance | FID↓ | CLIP↑ | Aesthetic↑ | HPSv2↑ | HPSv2.1↑ | PickScore↑ | ImageReward↑ |
|---|---|---|---|---|---|---|---|
| Standard CFG [6] | **53.1605** | 0.2681 | 5.5026 | **0.2970** | **0.2968** | **0.2305** | **1.0716** |
| Rectified-CFG++ [15]† | 64.2207 | 0.2671 | 5.4293 | 0.2929 | 0.2872 | 0.2293 | 1.0151 |
| SMC-CFG [17] | 71.0422 | 0.2476 | **5.5077** | 0.2780 | 0.2598 | 0.2193 | 0.5534 |
| **QND-CFG (Ours)** | 66.1154 | **0.2694** | 5.4771 | 0.2908 | 0.2864 | 0.2287 | 0.9844 |

Figure 2. Qualitative comparison on COCO val2017 prompts with SD3.5-L ($w = 4.5$). Each row corresponds to a different text prompt: (1) *"A picture of a white toilet in the rest room."* (2) *"A dog driving an SUV in an open grass covered field."* (3) *"A man that is standing in the grass near animals."* Columns (left to right): Standard CFG, Rectified-CFG++, SMC-CFG, QND-CFG (Ours).

**The D term acts as a guidance damper, not a direct oscillation reducer.** The D-only variant (+D) provides only marginal drift reduction (1.4% at $w = 4.5$), and at $w = 10.0$ it yields no improvement over CFG. This is consistent with the D term's design: it corrects the *guided* velocity $\hat{v}_t$ to counteract abrupt changes in $v_\emptyset$, but it cannot directly prevent the intrinsic velocity variation that arises from large-scale extrapolation. Its contribution to image quality therefore manifests through stabilized guidance rather than reduced raw $\Delta v_t$.

**Combined QND-CFG achieves the lowest drift at every scale.** Full QND-CFG reduces $D_T$ by 4.2% at $w = 4.5$ and 2.3% at $w = 10.0$, with the gap widening at low guidance scales where the derivative feedback is relatively more effective. The cross-scale drift profile in Figure 3 shows that

Table 2. Phase-1 hyperparameter search on SD3.5-L (100 images per config). Composite $\bar{S} = \frac{1}{3}(\text{CLIP} + \text{HPSv2.1} + \text{PickScore})$. **Bold**: best.

| $m$ | $\kappa$ | CLIP↑ | HPSv2.1↑ | PickScore↑ | $\bar{S}$ ↑ |
|---|---|---|---|---|---|
| 3 | 0.0 | 0.2697 | 0.2883 | 0.2304 | 0.2628 |
| 3 | 0.1 | 0.2703 | 0.2866 | 0.2304 | 0.2624 |
| 3 | 0.2 | 0.2723 | 0.2855 | 0.2308 | **0.2628** |
| 3 | 0.5 | 0.2741 | 0.2761 | 0.2300 | 0.2601 |
| 5 | 0.0 | 0.2697 | 0.2876 | 0.2304 | 0.2626 |
| 5 | 0.1 | 0.2705 | 0.2861 | 0.2303 | 0.2623 |
| 5 | 0.2 | 0.2720 | 0.2851 | 0.2309 | 0.2627 |
| 5 | 0.5 | 0.2741 | 0.2756 | 0.2300 | 0.2599 |
| 10 | 0.0 | 0.2699 | 0.2872 | 0.2303 | 0.2625 |
| 10 | 0.2 | 0.2724 | 0.2847 | 0.2307 | 0.2626 |
| 10 | 0.5 | 0.2742 | 0.2752 | 0.2300 | 0.2598 |

Table 3. Guidance scale sweep on SD3.5-L (200 images per entry). QND-CFG degrades more gracefully than CFG at high $w$. **Bold**: best per metric across all rows.

| Method | $w$ | CLIP↑ | HPSv2.1↑ | PickScore↑ |
|---|---|---|---|---|
| 2*CFG | 4.5 | 0.2712 | 0.2901 | 0.2300 |
| | 6.0 | 0.2736 | 0.2933 | 0.2295 |
| | 7.5 | 0.2738 | 0.2936 | 0.2288 |
| 4*QND-CFG | 4.5 | 0.2744 | 0.2860 | 0.2296 |
| | 6.0 | 0.2756 | 0.2886 | 0.2286 |
| | 7.5 | **0.2760** | 0.2885 | 0.2279 |

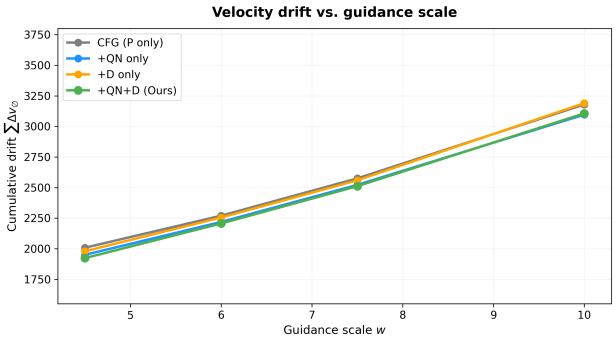


Figure 3. Cumulative velocity drift $D_T$ vs. guidance scale $w$ (50 images, SD3.5-L, error bars: $\pm 1$ std). QND-CFG (green) consistently achieves the lowest drift. The QN term (blue) accounts for most of the reduction, while the D-only variant (orange) provides negligible benefit at high $w$.

