TIME-WARPING GUIDANCE IN FLOW MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Inference-time guidance methods for diffusion and flow-matching image generators, from Classifier-Free Guidance (CFG) to recent autoguidance variants, form contrastive velocities between a model prediction and a less-informed baseline. We introduce **TWiG** (Time-Warping Guidance), which augments CFG with a second contrastive term whose probe lies at a different point in latent-time space, namely the conditional branch evaluated at a time-warped query that approximates a forward-process sample at a different sampling time. The warped query is built from a Tweedie noise estimate cached from the previous step, with a closed-form proxy and an explicit error bound. The schedule that selects the warped time is governed by a primary calibration scalar, together with two clip bounds. We hypothesize on geometric grounds that the calibration scalar is near $1/2$ and observe a corresponding empirical alignment peak on five backbones. We evaluate TWiG using automatic metrics spanning distribution, quality, and diversity (FID, CMMD, IS, Precision/Recall, CLIP, Image Reward, Aesthetic, Vendi, LPIPS, CVS), the compositional benchmarks GenEval and GenEval2, and a blinded pairwise human study. Across two evaluation backbones, TWiG improves CFG and shows favorable results against the evaluated baselines.

1 INTRODUCTION

Conditional generative models based on diffusion and flow matching are now the dominant paradigm for high-fidelity image synthesis (Ho et al., 2020; Song et al., 2021; Lipman et al., 2023; Liu et al., 2023; Esser et al., 2024). At sampling time, these models are commonly steered with Classifier-Free Guidance (CFG) (Ho & Salimans, 2021). CFG runs the model twice at every step of the denoising trajectory: once with the prompt, producing the conditional velocity, and once without it, producing an unconditional velocity. The two are combined into a *contrastive update* that pushes the trajectory toward the conditional velocity and away from the unconditional one, with a single scalar weight controlling how aggressively. This is widely used across architectures and tasks (Rombach et al., 2022; Podell et al., 2024; Esser et al., 2024; Black Forest Labs, 2025; Z-Image Team et al., 2025).

Several recent refinements of CFG stay inside this per-step template. One line of work post-processes the combined velocity so that its magnitude does not drift off the manifold the model was trained on (Zhang et al., 2024; Sadat et al., 2025a). Another replaces the unconditional branch with a deliberately weakened version of the model, on the intuition that subtracting a weaker self should sharpen the prediction more cleanly than subtracting an empty prompt; the weakening is implemented by perturbing attention, dropping blocks, or re-enabling dropout at inference (Ahn et al., 2024; Chen et al., 2026; Gu & Hou, 2025). A related method instead refines the initial noise along the guidance direction before sampling begins (Mannering et al., 2025). Self-Guidance (SG) forms cross-time contrast using a shifted-time prediction, and its variant SG-prev reuses the prediction from the preceding step (Li et al., 2024b). Time-Shift Sampler changes the base sampler's time assignment to mitigate exposure bias rather than as an auxiliary contrastive branch (Li et al., 2024a). These methods establish the value of cross-time inference. TWiG asks a complementary geometric question: how should the latent itself change when an auxiliary branch is evaluated at another time?

What if we asked the model at a different point? Picking a different point in the latent-time plane is awkward to do directly. However, for a linear flow, this can be precisely characterized: the forward process places the noisy latent z_t near a sphere of radius $t\sqrt{d}$ around a scaled version of the clean image z_0 in d dimensions. Lemma 1 shows that scaling z_t by a positive scalar β, followed by a correction along the noise direction, yields a forward sample at the warped time $t_\beta = 1 - \beta(1 - t)$.

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.

TWiG. Time-Warping Guidance (TWiG) adds a second contrastive term to CFG, formed against the conditional branch evaluated at exactly such a time-warped point. The augmented update is

$$v_t^{\mathrm{TWiG}} \;=\; v^{+} \;+\; w\,(v^{+} - v^{-}) \;+\; \eta\,(v^{+} - v^{\beta}),$$

where v^{+} and v^{-} are the conditional and unconditional velocities used by CFG, v^{β} is the time-warped probe, w is the CFG weight, and η is a new sampling-time scalar that controls the second contrast. The probe v^{β} adds one model evaluation per step (like most recent CFG-stacked autoguidance baselines), but the warped query point itself is constructed from quantities the sampler already maintains.

A schedule with a hypothesis attached. The remaining choice is which warped time t_β to aim at. The trade-off is concrete: at high t the latent is too noisy for the model to pin down the clean image, and at low t it is too clean for the model to pin down the noise; the cosine alignment between v^{CFG} and a trajectory-aware oracle direction therefore peaks at some interior time. We hypothesize, on geometric grounds, that this peak sits at $t^\star = 1/2$, where signal and noise scales are equal and the model is asked to identify both at once. We test this hypothesis directly and measure the cosine alignment as a function of time, and find the empirical peak $\hat{t}^\star$ very close to $1/2$ on five of five models tested. The schedule that pulls the warped time toward $t^\star$ has a closed form, $\beta(t) \propto (1 - t^\star)/(1 - t)$, with clipping to keep the warped time in range and the proxy error bounded. This introduces one calibration scalar, $t^\star$, and two clip bounds, $(\beta_{\min}, \beta_{\max})$, selected once per backbone.

Contributions. Our work introduces several novel insights and techniques. (i) **A second contrastive term against a time-warped probe.** We augment CFG with an additional contrast $\eta\,(v^{+} - v^{\beta})$, where v^{β} is the conditional branch evaluated at a single time-warped query point. The probe is one extra model evaluation per step. (ii) **Matched latent-time construction.** The matched-signal identity (Lemma 1) places the warped query exactly on a forward-process sample at t_β, given the noise; the linear-flow analogue of the Tweedie denoising estimator gives a closed-form proxy with an explicit error bound (Lemma 2); and the schedule $\beta^\star(t) = (1 - t^\star)/(1 - t)$ moves the warped time toward the cosine-alignment maximum under a single named assumption (Assumption 1). (iii) **Peak-seeking hypothesis.** We hypothesize that the optimal target time should be $t^\star = 1/2$, then verify on five unrelated rectified-flow text-to-image backbones that the empirical peak $\hat{t}^\star$ falls very close to $1/2$. We discuss this regularity as a non-trivial property of pretrained flow-matching models in Section 5. Finally, (iv) **Empirical validation.** We evaluate TWiG with metrics spanning distribution, quality, diversity, and compositional alignment. Across two backbones, Flux2-Klein-4B-Base and Z-Image, TWiG improves over CFG and compares favorably with the evaluated baselines. We further evaluate TWiG against nine of these baselines in a blinded pairwise human preference study; TWiG was preferred over every baseline on both axes.

2 Related Work

Diffusion and flow matching. Denoising diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021; Karras et al., 2022) and their latent variants (Rombach et al., 2022) underpin most large-scale text-to-image systems (Ramesh et al., 2022; Saharia et al., 2022; Podell et al., 2024; Esser et al., 2024). Flow matching (Lipman et al., 2023; Albergo & Vanden-Eijnden, 2023; Liu et al., 2023; Liu, 2022) reframes training as regression onto a target velocity field. Diffusion and rectified-flow Transformers (Peebles & Xie, 2023; Esser et al., 2024) are now the dominant architecture, and the Flux2 (Black Forest Labs, 2025) and Z-Image (Z-Image Team et al., 2025) backbones we use embody this design.

Classifier-Free Guidance. CFG (Ho & Salimans, 2021) replaces the external gradient of classifier guidance (Dhariwal & Nichol, 2021) with the contrast between a conditional and an unconditional prediction from the same network, and is widely used in text-to-image generation. A long line of work analyzes its mode-coverage trade-offs and its tendency to push samples off the training manifold (Karras et al., 2024; Kynkäänniemi et al., 2024; Sadat et al., 2025a; Zhang et al., 2024). Most refinements post-process CFG or modify its negative branch. Cross-time methods discussed below broaden that template by changing the time at which the model is evaluated. TWiG further couples such a time change with a matched transformation of the auxiliary latent.

Velocity normalization. A first response to the off-manifold drift of CFG is to keep the magnitude of the guided velocity in distribution. EP-CFG (Zhang et al., 2024) renormalizes the guided field to the magnitude of the conditional branch post-hoc, and APG (Sadat et al., 2025a) downweights

the component of the guidance delta parallel to the conditional velocity, the direction along which CFG over-stretches the latent. Per-frequency and per-region variants (Sadat et al., 2025b; Shen et al., 2024) apply different scales across spectral or spatial locations, and ReCFG (Xia et al., 2025) rectifies the two guidance coefficients in closed form so that the guided update remains consistent with the underlying diffusion theory. We compare TWiG directly to APG and EP-CFG. TWiG is structurally distinct: rather than correcting the magnitude of the guided velocity post-hoc, it adds a contrastive term against a time-warped probe of the same network, with the probe's energy rescaled to match the conditional branch in the spirit of EP-CFG.

Inference-time autoguidance. The line of work most directly relevant to TWiG is *autoguidance*: contrast against a deliberately weakened version of the same model rather than against an empty unconditional. The trained variants weaken via a smaller or undertrained twin (Karras et al., 2024) or via intermediate layers trained with auxiliary supervision (Zhou et al., 2026); the training-free variants extract a weaker self from the same network on the fly: PAG (Ahn et al., 2024) replaces self-attention with the identity, SLG (Stability AI, 2024a) skips transformer layers, ISG (Gu & Hou, 2025) re-enables dropout at inference for a stochastic second pass, SSG (Chen et al., 2026) drops blocks at random to form stochastic sub-networks, and SAG (Hong et al., 2023) adversarially blurs the regions the model attends to. NLG (Mannering et al., 2025) instead refines the initial noise toward better guidance alignment before standard sampling begins. Self-Guidance is the closest prior cross-time guidance method (Li et al., 2024b). SG contrasts the conditional prediction at (z_t, t) with a prediction at the same latent and a shifted time $t + \delta(t)$; SG-prev approximates this contrast by reusing the prediction from the preceding point (z_{t_p}, t_p). TWiG also forms a conditional self-contrast across time, but evaluates the auxiliary branch at a jointly transformed pair (z_{t_β}, t_β) derived from the matched-signal identity. Its schedule moves the auxiliary time toward an empirically observed interior alignment peak and can shift the query in either temporal direction.

Time-shifted sampling. Time-Shift Sampler assigns a generated state an alternative timestep selected to improve its coupling with the learned diffusion marginal and reduce exposure bias (Li et al., 2024a). This changes the timestep used by the base sampler with little additional computation. TWiG retains the base solver update at (z_t, t) and uses the changed-time probe (z_{t_β}, t_β) only as an auxiliary query. Thus, both methods use non-current time indices, but at different levels: Time-Shift Sampler adjusts the sampler state, whereas TWiG constructs a separate guidance probe.

3 METHOD

This section develops TWiG layer by layer. We ask what would be needed to query the conditional branch at a different point on the denoising trajectory, and show that a simple radial scaling of the latent corresponds, up to one explicit correction, to a forward-process sample at a different time (Section 3.1). The correction depends on the noise component of the latent, which we estimate locally via Tweedie's identity; the resulting query point is a closed-form proxy with an explicit error bound (Section 3.2). The remaining choice is which warped time to aim at; we hypothesize and then verify the natural geometric answer (Section 3.3). A small per-step rescaling of the time-warped probe (Section 3.4) and an algorithm summary (Section 3.5) close the section.

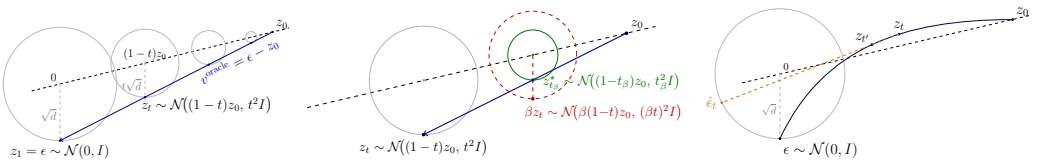
We work in the flow-matching framework with the linear schedule. Given a clean image z_0 drawn from the data distribution conditional on a prompt c, and Gaussian noise $\epsilon \sim \mathcal{N}(0, I_d)$ independent of z_0, the forward-process latent at time $t \in [0, 1]$ is

$$z_t \;=\; (1-t)\, z_0 + t\, \epsilon, \tag{1}$$

with denoising time running from $t = 1$ at the noise endpoint to $t = 0$ at the data endpoint. Differentiating along this line gives the (sampling-direction) *oracle velocity* $v^{\text{oracle}} = \dot{z}_t = \epsilon - z_0$, which is constant in t for fixed (z_0, ϵ). A flow-matching model M is trained to approximate the conditional expectation of this oracle, $u(z_t, t, c) = \mathbb{E}[\,\epsilon - z_0 \mid z_t, c\,]$. We write $v^+ = M(z_t, t, c)$ for the conditional output of the model and $v^- = M(z_t, t, \varnothing)$ for the unconditional one, obtained by passing the null condition. Standard CFG (Ho & Salimans, 2021) forms the contrastive update

$$v_t^{\text{CFG}} \;=\; v^+ + w\,(v^+ - v^-), \tag{2}$$

with a single sampling-time scalar $w \geq 0$. The trajectory is then traced by an ODE solver that integrates v_t^{CFG} from $t = 1$ down to $t = 0$.



(a) Time progression. (b) Scaling as time-warping. (c) Real trajectories bend.

Figure 1: Geometry behind TWiG. (a) For fixed (z_0, ϵ), z_t lies near a sphere of radius $t\sqrt{d}$ around $(1-t)z_0$. (b) Scaling z_t by β (red sphere) is reinterpreted as a latent at the warped time $t_\beta = 1 - \beta(1-t)$, where the matched-signal forward sample $z^\star_{t_\beta}$ lies on the inner green sphere. (c) Real sampler trajectories bend, so local estimates $\hat{\epsilon}_t$ differ from the original ϵ that initialized the trajectory.

3.1 Geometric motivation: scaling as time-warping

A folklore concentration fact in high dimensions (Appendix E) puts ϵ on a sphere of radius $\sqrt{d}$, so $z_t = (1-t)z_0 + t\,\epsilon$ lies near a sphere of radius $t\sqrt{d}$ around $(1-t)z_0$ (Figure 1(a)). The forward process is therefore a sweep through spheres of decreasing radii as t runs from 1 to 0. Multiplying z_t by a positive scalar β moves it from one sphere to another with the same intrinsic SNR (signal and noise scales each multiplied by β), but the result is no longer on the manifold the model was trained on: the signal scale $\beta(1-t)$ and noise scale βt no longer sum to 1. There is, however, a unique sampling time at which they would, namely $t_\beta = 1 - \beta(1-t)$. The corresponding "true" realization of the forward process at that time, sharing the same (z_0, ϵ) as z_t, is

$$z^\star_{t_\beta} \;=\; (1 - t_\beta)\, z_0 + t_\beta\, \epsilon.$$

The relation between the radially scaled point βz_t and this matched-signal forward sample is exact:

Lemma 1 (Matched-signal identity). *For any (z_0, ϵ, t) and any $\beta > 0$ with $t_\beta = 1-\beta(1-t) \in (0,1)$,*

$$\beta z_t \;=\; z^\star_{t_\beta} + (\beta - 1)\, \epsilon. \tag{3}$$

In particular, $\|\beta z_t - z^\star_{t_\beta}\| = |\beta - 1| \cdot \|\epsilon\|$: the offset is along the noise direction ϵ, with magnitude that does not depend on t.

Proof. $\beta z_t \;=\; \beta(1-t)\, z_0 + \beta t\, \epsilon \;=\; (1 - t_\beta)\, z_0 + \beta t\, \epsilon$, using $1 - t_\beta \;=\; \beta(1-t)$. Subtracting $z^\star_{t_\beta} = (1 - t_\beta)\, z_0 + t_\beta\, \epsilon$ leaves $(\beta t - t_\beta)\, \epsilon = (\beta - 1)\, \epsilon$. □

The identity is exact and assumption-free. By the high-dimensional concentration of ϵ (Appendix E), $\|\epsilon\| \approx \sqrt{d}$ with high probability, so the offset between βz_t and the matched-signal sphere has magnitude $\approx |\beta - 1|\sqrt{d}$, the visual gap between the two spheres in Figure 1(b). Rearranged, $z^\star_{t_\beta} = \beta z_t - (\beta - 1)\, \epsilon$: knowing the noise ϵ, we can take the radially scaled point and pull it back to a forward-process sample at the warped time exactly. The remainder of the construction estimates ϵ.

3.2 The TWiG query: estimating the noise locally

On a real CFG-driven sampler the trajectory bends, and the effective noise direction at later steps drifts from the initial ϵ (Figure 1(c)). Therefore, at sampling time we should not use the original noise ϵ that initialized the trajectory to correct the scaled latent. Nevertheless, we have access to the model's local velocity prediction. The linear-flow analogue of the Tweedie denoising identity (Efron, 2011) turns this prediction into a noise estimate. The oracle relation follows exactly from $z_t = (1-t)z_0 + t\,\epsilon$ and $v_t = \epsilon - z_0$, which give $\epsilon = z_t + (1-t)v_t$. Substituting the conditional model output for the unknown oracle velocity yields

$$\hat{\epsilon}_t \;=\; z_t + (1-t)\, v_t^+, \tag{4}$$

This is a local conditional tangent estimate. In the implementation evaluated throughout, we cache it and use it to construct the auxiliary query at the following sampling step. The estimate uses only quantities the sampler already computes.

Plugging $\hat{\epsilon}_{t_p}$ into the matched-signal identity gives the TWiG query point and the time-warped probe,

$$z_{t_\beta} \;=\; \beta\, z_t \;-\; (\beta-1)\,\hat{\epsilon}_{t_p}, \qquad v^\beta \;=\; M(z_{t_\beta},\, t_\beta,\, c), \tag{5}$$

and the TWiG velocity adds a second contrastive term to CFG:

$$v_t^{\mathrm{TWiG}} \;=\; v^+ \;+\; w\,(v^+ - v^-) \;+\; \eta\,(v^+ - v^\beta). \tag{6}$$

The probe v^β is one extra forward pass through the model per step (so TWiG runs three model calls per step in total, the same number as most CFG-stacked autoguidance baselines), but the query point at which it is evaluated is built from cached quantities. The query uses the most recently available local estimate, which is refreshed once per sampling step. The error of the proxy is bounded directly by the noise-estimation gap:

Lemma 2 (Proxy error)**.** *With $z_{t_\beta} = \beta z_t - (\beta-1)\,\hat{\epsilon}$ for any noise estimate $\hat{\epsilon}$, and $z^\star_{t_\beta}$ as in Lemma 1,*

$$z_{t_\beta} - z^\star_{t_\beta} \;=\; -(\beta-1)\,\big(\hat{\epsilon}-\epsilon\big), \qquad \big\|z_{t_\beta} - z^\star_{t_\beta}\big\| \;=\; |\beta-1|\cdot\big\|\hat{\epsilon}-\epsilon\big\|. \tag{7}$$

Proof. Subtract Lemma 1 from the definition of z_{t_β}. □

The proxy is exact when the local noise estimate matches the trajectory's underlying noise (e.g. on a perfectly straight forward trajectory). Lemma 2 shows that its error scales linearly with both $|\beta-1|$ and the noise-estimation gap. The schedule clips the first factor, while the consecutive-velocity diagnostics in Figure 2 provide empirical evidence that the velocity field changes smoothly over individual sampling steps. The geometry determines a structured cross-time direction, but it does not determine which sign of that direction should improve sampling. On real trajectories, the deployed probe remains close to v^+ while being slightly less aligned with the trajectory-aware oracle. We therefore interpret $v^+ - v^\beta$ as a structured self-contrast: unlike architectural autoguidance, its auxiliary prediction is weakened by the approximate matched query rather than by modifying the network. Appendix K tests this interpretation and the beneficial sign directly.

3.3 Choosing a $\beta(t)$ schedule

The remaining design choice is the schedule $\beta(t)$. We start from a hypothesis about where the time-warped probe is most informative, and then test it.

Since latents are a mixture of noise and signal, the trade-off is concrete: at high t, the latent is too noisy for the model to pin down the clean image accurately, and at low t, it is too clean for the model to pin down the noise; the cosine alignment between the guided velocity v^{CFG} and a trajectory-aware oracle direction v_t^{oracle} (Eq. (10) in Appendix A) therefore peaks at an interior point, which we assume is unique.

Assumption 1 (Unimodal cosine alignment)**.** *On real CFG-bent trajectories, the cosine similarity $\cos(v^{\mathrm{CFG}}(z_t,t,c),\, v_t^{\mathrm{oracle}}(z_t,t))$ is unimodal in t with a unique interior maximum at some $t^\star \in (0,1)$.*

On geometric grounds, we expect $t^\star$ to sit at $t^\star = 1/2$, the SNR-one regime where the signal scale $1-t$ and the noise scale t are equal. At this point the model is asked to identify the clean image and the noise simultaneously, which is exactly the prediction the cosine alignment with the oracle is sensitive to. To verify our hypothesis, we measure the cosine alignment directly on Flux2-Klein (Figure 2, left). The curve is unimodal in t with a clear interior peak, and the empirical peak $\hat{t}^\star$ falls very close to $1/2$. The same measurement on Z-Image (Appendix C) and three additional backbones (Appendix D) places the peak at the same location. The alignment peak sits at $t^\star \approx 1/2$ on five out of five models tested. We discuss this regularity in Section 5.

The schedule. With $t^\star$ in hand, the schedule that steers the warped time $t_\beta = 1-\beta(1-t)$ to $t^\star$ at every step is closed-form: $\beta^\star(t) = (1-t^\star)/(1-t)$. For $t > t^\star$ (early in denoising) this gives $\beta > 1$, so the probe is queried at a later, lower-noise time (a *speculative* probe, looking ahead along the trajectory); for $t < t^\star$ (late in denoising) it gives $\beta < 1$, so the probe is queried at an earlier, higher-noise time (a *retrospective* probe, looking back). To keep the warped time inside $(0,1)$ and to bound the proxy error of Lemma 2, we clip:

$$\beta(t) \;=\; \mathrm{clip}\!\left(\frac{1-t^\star}{1-t},\; \beta_{\min},\; \beta_{\max}\right). \tag{8}$$

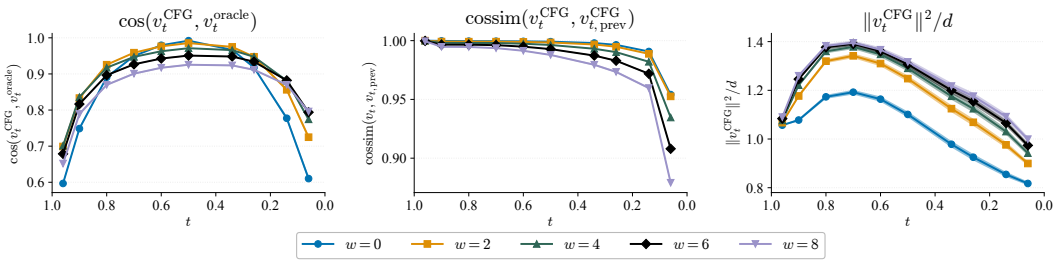


Figure 2: Empirical diagnostics on Flux2-Klein across CFG weights $w \in \{0, 2, 4, 6, 8\}$. **Left:** cosine alignment of v^{CFG} with the trajectory-aware oracle (Eq. (10)) is unimodal in t with peak $\hat{t}^\star \approx 1/2$. **Middle:** consecutive-step velocity cosine similarity stays close to 1, the regime in which the proxy error of Lemma 2 is small. **Right:** the guided velocity norm $\|v^{\mathrm{CFG}}\|$ varies systematically with t; the $w = 0$ curve equals $\|v^+\|$ and motivates Eq. (9). Protocol details and the same diagnostics on Z-Image are in Appendix C.