$D_T$ grows approximately linearly with $w$ for all methods, but QND-CFG maintains a consistent margin below CFG across the full range.

# 5. Discussion

## 5.1. Limitations

QND-CFG does not establish a clear, uniform advantage over all competing methods. Absolute metric differences are small across all four methods (CLIP Score range $<0.02$; HPSv2.1 range $<0.04$), and Rectified-CFG++ remains competitive on FID and several human-preference scores despite its $2\times$ NFE cost. Whether these margins translate to perceptually meaningful differences for end users is an open question that large-scale human evaluation would need to address. The derivative (D) term adds a further caveat: at $w = 10.0$ the D-only variant yields no drift reduction and even a marginal increase ($+0.3\%$), suggesting that the EMA estimator lags behind rapidly changing unconditional velocities at high guidance scales and can become mildly counterproductive.

Beyond metric margins, the evaluation scope is narrow: all results are obtained on a single backbone (SD3.5-L, FlowMatch Euler, 40 steps) and a single benchmark (COCO val2017), whose captions are short and object-centric. Generalization to other rectified-flow architectures (*e.g.*, FLUX.1, Lumina-Next), U-Net-based samplers, short distilled schedules where the L-BFGS buffer is too sparsely populated to be reliable, or compositionally richer prompts typical of creative use cases remains unverified. Broader benchmarks such as DrawBench, T2I-CompBench, or GenAI-Bench would be necessary to establish wider validity.

Table 4. Mean cumulative drift $D_T$ ($\pm$ std) over 50 images on SD3.5-L. Percentage reduction relative to CFG at each $w$ is shown in parentheses. Lower is better. **Bold**: best per column.

| Method | $w = 4.5$ | $w = 6.0$ | $w = 7.5$ | $w = 10.0$ |
|---|---|---|---|---|
| CFG | 2008 | 2270 | 2575 | 3179 |
| +QN only | 1950 (−2.9%) | 2219 (−2.2%) | 2522 (−2.1%) | 3096 (−2.6%) |
| +D only | 1979 (−1.4%) | 2255 (−0.7%) | 2559 (−0.6%) | 3190 (+0.3%) |
| +QN+D (Ours) | **1923** (−4.2%) | **2206** (−2.8%) | **2511** (−2.5%) | **3106** (−2.3%) |

## References


[1] Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2209.15571*, 2022. 3

[2] 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. 2, 6

[3] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, et al. Scaling rectified flow transformers for high-resolution image synthesis. *International conference on machine learning (ICML)*, 2024. 3, 5

[4] Weichen Fan, Amber Yijia Zheng, Raymond A Yeh, and Ziwei Liu. Cfg-zero*: Improved classifier-free guidance for flow matching models. *arXiv preprint arXiv:2503.18886*, 2025. 2, 6

[5] 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. *Advances in neural information processing systems (NeurIPS)*, 30, 2017. 5

[6] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2, 3, 5, 6, 7

[7] Zexi Jia, Pengcheng Luo, Zhengyao Fang, Zhang. Jinchao, and Jie Zhou. Manifold-optimal guidance: A unified riemannian control view of diffusion guidance. *arXiv preprint arXiv:2603.11509*, 2026. 3, 4

[8] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preference for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1957–1967, 2023. 5

[9] Black Forest Labs. Flux.1: Text-to-image generation with flow-based transformers. https://blackforestlabs.ai/, 2024. 3

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 740–755, 2014. 4

[11] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for scalable generative modeling. In *International conference on machine learning (ICML)*, 2023. 3

[12] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical Programming*, 45(1-3):503–528, 1989. 3

[13] 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. 3

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning (ICML)*, pages 8748–8763. PMLR, 2021. 5

[15] Shreshth Saini, Shashank Gupta, and Alan Bovik. Rectified cfg++ for flow based models. *Advances in Neural Information Processing Systems (NeurIPS)*, 38:149034–149074, 2026. 2, 5, 7

[16] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:25278–25294, 2022. 5

[17] Hanyang Wang, Yiyang Liu, Jiawei Chi, Fangfu Liu, Ran Xue, and Yueqi Duan. Cfg-ctrl: Control-based classifier-free diffusion guidance. *arXiv preprint arXiv:2603.03281*, 2026. 2, 3, 5, 7

[18] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A better evaluation metric for text-to-image generation. *arXiv preprint arXiv:2306.09341*, 2023. 5

[19] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 5