The base sampler remains at (z_t, t); only the auxiliary conditional query is evaluated at (z_{t_β}, t_β). The target $t^\star$ is calibrated once per backbone, while the clip bounds $(\beta_{\min}, \beta_{\max})$ are selected once and are not tuned per prompt.

3.4 Per-step norm correction of the time-warped probe

The probe v^β is queried at $t_\beta \neq t$, and Figure 2 (right) shows that guided velocity norms vary systematically with sampling time; at $w = 0$, the plotted velocity is exactly v^+ (for a rigorous analysis and proof, see Appendix B). Consequently, $\|v^\beta\|$ does not match $\|v^+\|$, and the raw difference $v^+ - v^\beta$ would be dominated by this magnitude mismatch rather than by the directional disagreement we want to use. Several corrections are possible (rescale v^β, rescale the entire contrastive term, or set a target dictated by the schedule); we adopt the simplest, a per-sample rescaling of v^β to the norm of v^+,

$$v^\beta \leftarrow v^\beta \cdot \frac{\|v^+\|}{\|v^\beta\|}, \tag{9}$$

and ablate the choice in Section 4.

Algorithm 1 TWiG sampling.

Require: $z_{t_0} \sim \mathcal{N}(0, I)$, M, c, $\beta(\cdot)$, w, η, $1 = t_0 > \ldots > t_N = 0$

1: $\hat{\epsilon} \leftarrow z_{t_0}$
2: **for** $i = 0, \ldots, N-1$ **do**
3: $\quad t \leftarrow t_i$
4: $\quad v^+ \leftarrow M(z_t, t, c), \;\; v^- \leftarrow M(z_t, t, \varnothing)$
5: $\quad \beta \leftarrow \beta(t), \;\; t_\beta \leftarrow 1 - \beta(1 - t)$
6: $\quad z_{t_\beta} \leftarrow \beta\, z_t - (\beta - 1)\, \hat{\epsilon}$ ▷ using the cached $\hat{\epsilon}$
7: $\quad v^\beta_{\mathrm{raw}} \leftarrow M(z_{t_\beta}, t_\beta, c)$
8: $\quad v^\beta \leftarrow v^\beta_{\mathrm{raw}} \cdot \|v^+\| / \|v^\beta_{\mathrm{raw}}\|$
9: $\quad v_t \leftarrow v^+ + w(v^+ - v^-) + \eta(v^+ - v^\beta)$
10: $\quad \hat{\epsilon} \leftarrow z_t + (1 - t)\, v^+$ ▷ cache for next step
11: $\quad z_t \leftarrow z_t + (t_{i+1} - t_i)\, v_t$
12: **end for**
13: **return** z_t

3.5 The TWiG algorithm

Algorithm 1 summarizes the sampler. Initialization (line 1) seeds the cached noise estimate with the starting latent, exploiting the linear-schedule identity $z_1 = \epsilon$. The main loop unfolds in five stages. First, the conditional and unconditional CFG branches are evaluated at the current latent to obtain v^+ and v^- (line 4). Second, the schedule produces $\beta = \beta(t)$ and the warped time $t_\beta = 1 - \beta(1 - t)$ (line 5). Third, the warped query point $z_{t_\beta} = \beta z_t - (\beta - 1)\, \hat{\epsilon}$ is constructed from the cached noise estimate (line 6), and the model is evaluated at it and norm-corrected to obtain v^β (lines 7–8). Fourth, the additive TWiG velocity $v_t = v^+ + w(v^+ - v^-) + \eta(v^+ - v^\beta)$ is formed (line 9). Fifth, the cached noise estimate is refreshed and an Euler step advances the latent to time t_{i+1} (lines 10–11). The sampler exposes two sampling-time weights, the standard CFG weight w and the new TWiG weight η. The schedule uses the calibrated target $t^\star$ and the selected clip bounds $(\beta_{\min}, \beta_{\max})$, all fixed per backbone rather than per prompt.

4 EXPERIMENTS

We evaluate TWiG quantitatively against CFG and the recent autoguidance and velocity-normalization baselines, and follow up with a blinded pairwise human study on the same prompt set. Side-by-side qualitative comparisons are deferred to Appendix M for space.

Models and pipeline. We evaluate on Flux2-Klein-4B-Base (Black Forest Labs, 2025) (Flux2-Klein for short) and Z-Image (Z-Image Team et al., 2025), both rectified-flow latent transformers. TWiG is implemented as a drop-in modification of the standard guidance call inside the `diffusers` library (von Platen et al., 2022); pseudocode in Algorithm 1. Unless stated otherwise, experiments ran for 50 diffusion steps; the matched-compute CFG baseline ran for 75. We use Diffusers CFG scales $s = 4.5/5.0$, corresponding to $w = s - 1 = 3.5/4.0$ in Eq. (2), for Flux2-Klein and Z-Image respectively. TWiG additionally uses $\eta = 1.5$, $\beta_{\min} = 0.8$, $\beta_{\max} = 1.2$, and the calibrated $t^\star = 1/2$. The full evaluation protocol (image counts, resolutions, and seeds) and the per-method hyperparameters are given in Appendix F.

Baselines and metrics. We compare TWiG against standard CFG, against inference-time autoguidance methods (ISG (Gu & Hou, 2025), SSG (Chen et al., 2026), NLG (Mannering et al., 2025), PAG (Ahn et al., 2024), and SG and SG-prev (Li et al., 2024b)), and against velocity-normalization methods (APG (Sadat et al., 2025a) and EP-CFG (Zhang et al., 2024)). We also compare with CFG run for 75 steps, which matches TWiG's 150 total model evaluations. We report distribution metrics, FID (Heusel et al., 2017), CMMD (Jayasumana et al., 2024), Inception Score (Salimans et al., 2016), and Improved Precision/Recall (Kynkäänniemi et al., 2019); quality metrics, Aesthetic Score (AS) (Schuhmann, 2022), CLIP Score (Hessel et al., 2021), and Image Reward (IR) (Xu et al., 2023); diversity metrics, Vendi Score (Friedman & Dieng, 2023), mean pairwise LPIPS (Zhang et al., 2018), and the Conditional Vendi Score (CVS) (Jalali et al., 2026); and compositional benchmarks GenEval (Ghosh et al., 2023) and GenEval2 (Kamath et al., 2025).

Quantitative results Table 1 reports all metrics on Flux2-Klein (top) and Z-Image (bottom). Relative to 50-step CFG, TWiG lowers FID from 23.10 to 19.64 on Flux2-Klein and from 23.73 to 21.88 on Z-Image, with corresponding improvements in CMMD, Inception Score, CLIP Score, and Image Reward. It also improves all three reported diversity metrics on both backbones; Appendix J offers a theoretical argument to explain the observed increase in the diversity metrics. The trade-off is not uniformly dominant: APG attains the lowest FID on both models, EP-CFG leads selected precision and quality metrics, SG-prev leads Flux precision, and SG gives the highest Z-Image Aesthetic Score. TWiG's main advantage is therefore its balanced profile across distributional fidelity, semantic quality, and diversity rather than the number of individually winning columns.

We further evaluate all methods on GenEval (Ghosh et al., 2023) and GenEval2 (Kamath et al., 2025). TWiG has the highest overall GenEval and GenEval2 scores on both backbones (on Z-Image, GenEval2 is tied with SG-prev). This is important because the strongest FID improvements do not consistently preserve compositional alignment: relative to CFG, APG loses 16.6 and 6.2 GenEval points on Flux2-Klein and Z-Image, while PAG loses 5.6 and 22.8. TWiG instead improves FID, CLIP Score, Image Reward, and all reported diversity metrics over CFG while preserving or slightly improving both compositional benchmarks. The results suggest a favorable quality-diversity-alignment trade-off rather than uniform dominance. The protocol and per-task breakdowns are in Appendix L.

TWiG performs three model calls per step (conditional, unconditional, time-warped) where CFG performs two. At 50 steps, TWiG takes 5.22 and 6.18 seconds per image on Flux2-Klein and Z-Image. Full SG also uses 150 model evaluations and is slower at 5.55 and 7.66 seconds. SG-prev caches the preceding conditional prediction, uses the same 100 evaluations as CFG, and is cheaper than both TWiG and full SG. At matched total NFE, 75-step CFG takes 5.60 and 6.80 seconds and does not recover TWiG's overall gains: TWiG improves every reported metric except precision and Aesthetic Score on Flux2-Klein, where GenEval is tied, and every metric except Aesthetic Score on Z-Image. Full per-method compute accounting is given in Appendix H.

User study We additionally performed a blinded pairwise user study comparing TWiG with nine baselines. The responses were collected from 19 participants who ranked $1{,}949$ randomized TWiG-versus-baseline pairs drawn from the two backbones; study assignments were organized around 100 comparisons. Participants rated every pair on overall image quality and prompt alignment, with

Table 1: Main comparison on **Flux2-Klein-4B-Base** (top) and **Z-Image** (bottom). Methods are grouped by category: CFG baseline (black), inference-time autoguidance (blue), and velocity norm correcting (red), with TWiG listed last in each block. FID, CMMD: lower is better. All others: higher is better. **Bold** = best, underline = second best per column, computed within each backbone. The GE and GE2 columns report the overall scores on the compositional benchmarks GenEval and GenEval2, measured at 512×512 (per-task breakdowns in Appendix L). Unless marked otherwise, methods use 50 sampling steps. TWiG and SG use 150 NFE; CFG, SG-prev, and most two-pass baselines use 100; CFG-75 is the 150-NFE matched-compute CFG baseline. The Time column reports end-to-end wall-clock seconds per image (point estimates; full accounting is in Appendix H).

		Distribution					*Quality*			*Diversity*			*GenEval/2*		*Time*
Model	Method	FID↓	CMMD↓	IS↑	Prec↑	Rec↑	AS↑	CLIP↑	IR↑	Vendi↑	LPIPS↑	CVS↑	GE↑	GE2↑	s/img↓
Flux2-Klein-4B-Base	CFG	23.10	0.0119	38.50	0.708	0.390	5.27	27.05	0.828	5.91	1.580	2.444	0.779	0.672	3.78
	CFG-75	23.06	0.0121	38.89	0.721	0.399	5.27	27.06	0.827	5.92	1.619	2.499	**0.782**	0.666	5.60
	PAG	30.80	0.0325	29.25	0.537	0.439	**5.39**	26.05	0.426	4.86	1.584	2.443	0.723	0.624	5.14
	NLG	22.85	0.0117	38.53	0.708	0.396	5.26	27.07	0.831	5.92	1.404	2.000	0.773	0.672	4.59
	ISG	22.94	0.0118	38.46	0.711	0.385	5.26	27.07	0.831	5.92	1.580	2.445	0.762	0.629	5.49
	SSG	30.73	0.0265	30.41	0.598	0.252	**5.39**	26.90	0.577	5.12	1.569	2.388	0.749	0.631	5.49
	SG	21.64	0.0106	39.15	0.730	0.425	5.24	27.02	0.813	6.01	1.638	2.547	0.759	0.661	5.55
	SG-prev	21.76	0.0099	39.37	**0.732**	0.417	5.23	27.13	0.823	6.08	1.629	2.525	0.781	0.672	3.80
	APG	**16.12**	**0.0080**	37.95	0.708	0.455	5.08	26.90	0.682	6.14	1.627	2.608	0.613	0.575	3.82
	EP-CFG	20.07	0.0087	**39.46**	0.720	0.422	5.16	27.15	0.823	6.12	1.595	2.502	0.772	0.664	**3.75**
	TWiG	19.64	0.0084	**39.46**	0.705	**0.461**	5.20	**27.69**	**0.834**	**6.16**	**1.646**	**2.612**	**0.782**	**0.676**	5.22
Z-Image	CFG	23.73	0.0172	34.24	0.638	0.367	5.17	27.08	+0.710	5.66	0.650	1.386	0.693	0.752	**4.41**
	CFG-75	23.44	0.0176	34.68	0.644	0.375	5.19	26.94	+0.682	5.67	0.663	1.400	0.701	0.755	6.80
	PAG	88.69	0.1002	11.46	0.195	0.197	4.09	23.12	−1.348	3.18	0.611	**1.453**	0.465	0.600	6.10
	NLG	23.71	0.0171	34.38	0.640	0.371	5.18	27.07	+0.715	5.66	0.644	1.388	0.689	0.762	5.29
	ISG	23.56	0.0161	34.57	0.639	0.378	5.11	27.26	+0.713	5.72	0.650	1.401	0.674	0.742	6.87
	SSG	25.11	0.0182	31.87	0.604	0.349	5.11	27.23	+0.634	5.75	0.653	1.441	0.676	0.749	6.68
	SG	22.53	0.0171	35.13	0.636	0.386	**5.20**	26.90	+0.689	5.78	0.681	1.413	0.674	0.746	7.66
	SG-prev	22.31	0.0148	35.19	0.644	0.385	4.82	27.65	+0.649	5.75	0.631	1.441	0.682	**0.764**	5.17
	APG	**18.69**	0.0154	34.07	0.643	**0.434**	5.12	26.69	+0.624	5.58	**0.686**	1.426	0.631	0.707	4.48
	EP-CFG	22.26	0.0145	35.15	**0.654**	0.388	5.06	27.09	**+0.747**	5.78	0.662	1.420	0.698	0.758	4.44
	TWiG	21.88	**0.0141**	**35.27**	0.650	0.386	5.10	**27.69**	+0.734	**5.83**	**0.686**	1.426	**0.704**	**0.764**	6.18

a tie option. Appendix N details the study design and Appendix O shows the participant screens. We analyze the study with a variance-corrected clustered McNemar test (Durkalski et al., 2003) to account for correlations among rankings from the same participant (Appendix N). Pooling over backbones and baselines, TWiG is preferred on 43.0% of quality pairs versus 29.1% for the baselines ($\chi^2 = 5.16$, $p = 0.023$), and on 8.0% versus 4.2% of prompt-alignment pairs ($\chi^2 = 9.49$, $p = 0.002$). Descriptively, TWiG's win share exceeds its loss share in every per-backbone, per-baseline cell.

Ablations We ablate each design choice quantitatively on Flux2-Klein (256×256, 50 steps, $N = 5{,}000$ COCO prompts per configuration, identical prompt and seed set across configurations), reporting Image Reward and CLIP; Table 2 collects the four studies, and Appendix I repeats them with 95% confidence intervals. At the level of the point estimates, (a) every row of the $s \times \eta$ grid peaks at $\eta = 1.5$ on both metrics, with higher IR and CLIP than CFG ($\eta = 0$) at every fixed Diffusers CFG scale s. The different response to sweeping s suggests that η does more than simply rescale CFG guidance. (b) The clip $[0.8, 1.2]$ gives the highest point estimates on both metrics. Performance drops sharply when $\beta_{\max}$ exceeds 1.4 and more gradually as $\beta_{\min}$ moves below 0.8, a pattern consistent with the sensitivity in Lemma 2. (c) Without the norm correction of Eq. (9), IR drops to 0.828, compared with 0.882 with the correction and 0.876 for CFG, suggesting that the correction is important. (d) Later TWiG activation is associated with progressively smaller point-estimate gains, suggesting that the early, high-noise steps can contribute usefully despite the cached noise estimate.

5 Discussion and Limitations

One empirical regularity relevant to TWiG is that the measured cosine-alignment peak $t^\star$ lies near 0.5 on the five rectified-flow text-to-image models we tested, despite differences in architecture, training data, and organization. We use this as an empirical calibration rule, not as a universal constant. We do not currently have a proof, and deriving or more precisely characterizing the peak under the unimodality assumption (Assumption 1) remains future work.

Table 2: Quantitative ablations on Flux2-Klein ($N = 5{,}000$ COCO prompts per configuration; cells report IR/CLIP). **Bold** = the paper's configuration (Diffusers scale $s = 4.5$, equivalently $w = 3.5$ in Eq. (2); $\eta = 1.5$, $\beta \in [0.8, 1.2]$, norm correction on). Versions with 95% CIs are in Appendix I.

(a) Diffusers CFG scale s vs. TWiG weight η

$s\backslash\eta$	0 (CFG)	0.5	1	1.5	2
3	0.695/26.65	0.713/26.72	0.721/26.75	0.725/26.80	0.721/26.58
4.5	0.876/26.93	0.880/26.99	0.879/27.00	**0.882/27.03**	0.878/26.91
6	0.830/27.06	0.835/27.06	0.838/27.11	0.844/27.13	0.839/27.03

(b) Schedule clip $\beta_{\min}$ vs. $\beta_{\max}$

$\beta_{\min}\backslash\beta_{\max}$	1	1.2	1.4	1.6	1.8
1	0.876/26.93	0.879/26.95	0.865/26.64	0.802/26.60	0.777/26.02
0.9	0.877/26.98	0.880/26.99	0.867/26.65	0.785/26.57	0.750/26.04
0.8	0.879/27.00	**0.882/27.03**	0.870/26.61	0.744/26.57	0.728/26.03
0.7	0.848/26.97	0.866/26.54	0.851/26.03	0.712/25.95	0.702/25.87
0.6	0.809/26.88	0.796/26.13	0.725/25.99	0.688/25.91	0.678/25.84

(c) Norm correction (Eq. (9))

	IR	CLIP
Off	0.828	26.92
On (TWiG)	**0.882**	**27.03**

(d) CFG-then-TWiG warm start

Start step	IR	CLIP
K=0 (TWiG)	**0.882**	**27.03**
K=5	0.881	27.02
K=15	0.879	26.98
K=25	0.878	26.94
K=50 (CFG)	0.876	26.93

A practical caveat is that the time-warped proxy and the norm correction inherit error from the noise-estimation gap $\|\epsilon - \hat{\epsilon}_{t_p}\|$, so TWiG is best suited to backbones whose conditional flow is approximately straight, in particular rectified-flow models (Appendix C). The matched-signal identity also uses the linear schedule $(\alpha_t, \sigma_t) = (1 - t, t)$; non-linear monotone schedules can be accommodated by redefining t_β as the time at which $\alpha_{t_\beta} = \beta\,\alpha_t$, but we have not validated this empirically.

The other practical caveat is wall-clock cost: TWiG runs three model calls per sampling step where CFG runs two. Its measured runtime is $1.380\times$ and $1.400\times$ that of 50-step CFG on Flux2-Klein and Z-Image. Full SG has comparable or higher cost, and a 75-step CFG baseline matches TWiG's 150 total model evaluations and is slower in our measurements; it improves little over 50-step CFG and does not recover TWiG's overall quality-diversity-alignment profile (Table 1).

Several extensions follow naturally from the geometric perspective we use, "warp the time, not just the latent". The peak $t^\star$ may vary across prompts, so a per-prompt or per-cluster $t^\star$ may further improve TWiG. The time-warped probe is a natural teacher signal for guidance distillation, which would eliminate the warped query at deployment and amortize the wall-clock overhead. More broadly, Lemma 1 provides a general way of querying the model at non-current times, potentially useful for editing and anchoring beyond guidance.

6 Conclusion

We introduced Time-Warping Guidance, a sampling-time intervention for conditional flow-matching and diffusion models that augments CFG with a second contrastive term against the conditional branch evaluated at a time-warped query point: a proxy for the corresponding matched forward-process sample, built from a Tweedie-style noise estimate cached from the previous step, with a closed-form error bound (Lemma 2). A clipped peak-seeking schedule steers the warped time toward the cosine-alignment peak $t^\star$, using one calibrated target and two selected clip bounds. Across all five backbones we tested, the calibration recovers $\hat{t}^\star \approx 1/2$, a non-trivial empirical regularity that we leave open as a theoretical question.

TWiG keeps CFG's standard tuning interface and adds one new sampling-time scalar η on top of the fixed schedule parameters. Across two evaluation backbones, TWiG improves several distribution and semantic-quality metrics, all reported diversity metrics, and the overall GenEval and GenEval2 scores relative to CFG. Compared with the evaluated alternatives, including SG and SG-prev, its main empirical advantage is a balanced improvement without the pronounced compositional trade-off exhibited by some distribution-focused methods. The human evaluation points in the same direction: pooled preferences favor TWiG for both quality and prompt alignment, and its win share exceeds its loss share in every per-backbone, per-baseline comparison. Together, these results show TWiG as a practical way to extend CFG without modifying model weights or the base solver.

AI USE STATEMENT

We used generative AI tools to provide feedback on research methodology and experimental design, and to assist with writing and refining implementation code. The method, mathematical claims, proofs, hypotheses, and conceptual framework were developed by the authors. AI tools did not generate synthetic datasets, translate material, clean or reformat datasets, conduct qualitative or thematic data analysis, or interpret experimental results. All AI-assisted code was reviewed and tested by a human author.

Generative AI tools also assisted with drafting and editing parts of the manuscript for clarity and readability; writing the instructions for the user study; creating or modifying scientific figures; summarizing and analyzing existing literature; identifying potentially relevant literature as a complement to the authors' manual literature review; and organizing research notes. They were not used to suggest experimental parameters, create other research artifacts, determine the paper's structure, or transcribe research recordings. The remaining disclosure categories are not applicable to this work. The authors reviewed all AI-assisted material, verified literature-related suggestions against the original sources, and take responsibility for the final content of the paper, including its text, claims, code, analyses, and artifacts.

ETHICS STATEMENT

This paper introduces an inference-time guidance method for existing diffusion and flow-matching models. It does not train a new model, update model weights, or introduce a new dataset; the experiments use publicly available models and benchmarks under their respective licenses. The method therefore does not create a distinct category of ethical risk beyond those already associated with the base models, including the possibility of biased, harmful, misleading, or otherwise inappropriate generated content. Users remain responsible for applying the base models and the proposed guidance method consistently with the models' licenses and appropriate safeguards.

REPRODUCIBILITY STATEMENT

Section 3 derives the TWiG query, schedule, and norm correction, and Algorithm 1 specifies the complete sampling procedure. Section 4 describes the evaluation setting, baselines, metrics, quantitative comparisons, human evaluation, and ablations. Appendix F gives the implementation details, equations, and hyperparameters for TWiG and every baseline; Appendix G documents the cached quantities and numerical checks used for empirical validation; and Appendix H reports model-evaluation and timing costs. Appendices I, K, and L provide confidence intervals, probe diagnostics, and compositional-benchmark protocols and breakdowns. Appendices M, N, and O provide the extended qualitative comparison and the full human-evaluation design, analysis, outcome rates, and participant screens. Together, these sections specify the algorithms, configurations, evaluation data, sample counts, metrics, statistical analyses, and compute accounting needed to reproduce the reported experiments.

REFERENCES

Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. In *European Conference on Computer Vision (ECCV)*, pp. 1–17, 2024. doi: 10.1007/978-3-031-73464-9_1. URL https://doi.org/10.1007/978-3-031-73464-9_1. PAG; arXiv:2403.17377.

Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=li7qeBbCR1t.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin,

Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

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

Chubin Chen, Jiashu Zhu, Xiaokun Feng, Nisha Huang, Chen Zhu, Meiqi Wu, Fangyuan Mao, Jiahong Wu, Xiangxiang Chu, and Xiu Li. Stochastic self-guidance for training-free enhancement of diffusion models. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=9RyRIyzgT5. SSG (S^2-Guidance); arXiv:2508.12880.

Geoffrey Cideron, Andrea Agostinelli, Johan Ferret, Sertan Girgin, Romuald Elie, Olivier Bachem, Sarah Perrin, and Alexandre Ramé. Diversity-rewarded CFG distillation. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=lJ66m0ibQL.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 8780–8794, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html.

Valerie L. Durkalski, Yuko Y. Palesch, Stuart R. Lipsitz, and Philip F. Rust. Analysis of clustered matched-pair data. *Statistics in Medicine*, 22(15):2417–2428, 2003. doi: 10.1002/sim.1438. URL https://doi.org/10.1002/sim.1438.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. doi: 10.1198/jasa.2011.tm11181. URL https://doi.org/10.1198/jasa.2011.tm11181.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, pp. 12606–12633, 2024. URL https://proceedings.mlr.press/v235/esser24a.html.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research (TMLR)*, 2023. URL https://openreview.net/forum?id=g97OHbQyk1.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 36, pp. 52132–52152, 2023. doi: 10.52202/075280-2270. URL https://papers.nips.cc/paper_files/paper/2023/hash/a3bf71c7c63f0c3bcb7ff67c67b1e7b1-Abstract-Datasets_and_Benchmarks.html.

Enhao Gu and Haolin Hou. In-situ autoguidance: Eliciting self-correction in diffusion models. *arXiv preprint arXiv:2510.17136*, 2025. URL https://arxiv.org/abs/2510.17136. ISG.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7514–7528, 2021. doi: 10.18653/v1/2021.emnlp-main.595. URL https://doi.org/10.18653/v1/2021.emnlp-main.595.

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 *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6626–6637, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html. FID.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI. Extended version: arXiv:2207.12598 (2022).

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 6840–6851, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7428–7437, 2023. doi: 10.1109/ICCV51070.2023.00686. URL https://doi.org/10.1109/ICCV51070.2023.00686. SAG; arXiv:2210.00939.

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. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20349–20360, 2023. doi: 10.1109/ICCV51070.2023.01866. URL https://doi.org/10.1109/ICCV51070.2023.01866.

Mohammad Jalali, Azim Ospanov, Amin Gohari, and Farzan Farnia. Conditional vendi score: Prompt-aware diversity evaluation for generative AI models and LLMs. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 300, pp. 1369–1377, 2026. URL https://proceedings.mlr.press/v300/jalali26a.html. CVS; conditional extension of the Vendi Score.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking FID: Towards a better evaluation metric for image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9307–9315, 2024. doi: 10.1109/CVPR52733.2024.00889. URL https://doi.org/10.1109/CVPR52733.2024.00889. CMMD.

Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. GenEval 2: Addressing benchmark drift in text-to-image evaluation. *arXiv preprint arXiv:2512.16853*, 2025. URL https://arxiv.org/abs/2512.16853.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 26565–26577, 2022. doi: 10.52202/068431-1926. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/a98846e9d9cc01cfb87eb694d946ce6b-Abstract-Conference.html.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 52996–53021, 2024. doi: 10.52202/079017-1679. URL https://doi.org/10.52202/079017-1679. Autoguidance; arXiv:2406.02507.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3929–3938, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/0234c510bc6d908b28c70ff313743079-Abstract.html.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 122458–122483, 2024. doi: 10.52202/079017-3892. URL https://papers.nips.cc/paper_files/paper/2024/hash/dd540e1c8d26687d56d296e64d35949f-Abstract-Conference.html. arXiv:2404.07724.

Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000. doi: 10.1214/aos/1015957395. URL https://doi.org/10.1214/aos/1015957395.

Mingxiao Li, Tingyu Qu, Ruicong Yao, Wei Sun, and Marie-Francine Moens. Alleviating exposure bias in diffusion models through sampling with shifted time steps. In *International Conference on Learning Representations (ICLR)*, 2024a. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/483f8d2018d7025c87dd07e9b02fe4bf-Abstract-Conference.html. Time-Shift Sampler; arXiv:2305.15583.

Tiancheng Li, Weijian Luo, Zhiyang Chen, Liyuan Ma, and Guo-Jun Qi. Self-guidance: Boosting flow and diffusion generation on their own. *arXiv preprint arXiv:2412.05827*, 2024b. URL https://arxiv.org/abs/2412.05827. SG and SG-prev.

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 (ECCV)*, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://doi.org/10.1007/978-3-319-10602-1_48.

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. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Qiang Liu. Rectified flow: A marginal preserving approach to optimal transport. *arXiv preprint arXiv:2209.14577*, 2022. URL https://arxiv.org/abs/2209.14577.

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. URL https://openreview.net/forum?id=XVjTT1nw5z. arXiv:2209.03003.

Harvey Mannering, Zhiwu Huang, and Adam Prugel-Bennett. Noise-level diffusion guidance: Well begun is half done. *arXiv preprint arXiv:2509.13936*, 2025. URL https://arxiv.org/abs/2509.13936. NLG.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182, 2023. doi: 10.1109/ICCV51070.2023.00387. URL https://doi.org/10.1109/ICCV51070.2023.00387.

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

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. URL https://arxiv.org/abs/2204.06125.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, 2022. doi: 10.1109/CVPR52688.2022.01042. URL https://doi.org/10.1109/CVPR52688.2022.01042.

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

Seyedmorteza Sadat, Tobias Vontobel, Farnood Salehi, and Romann M. Weber. Guidance in the frequency domain enables high-fidelity sampling at low CFG scales. *arXiv preprint arXiv:2506.19713*, 2025b. URL https://arxiv.org/abs/2506.19713. Frequency-Decoupled Guidance (FDG).

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 36479–36494, 2022. doi: 10.52202/068431-2643. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/ec795aeadae0b7d230fa35cbaf04c041-Abstract-Conference.html.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2226–2234, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/8a3363abe792db2d8761d6403605aeb7-Abstract.html. Inception Score.

Christoph Schuhmann. Improved aesthetic predictor. GitHub repository, 2022. URL https://github.com/christophschuhmann/improved-aesthetic-predictor.

Dazhong Shen, Guanglu Song, Zeyue Xue, Fu-Yun Wang, and Yu Liu. Rethinking the spatial inconsistency in classifier-free diffusion guidance. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9370–9379, 2024. doi: 10.1109/CVPR52733.2024.00895. URL https://doi.org/10.1109/CVPR52733.2024.00895. S-CFG; arXiv:2404.05384.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, pp. 2256–2265, 2015. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

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)*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Stability AI. Skip-layer guidance in stable diffusion 3.5. https://github.com/Stability-AI/sd3.5, 2024a. SLG; introduced with the Stable Diffusion 3.5 release.

Stability AI. Introducing stable diffusion 3.5. https://stability.ai/news-updates/introducing-stable-diffusion-3-5, 2024b.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models, 2022. https://github.com/huggingface/diffusers.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025. URL https://arxiv.org/abs/2508.02324.

Mengfei Xia, Nan Xue, Yujun Shen, Ran Yi, Tieliang Gong, and Yong-Jin Liu. Rectified diffusion guidance for conditional generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13371–13380, 2025. doi: 10.1109/CVPR52734.2025.01248. URL https://doi.org/10.1109/CVPR52734.2025.01248. ReCFG; arXiv:2410.18737.

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 text-to-image synthesis with linear diffusion transformers. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=N8Oj1XhtYZ.

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)*, volume 36, pp. 15903–15935, 2023. doi: 10.52202/075280-0700. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/33646ef0ed554145eab65f6250fab0c9-Abstract-Conference.html.

Z-Image Team, Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Yuming Jiang, Xin Jin, Liangchen Li, Zhen Li, Zhong-Yu Li, David Liu, Dongyang Liu, Qilong Wu, Feng Yu, Zechao Zhan, Chi Zhang, Shifeng Zhang, Ruikai Zhou, and Shilin Zhou. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025. URL https://arxiv.org/abs/2511.22699.

Kai Zhang, Fujun Luan, Sai Bi, and Jianming Zhang. EP-CFG: Energy-preserving classifier-free guidance. *arXiv preprint arXiv:2412.09966*, 2024. URL https://arxiv.org/abs/2412.09966.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068. URL https://doi.org/10.1109/CVPR.2018.00068. LPIPS; used here as a pairwise diversity metric (mean LPIPS between generations).

Xingyu Zhou, Qifan Li, Xiaobin Hu, Hai Chen, and Shuhang Gu. Guiding a diffusion transformer with the internal dynamics of itself. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11536–11545, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhou_Guiding_a_Diffusion_Transformer_with_the_Internal_Dynamics_of_Itself_CVPR_2026_paper.html. Internal Guidance (IG); arXiv:2512.24176.

A PROPERTIES OF v^+ UNDER CFG SAMPLING

The methodology evaluates $v^+ = M(z_t, t, c)$ on inputs from the CFG-induced trajectory marginal, not from the (idealized) forward process. This appendix collects three properties of v^+ and of the trajectory-aware oracle that justify the operational story in the body: in particular, that the schedule $\beta^\star(t)$ and the CFG weight w decouple in the practical w range. Each property is empirically verified across $w \in \{0, 2, 4, 6, 8\}$ on Flux2-Klein.

Setup: local trajectory oracle. Starting from $z_1 = \epsilon$, the CFG sampler integrates $\dot{z}_s = v_w^{\text{cfg}}(z_s, s, c)$ from $s = 1$ down to $s = 0$, with $v_w^{\text{cfg}} = (1 + w)\, v^+ - w\, v^-$, reaching some z_0^{final} at $s = 0$. At any $t \in (0, 1)$, define the *local trajectory oracle*

$$v_t^{\text{oracle}}(z_t, t) \;=\; \frac{z_t - z_0^{\text{final}}}{t} \;=\; \frac{1}{t}\int_0^t v_w^{\text{cfg}}(z_s, s, c)\, ds. \tag{10}$$

On a perfectly straight forward trajectory, $z_0^{\text{final}} = z_0$ and (10) reduces to the global oracle $v^{\text{oracle}} = \epsilon - z_0$, constant in t; on bent trajectories it varies with t.

Property 1: the oracle's squared norm grows at most quadratically in w. Assume that $\mathbb{E}\|v^+\|^2$ and $\mathbb{E}\|v^-\|^2$ are bounded above by a constant C over the time and guidance range under consideration. Equation (10), Jensen's inequality, and $\|a + b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ give

$$\mathbb{E}\|v_t^{\text{oracle}}\|^2 \;\le\; \frac{1}{t}\int_0^t \mathbb{E}\|v_w^{\text{cfg}}(z_s, s, c)\|^2\, ds \;\le\; 2C\big((1 + w)^2 + w^2\big) \;=\; O((1 + w)^2).$$

This is an upper bound, not a claim of exact quadratic scaling. At $w = 0$ the oracle reduces to the model's own conditional prediction averaged over the trajectory.

Property 2: $\|v^+\|^2$ retains its peak shape across w. $\mathbb{E}\|v^+(z_t, t, c)\|^2$ is unimodal in t with a peak location approximately invariant in w. Re-expressing $v^+ = (z_t - \hat{z}_0(z_t, c))/t$ via the conditional Tweedie estimator $\hat{z}_0(z_t, c) = z_t - t\, v^+$, the norm of v^+ is determined by the model's denoising response on (z_t, c), a (data, model) quantity rather than a sampling-procedure quantity. Hence $\|v^+\|^2$ depends on w only indirectly, through how CFG drifts the latent marginal $p_t(z_t \mid w)$, and not through the per-input value of v^+ itself.

Property 3: CFG-invariance of the alignment-error surface. Define $\sigma_{\text{traj}}(t, w) = \mathbb{E}\|v^+ - v_t^{\text{oracle}}\|^2$. Substituting the Tweedie identity,

$$\sigma_{\text{traj}}(t, w) \;=\; \frac{\mathbb{E}\|\hat{z}_0(z_t, c) - z_0^{\text{final}}\|^2}{t^2}. \tag{11}$$

On the master cache (Appendix G), the bulk of σ_{traj}'s variation across the sweep $w \in \{0, 2, 4, 6, 8\}$ comes from the prefactor $1/t^2$ rather than from $\mathbb{E}\|\hat{z}_0(z_t, c) - z_0^{\text{final}}\|^2$. We treat this as supporting evidence rather than a verified quantitative property and leave a precise plot to future work; importantly, the locally-constant-velocity approximation $\sigma_{\text{naive}} \sim w^2\, \|v^+ - v^-\|^2$ would instead predict quadratic growth in w, which we do not observe.

Self-consistency hypothesis. Both $\hat{z}_0(z_t, c)$ (from the model at the current latent) and z_0^{final} (the trajectory's actual endpoint) approximate the conditional mode $m(c) = \arg\max p(z_0 \mid c)$. Decompose

$$\hat{z}_0 \;=\; m + e_{\text{model}}(z_t, c), \qquad z_0^{\text{final}} \;=\; m + e_{\text{cfg}}(w, c, z_1).$$

The two residuals come from independent sources (model uncertainty at the current latent vs. CFG mode-pull along the integrated trajectory) and are approximately uncorrelated. The first contributes a w-independent term; the second saturates at the conditional-mode basin width as w grows. The leading-order picture is $\sigma_{\text{traj}}(t, w) \approx \sigma_{\text{intrinsic}}(t) + \sigma_{\text{cfg-sat}}(t, w)$, with the second term small and slowly growing across the practical w range.

Implication: the two design knobs decouple. Properties 2 and 3 jointly support the operational story used in the method section. The CFG weight w is tuned for image quality (saturation, contrast, mode-sharpness, exactly as in standard CFG); the schedule $\beta^\star(t)$ is determined per-model from $t^\star$

measured once. The two choices are nearly independent across the practical w range: changing w can change the oracle's magnitude, with at most quadratic growth of its squared norm under the bounded-moment condition of Property 1, but does not substantially shift the peak location of $\|v^+\|^2$ (Property 2) or the alignment-error surface (Property 3).

B WHY LATENT AND VELOCITY NORMS VARY IN t

This appendix formalizes the empirical observations behind the right panel of Figure 2 (the guided velocity norm $\|v^{\mathrm{CFG}}\|$ varies systematically with the sampling time t) and the per-step norm correction of Section 3.4. At $w = 0$, $v^{\mathrm{CFG}} = v^+$, so the conditional-field analysis below applies directly to one of the plotted curves. The latent-norm variation is exact algebra (Lemma 3). The velocity-norm variation is governed by an exact identity (Lemma 4) that ties the squared norm of the conditional velocity field to its inner-product alignment with the oracle, and then inherits its t-dependence from a single named assumption (Assumption 2). All quantities below are taken with respect to the joint distribution over (z_0, ϵ, c) for $z_0 \sim p(\cdot \mid c)$, $\epsilon \sim \mathcal{N}(0, I_d)$ independent of z_0, and c from the prompt distribution.

Setup. We work in the flow-matching framework with the linear schedule $z_t = (1-t)z_0 + t\,\epsilon$, with z_0 a VAE-encoded data point and $\epsilon \sim \mathcal{N}(0, I_d)$ independent of z_0. Let $\mu d := \mathbb{E}\|z_0\|^2$ denote the data's mean squared norm, a (VAE, data) constant computable once per backbone; on Flux2-Klein, $\hat{\mu} \approx 0.97$. Since ϵ is centered Gaussian, $\mathbb{E}\|\epsilon\|^2 = d$ exactly. The oracle velocity is $v^{\mathrm{oracle}} = \epsilon - z_0$ and the model is trained to approximate the conditional expectation $u(z_t, t, c) := \mathbb{E}[v^{\mathrm{oracle}} \mid z_t, c]$.

B.1 LATENT-NORM VARIATION

Lemma 3 (Latent norm is quadratic in t)**.** *For the linear schedule with $z_0 \perp \epsilon$ and $\mathbb{E}\epsilon = 0$,*

$$\mathbb{E}\|z_t\|^2 \;=\; (1-t)^2\,\mu d \;+\; t^2\,d \;=\; d\big[(1-t)^2\mu + t^2\big]. \tag{12}$$

The minimum is attained at $t^\star_{\mathrm{lat}} = \mu/(1+\mu)$, with value $\mu d/(1+\mu)$.

Proof. Expand $\mathbb{E}\|z_t\|^2 = (1-t)^2\,\mathbb{E}\|z_0\|^2 + t^2\,\mathbb{E}\|\epsilon\|^2 + 2t(1-t)\,\mathbb{E}\langle z_0, \epsilon\rangle$. The cross-term vanishes by independence and centered noise: $\mathbb{E}\langle z_0, \epsilon\rangle = \langle \mathbb{E}z_0, \mathbb{E}\epsilon\rangle = 0$. Substituting $\mathbb{E}\|z_0\|^2 = \mu d$ and $\mathbb{E}\|\epsilon\|^2 = d$ gives (12); the minimizer follows from setting $\partial_t \mathbb{E}\|z_t\|^2 = 0$. □

For typical VAE latents μ is close to 1 (Flux2-Klein: $\hat{\mu} = 0.97$, giving $t^\star_{\mathrm{lat}} \approx 0.49$), so the minimum is essentially at the midpoint. Geometrically, the data manifold and the noise sphere lie on shells of comparable radii, $\sqrt{\mu d}$ and $\sqrt{d}$, and the forward process traces the chord between them; $\|z_t\|^2$ is smallest at that chord's midpoint.

B.2 VELOCITY-NORM VARIATION

The conditional velocity field $u(z_t, t, c) = \mathbb{E}[v^{\mathrm{oracle}} \mid z_t, c]$ is a conditional expectation, so it satisfies a Pythagorean identity that immediately ties its squared norm to its alignment with the oracle.

Lemma 4 (Alignment identity)**.** *For every $t \in (0, 1)$,*

$$\mathbb{E}\langle u, v^{\mathrm{oracle}}\rangle \;=\; \mathbb{E}\|u\|^2. \tag{13}$$

Proof. Decompose $v^{\mathrm{oracle}} = u + r$ with $r := v^{\mathrm{oracle}} - u$. By definition of u as a conditional expectation, $\mathbb{E}[r \mid z_t, c] = 0$. Since u is $\sigma(z_t, c)$-measurable, the tower property gives $\mathbb{E}\langle u, r\rangle = \mathbb{E}\big[\langle u, \mathbb{E}[r \mid z_t, c]\rangle\big] = 0$. Therefore $\mathbb{E}\langle u, v^{\mathrm{oracle}}\rangle = \mathbb{E}\|u\|^2 + \mathbb{E}\langle u, r\rangle = \mathbb{E}\|u\|^2$. □

The identity is exact: as functions of t, $\mathbb{E}\|u\|^2$ and $\mathbb{E}\langle u, v^{\mathrm{oracle}}\rangle$ coincide point for point, with no model assumption and no schedule assumption. To pin down the t-shape we make a single named assumption that mirrors Assumption 1 of the body.

Assumption 2 (Unimodal velocity-norm profile)**.** $\mathbb{E}\|u(z_t, t, c)\|^2$ *is unimodal in $t \in (0, 1)$ with a unique interior maximum.*

The motivation is the same SNR trade-off used in Section 3.3: at extreme times the model cannot pin down z_0 and ϵ separately, so its alignment with v^{oracle} (and hence, by Lemma 4, its norm) drops.

Remark 1 (Inner product vs. cosine similarity, and the choice of oracle). The metric used in the body (Assumption 1, left panel of Figure 2) is the cosine similarity

$$\cos\big(v^{\text{CFG}}, v_t^{\text{oracle}}\big) \;=\; \frac{\langle v^{\text{CFG}}, v_t^{\text{oracle}}\rangle}{\|v^{\text{CFG}}\|\,\|v_t^{\text{oracle}}\|},$$

which normalizes out both magnitudes and isolates the directional alignment. The quantity in Lemma 4 is the inner product $\mathbb{E}\langle u, v^{\text{oracle}}\rangle$. The two are deeply related: by the alignment identity the inner product equals $\mathbb{E}\|u\|^2$, while the cosine equals the inner product divided by the two norms. The two metrics differ in three further respects, all of which matter when comparing this appendix to the body:

- *Different predicted velocities.* The body measures the guided velocity $v^{\text{CFG}} = (1+w)v^+ - wv^-$ along CFG trajectories. The identity concerns the conditional field u, approximated by v^+; these agree with the body's velocity only for $w = 0$.
- *Different oracles.* The body uses the *trajectory-aware* oracle $v_t^{\text{oracle}} = (z_t - z_0^{\text{final}})/t$ defined along the actual CFG-bent sampling trajectory (Eq. (10)), so it folds in the model's own bending. The lemma above uses the *global* oracle $v^{\text{oracle}} = \epsilon - z_0$, defined from the original (z_0, ϵ) that initialized the trajectory.
- *Different normalization.* The cosine sees only direction; the inner product is also stretched by $\|u\|$ and $\|v^{\text{oracle}}\|$. Since both norms vary with t, the cosine and the inner product can have different t-profiles even when the underlying directional pattern is the same.

We use the cosine of v^{CFG} in the body because the schedule design in Section 3.3 is directional and is calibrated on the deployed trajectory. We use the conditional-field inner product here because Lemma 4 ties it directly to $\mathbb{E}\|u\|^2$; empirically, the $w = 0$ norm curve in Figure 2 is the corresponding v^+ quantity and motivates the per-step norm correction (Section 3.4).

The oracle norm itself is constant in t:

$$\mathbb{E}\|v^{\text{oracle}}\|^2 \;=\; \mathbb{E}\|\epsilon - z_0\|^2 \;=\; \mathbb{E}\|\epsilon\|^2 + \mathbb{E}\|z_0\|^2 \;=\; (1+\mu)\,d,$$

again by independence and centered noise. So while $\|v^{\text{oracle}}\|^2$ is flat in t, $\|u\|^2$ is unimodal (Assumption 2), and Eq. (13) forces the alignment surface $\mathbb{E}\langle u, v^{\text{oracle}}\rangle$ to share that unimodality exactly.

Corollary 1 (MSE-from-oracle is the exact complement of the squared norm)**.** *Under the model approximation $v^+ \approx u$,*

$$\mathbb{E}\|v^+ - v^{\text{oracle}}\|^2 \;\approx\; (1+\mu)\,d \;-\; \mathbb{E}\|v^+\|^2, \tag{14}$$

i.e., $\mathbb{E}\|v^+\|^2$ and $\mathbb{E}\|v^+ - v^{\text{oracle}}\|^2$ are reflections of each other about the t-independent constant $(1+\mu)d$.

Proof. $\mathbb{E}\|v^+ - v^{\text{oracle}}\|^2 = \mathbb{E}\|v^+\|^2 - 2\mathbb{E}\langle v^+, v^{\text{oracle}}\rangle + \mathbb{E}\|v^{\text{oracle}}\|^2$. With $v^+ \approx u$ and Lemma 4, $\mathbb{E}\langle v^+, v^{\text{oracle}}\rangle \approx \mathbb{E}\|v^+\|^2$, and $\mathbb{E}\|v^{\text{oracle}}\|^2 = (1+\mu)d$, yielding (14). □

Where the model is most aligned with the oracle (peak of $\mathbb{E}\|v^+\|^2$), its MSE-from-oracle is lowest; at the extremes of t, alignment drops and MSE rises in lock step.

Empirical verification. Figure 3 verifies the three predictions on Flux2-Klein: the latent-norm quadratic of Lemma 3, the alignment identity of Lemma 4, and the MSE complement of Corollary 1. *Protocol:* we sample 100 COCO image-prompt pairs, encode each image through the Flux2-Klein VAE to obtain z_0, sample independent Gaussian noise ϵ, and form ideal forward samples $z_t = (1-t)z_0 + t\,\epsilon$ at every t on a sweep grid. We then evaluate $v^+ = M(z_t, t, c)$ and compare to the analytic oracle $v^{\text{oracle}} = \epsilon - z_0$; all quantities are normalized by the latent dimension d.

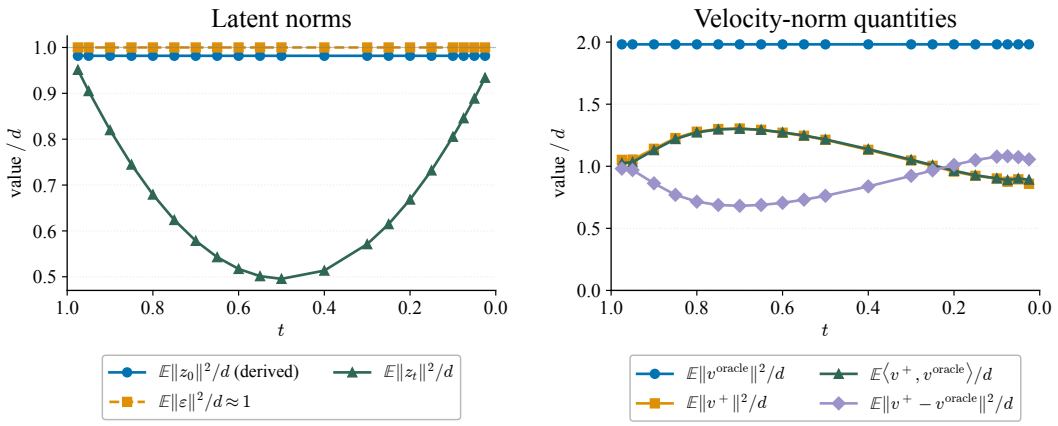


Figure 3: Norm and alignment behavior on Flux2-Klein, evaluated on ideal forward samples $z_t = (1-t)z_0 + t\,\epsilon$ from 100 COCO image-prompt pairs. **(a) Latent norms ($/d$).** The data norm $\|z_0\|^2/d \approx \hat{\mu} \approx 0.97$ and the noise norm $\|\epsilon\|^2/d \approx 1$ are both flat in t; the forward-sample norm $\|z_t\|^2/d$ is the quadratic of Lemma 3 with minimum at $t \approx \mu/(1+\mu) \approx 0.49$. **(b) Velocity norms and alignment ($/d$).** The oracle squared norm $\|v^{\text{oracle}}\|^2/d \approx 1 + \hat{\mu} \approx 1.97$ is flat in t; the conditional-output squared norm $\mathbb{E}\|v^+\|^2/d$ and the inner-product alignment $\mathbb{E}\langle v^+, v^{\text{oracle}}\rangle/d$ collapse onto the same unimodal curve, confirming Lemma 4; the MSE $\mathbb{E}\|v^+ - v^{\text{oracle}}\|^2/d$ is the exact complement (Corollary 1), mirrored about $(1+\hat{\mu}) \approx 1.97$.

C Trajectory straightness and Z-Image diagnostics

TWiG relies on the gap $\|\epsilon - \hat{\epsilon}_{t_p}\|$ remaining controlled along the sampling trajectory: the time-warped proxy of Eq. (5) loses accuracy when the gap grows (Lemma 2). For a perfectly straight forward-process trajectory, $\hat{\epsilon}_t = \epsilon$ exactly. On a sampled trajectory, slowly changing velocities provide indirect evidence that a one-step-cached local estimate remains useful, although consecutive-velocity alignment does not by itself measure the complete proxy error. Rectified-flow training (Liu, 2022; Esser et al., 2024) explicitly encourages straighter conditional flows, placing these models in a favorable regime for this approximation.

We diagnose local trajectory smoothness by measuring the cosine similarity between consecutive sampling-step velocities, $\cos(v_t, v_{t_p})$. A value close to 1 indicates that the tangent direction changes little over one step. The middle panel of Figure 2 reports this on Flux2-Klein and Figure 4 reports it on Z-Image; on both backbones the consecutive-velocity cosine similarity sits close to 1 across the schedule, consistent with the rectified-flow training objective these models are optimized for.

The same Z-Image diagnostic is reported alongside the other two diagnostics from Figure 2 (cosine alignment with the trajectory-aware oracle, and guided-velocity norm), which provide cross-backbone evidence for the empirical regularities discussed in the body: the cosine-alignment peak $\hat{t}^\star \approx 1/2$ on Z-Image too, and $\|v^{\text{CFG}}\|$ varies systematically with t in the same way. In particular, the $w = 0$ trace equals $\|v^+\|$ and motivates the per-step norm correction we apply on both backbones. Protocol for both figures: 100 CFG-sampled images with z_t and v^{CFG} cached at every step; the trajectory-aware oracle (Eq. (10)) is computed from the cached endpoints, and shaded bands report 95% CIs across samples.

D Calibration of $t^\star$ on additional backbones

We calibrated $t^\star$ on three additional backbones spanning distinct architectures: QwenImage-2512 (Wu et al., 2025), SD3.5-medium (Esser et al., 2024; Stability AI, 2024b) (an MMDiT), and Sana-0.6B (Xie et al., 2025) (a linear DiT). For each model we sample 64 CFG trajectories from noise at two guidance strengths, $w = 0$ (conditional-only) and $w = 4$, cache the per-step guided velocity v_t^{CFG}, latent z_t, and trajectory endpoint z_0^{final}, and compute the same statistic used in the body: $\cos(v_t^{\text{CFG}}, v_t^{\text{oracle}})$, where $v_t^{\text{oracle}} = (z_t - z_0^{\text{final}})/t$. The peak $\hat{t}^\star$ is the argmax over the full 50-step grid. Table 3 reports the peaks and the alignment curves on a common t-grid.

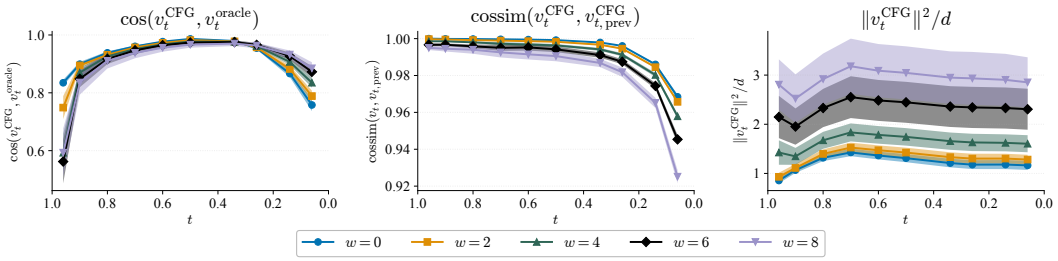


Figure 4: Empirical diagnostics on Z-Image across CFG weights $w \in \{0, 2, 4, 6, 8\}$. Same panels and protocol as Figure 2: **left** cosine alignment of v^{CFG} with the trajectory-aware oracle, **middle** consecutive guided-velocity cosine similarity (trajectory straightness), **right** guided velocity norm $\|v^{\mathrm{CFG}}\|$ as a function of t. The $\hat{t}^\star$ peak falls at the same location $\approx 1/2$ as on Flux2-Klein, and the consecutive-velocity cosine similarity sits close to 1 across the schedule.

Table 3: $t^\star$ on three additional backbones. Cells report $\cos(v_t^{\mathrm{CFG}}, v_t^{\mathrm{oracle}})$ as "w=0 / w=4"; the peak column is the argmax over the full 50-step grid. 64 prompts per configuration.

Model	peak $\hat{t}^\star$	t=0.2	t=0.3	t=0.4	t=0.5	t=0.6	t=0.7	t=0.8
QwenImage-2512	0.542/0.501	0.832/0.903	0.930/0.952	0.979/0.973	0.993/0.976	0.987/0.970	0.965/0.958	0.920/0.934
SD3.5-medium	0.474/0.476	0.976/0.972	0.992/0.987	0.997/0.992	0.997/0.992	0.995/0.990	0.992/0.984	0.987/0.975
Sana-0.6B	0.517/0.494	0.764/0.868	0.894/0.932	0.967/0.964	0.989/0.970	0.977/0.956	0.940/0.919	0.881/0.861

All six peaks fall within $[0.47, 0.55]$, clustering at $t^\star = 1/2$, across three architecturally distinct backbones and two guidance strengths, including $w = 0$. Together with the two backbones of the body, the alignment peak sits at $t^\star \approx 1/2$ on five out of five models tested, supporting Assumption 1 and the use of a single calibration constant.

E Folklore concentration of Gaussian noise

Claim 1 (Gaussian annulus). *For $\epsilon \sim \mathcal{N}(0, I_d)$, the squared norm $\|\epsilon\|^2$ is χ^2_d-distributed and concentrates on its mean: $\|\epsilon\|^2/d \to 1$ in probability with $\mathrm{Var}(\|\epsilon\|^2/d) = 2/d$. In particular $\|\epsilon\|/\sqrt{d} \to 1$ in probability, and a Chernoff bound gives, for any $\delta \in (0, 1)$,*

$$\Pr\left[\|\epsilon\|^2 \notin (1 \pm \delta)\, d\right] \;\le\; 2\exp\!\left(-\tfrac{d}{2}\left(\tfrac{\delta^2}{2} - \tfrac{\delta^3}{3}\right)\right).$$

Proof. $\|\epsilon\|^2 = \sum_{i=1}^{d} \epsilon_i^2 \sim \chi^2_d$. The Chernoff bound is Laurent & Massart (2000): setting $u = \frac{d}{2}(\frac{\delta^2}{2} - \frac{\delta^3}{3})$ in their two-sided χ^2 tail gives the stated bound. □

The geometric reading is that all but exponentially small probability mass lies in a thin spherical shell of radius $\sqrt{d}$, and the shell width tends to zero in relative terms as $d \to \infty$. Combined with $z_0 \perp \epsilon$, $z_t = (1-t)z_0 + t\,\epsilon$ has $\|z_t - (1-t)z_0\| \approx t\sqrt{d}$, so z_t lies near a sphere of radius $t\sqrt{d}$ around $(1-t)z_0$. This is the concentration that underlies Figure 1 of the method section.

F Implementation of baselines and TWiG

This appendix documents the per-step velocity update each method computes inside our `diffusers`-based pipeline. All methods use the same scheduler and 256×256 resolution. All use 50 sampling steps except the explicitly marked CFG-75 matched-compute baseline. We use Diffusers CFG scales $s = 4.5/5.0$, corresponding to $w = s - 1 = 3.5/4.0$ in Eq. (2). We use the same notation as in the body: $v^+ = M(z_t, t, c)$ is the conditional velocity, $v^- = M(z_t, t, \varnothing)$ is the unconditional velocity, and σ refers to the scheduler's noise level at the current step. Table 4 collects the per-method hyperparameters used to produce the numbers in Table 1.

Evaluation protocol. Distribution and quality metrics are computed from 30K 256×256 images per method, generated using 30K MS-COCO captions (Lin et al., 2014) with seed i for the i-th caption; diversity metrics are computed from 100 MS-COCO captions with 20 independent samples per caption per method. GenEval (Ghosh et al., 2023) and GenEval2 (Kamath et al., 2025) are computed at 512×512. For Aesthetic Score, we apply the OpenAI CLIP ViT-L/14 preprocessing, extract and ℓ_2-normalize the CLIP image embedding, and score it with the LAION Improved Aesthetic Predictor checkpoint `sac+logos+ava1-l14-linearMSE.pth` (Schuhmann, 2022). The checkpoint SHA-256 is `21dd590f3ccdc646f0d53120778b296013b096a035a2718c9cb0d511bff0f1e0`. We report the arithmetic mean across all 30K generated images.

Table 4: Per-method hyperparameters in the notation of Eq. (2). "3.5/4.0" indicates the value used on Flux2-Klein / Z-Image, respectively; the corresponding Diffusers CFG scales are $s = 4.5/5.0$.

Method	Hyperparameters
CFG	$w = 3.5/4.0$
CFG-75	$w = 3.5/4.0$, 75 sampling steps
APG	$w = 3.5/4.0$, momentum $\beta_{\text{apg}} = -0.5$, norm threshold $\tau = 15.0$, $\eta_{\text{apg}} = 0$
EP-CFG	$w = 3.5/4.0$, energy percentile band $[0.45, 0.55]$
ISG	$w = 3.5/4.0$, ISG scale $s_{\text{isg}} = 3.5$
NLG	$w = 3.5/4.0$, refinement steps $K = 10$, clip threshold 5.0, step size 0.005
SSG	$w = 3.5/4.0$, SSG scale $s_{\text{ssg}} = 0.5$, block-drop ratio 0.1
PAG	$w = 3.5/4.0$, PAG scale $s_{\text{pag}} = 1.0$
SG	$w = 3.5/4.0$, $s_{\text{sg}} = 3.0$, shifted-noise offset $\Delta\sigma = 0.01$
SG-prev	$w = 3.5/4.0$, $s_{\text{sg}} = 3.0$, active for $\sigma_t < 0.5$
TWiG	$w = 3.5/4.0$, $\eta = 1.5$, $\beta \in [0.8, 1.2]$, $t^\star = 1/2$

CFG (Ho & Salimans, 2021). Standard contrastive update, in the body's parameterization $w \geq 0$:

$$v_t \;=\; v^+ \;+\; w\,(v^+ - v^-).$$

APG (Sadat et al., 2025a). APG decomposes the guidance delta into components parallel and orthogonal to v^+, applies reverse-momentum smoothing, and clips the running delta's norm. With d_{prev} the delta from the previous step,

$$\begin{aligned} d \;&=\; \text{clip}_\tau\big(v^+ - v^- + \beta_{\text{apg}}\, d_{\text{prev}}\big)\,, \quad \tau = 15.0, \\ d_\parallel \;&=\; \frac{\langle d, v^+\rangle}{\|v^+\|^2}\, v^+, \qquad d_\perp \;=\; d - d_\parallel, \\ v_t \;&=\; v^+ + w\,\big(d_\perp + \eta_{\text{apg}}\, d_\parallel\big). \end{aligned}$$

With $\eta_{\text{apg}} = 0$, the parallel "stretching" component is fully suppressed.

EP-CFG (Zhang et al., 2024). EP-CFG forms the standard CFG update and rescales it to match the robust energy of v^+, where the robust energy is the mean squared entry in the central percentile band (between the 45th and 55th percentiles of squared values):

$$v_t \;=\; \big(v^+ + w\,(v^+ - v^-)\big) \cdot \sqrt{\frac{E_{\text{robust}}(v^+)}{E_{\text{robust}}(v^+ + w(v^+ - v^-))}}.$$

ISG (Gu & Hou, 2025). ISG adds a self-contrast against a stochastic re-evaluation of the same conditional branch with dropout active:

$$v_t \;=\; v^+ + w\,(v^+ - v^-) \;+\; s_{\text{isg}}\,(v^+ - v^{\text{stoch}}), \quad v^{\text{stoch}} = M_{\text{train}}(z_t, t, c).$$

NLG (Mannering et al., 2025). NLG runs a pre-loop refinement that nudges the initial latent in the direction of the CFG delta, then runs standard CFG denoising. With $K = 10$ refinement iterations at the noise level σ_0, clip threshold 5.0, and step size 0.005:

1: **for** $k = 1, \ldots, K$ **do**

2: $d \leftarrow \mathrm{clip}_{5.0}\big(v^+(z_T, \sigma_0) - v^-(z_T, \sigma_0)\big)$
3: $z_T \leftarrow z_T - 0.005\, d + \xi, \quad \xi \sim \mathcal{N}(0, 0.001^2 I)$
4: **end for**
5: Run standard CFG denoising starting from the refined z_T.

SSG (Chen et al., 2026). SSG contrasts the conditional branch against a structurally degraded variant in which a random subset of blocks is zeroed:

$$v_t \;=\; v^+ + w\,(v^+ - v^-) \;-\; s_{\mathrm{ssg}}\,(v^{\mathrm{drop}} - v^+), \quad v^{\mathrm{drop}} = M_{\mathrm{drop}\,0.1}(z_t, t, c),$$

where $M_{\mathrm{drop}\,0.1}$ randomly zeroes 10% of the blocks in the forward pass.

PAG (Ahn et al., 2024). PAG contrasts the conditional branch against a structure-blind variant in which self-attention maps are replaced by the identity:

$$v_t \;=\; v^+ + w\,(v^+ - v^-) \;+\; s_{\mathrm{pag}}\,(v^+ - v^{\mathrm{pag}}), \quad v^{\mathrm{pag}} = M_{\mathrm{id-attn}}(z_t, t, c).$$

Self-Guidance (Li et al., 2024b). Let $g_t = v^+ + w(v^+ - v^-)$ denote the standard CFG update. We implement SG as

$$v_t^{\mathrm{SG}} \;=\; g_t + s_{\mathrm{sg}}\big(v^+ - v^{+,\mathrm{shift}}\big), \qquad v^{+,\mathrm{shift}} = M(z_t, t_{\mathrm{shift}}, c).$$

The auxiliary branch uses the same latent and prompt embeddings at a noisier time. We set $s_{\mathrm{sg}} = 3.0$ and $\sigma_{\mathrm{shift}} = \min(\sigma_t + 0.01, 1)$. On Flux2-Klein this is $\tau_{\mathrm{shift}} = \min(\tau + 10, 1000)$ on the model's timestep scale from 0 to 1000, equal to the official `shift_t=10`; Z-Image receives $1 - \sigma_{\mathrm{shift}}$ as its time input. The Euler update retains the original scheduler time. SG therefore adds one conditional-only pass, for three transformer evaluations per step and 150 total evaluations over 50 steps.

SG-prev removes this extra pass by caching the conditional prediction after every step and using

$$v_t^{\text{SG-prev}} \;=\; g_t + \mathbf{1}\{\sigma_t < 0.5\}\, s_{\mathrm{sg}}\big(v_t^+ - v_{t_p}^+\big).$$

The correction is active only in the cleaner second half of the trajectory, corresponding to the official `sg_prev_max_t=500` on the scale from 0 to 1000. The first step is a no-op. Because the cache is updated even while the correction is inactive, its first active step uses the immediately preceding conditional prediction. SG-prev uses the same two evaluations per step as CFG, for 100 total evaluations. We use $s_{\mathrm{sg}} = 3.0$, $\Delta\sigma = 0.01$, and the 0.5 activation threshold for all reported SG and SG-prev experiments.

TWiG (ours). TWiG augments the standard CFG combination with a second contrastive term, $\eta\,(v^+ - v^\beta)$, where v^β is the conditional branch evaluated at a time-warped query point built from the cached Tweedie estimate $\hat{\epsilon}$. Per step, with $\hat{\epsilon}$ initialized to z_{t_0}:

1: $v^+ \leftarrow M(z_t, t, c), \;\; v^- \leftarrow M(z_t, t, \varnothing)$
2: $\beta \leftarrow \mathrm{clip}\big((1-t^\star)/(1-t),\ \beta_{\min},\ \beta_{\max}\big), \;\; t_\beta \leftarrow 1 - \beta(1-t)$
3: $z_{t_\beta} \leftarrow \beta\, z_t - (\beta - 1)\,\hat{\epsilon}$ ▷ TWiG query point
4: $v^\beta_{\mathrm{raw}} \leftarrow M(z_{t_\beta}, t_\beta, c)$
5: $v^\beta \leftarrow v^\beta_{\mathrm{raw}} \cdot \|v^+\| / \|v^\beta_{\mathrm{raw}}\|$ ▷ per-sample norm correction
6: $v_t \leftarrow v^+ + w\,(v^+ - v^-) + \eta\,(v^+ - v^\beta)$
7: $\hat{\epsilon} \leftarrow z_t + (1-t)\,v^+$ ▷ Tweedie cache for the next step
8: $z_t \leftarrow z_t + (t_{i+1} - t_i)\,v_t$ ▷ Euler step

This is exactly the body recipe of Algorithm 1; the target $t^\star$ is calibrated once per backbone, the clip bounds $(\beta_{\min}, \beta_{\max})$ are selected once, and the new sampling-time scalar is η.

G Master cache for empirical validation

The master cache used for the empirical validation is built from 500 MS-COCO image-prompt pairs with 8 random seeds per pair, evaluated at a 20-point time grid and a 7-point β grid. For every (prompt, seed, t, β) tuple, the cache stores the conditional output $v^+ = M(z_t, t, c)$, the unconditional

output $v^- = M(z_t, t, \varnothing)$, the estimated matched query $M(z_{t_\beta}, t_\beta, c)$ where $z_{t_\beta} = \beta z_t - (\beta - 1)\hat{\epsilon}_{t_p}$ with $\hat{\epsilon}_{t_p}$ the cached local Tweedie estimate at the previous step (Eq. (4)), the radially scaled query $M(\beta z_t, t_\beta, c)$ (used as a baseline against which the estimated matched query is compared), the ideal forward query $M(z^\star_{t_\beta}, t_\beta, c)$ (computable here because the ground-truth z_0 and ϵ are available), the noise ϵ, and the oracle velocity $v^{\text{oracle}} = \epsilon - z_0$.

The build script enforces three sanity checks before the full sweep, on a small 4-pair, 1-seed, 3-time subset:

1. At $\beta = 1$, $z^\star_{t_\beta} = z_t$ and the estimated query reduces to z_t (within $1\text{e}{-4}$ relative tolerance after fp16 round-trip).
2. For all (β, t), $\beta z_t - z^\star_{t_\beta} = (\beta - 1)\,\epsilon$ elementwise (Lemma 1).
3. Repeated runs of v^+ and v^- on identical inputs are deterministic.

The time grid concentrates near $t = 0$: $t \in \{0.025, 0.05, 0.075, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 0.975\}$. The β grid is $\{0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3\}$.

H COMPUTE ACCOUNTING

Table 5 reports the per-step number of function evaluations (NFE) and measured end-to-end wall-clock time on both backbones at 256×256, batch size 1, and a single A100. Methods use 50 steps except CFG-75, which uses 75. Across the repeated timing runs ($N = 100$ per method and backbone), sec/image is mean$\pm$std and the ratio to CFG includes a Fieller 95% confidence interval. Every measurement includes text encoding, denoising, and VAE decoding.

Table 5: Per-step NFE and end-to-end wall-clock time per image, measured at 256×256 on a single A100. Methods use 50 steps except CFG-75, which uses 75; sec/image is mean$\pm$std over 100 runs, and ratio intervals are Fieller 95% CIs.

		Flux2-Klein-4B-Base		*Z-Image*	
Method	NFE/step	sec/image	×CFG [95% CI]	sec/image	×CFG [95% CI]
CFG	2	3.78 ± 0.004	1.000 (ref)	4.41 ± 0.007	1.000 (ref)
CFG-75	2	5.60 ± 0.004	1.480 [1.480, 1.480]	6.80 ± 0.006	1.542 [1.541, 1.542]
EP-CFG	2	3.75 ± 0.020	0.992 [0.991, 0.993]	4.44 ± 0.003	1.007 [1.007, 1.007]
APG	2	3.82 ± 0.005	1.008 [1.008, 1.009]	4.48 ± 0.041	1.015 [1.014, 1.017]
NLG	2.4	4.59 ± 0.004	1.212 [1.211, 1.213]	5.29 ± 0.020	1.199 [1.198, 1.200]
PAG	3	5.14 ± 0.022	1.359 [1.357, 1.360]	6.10 ± 0.015	1.383 [1.382, 1.384]
TWiG	3	5.22 ± 0.006	1.380 [1.379, 1.380]	6.18 ± 0.010	1.400 [1.399, 1.400]
SG	3	5.55 ± 0.006	1.467 [1.466, 1.467]	7.66 ± 0.011	1.737 [1.736, 1.737]
SG-prev	2	3.80 ± 0.005	1.004 [1.004, 1.005]	5.17 ± 0.008	1.172 [1.172, 1.173]
SSG	3	5.49 ± 0.025	1.451 [1.449, 1.452]	6.68 ± 0.012	1.514 [1.514, 1.515]
ISG	3	5.49 ± 0.007	1.452 [1.451, 1.452]	6.87 ± 0.016	1.557 [1.556, 1.558]

The two-pass methods CFG, APG, EP-CFG, and SG-prev require 100 model evaluations at 50 steps. APG and EP-CFG sit near $1.0\times$ CFG on both backbones; SG-prev measures $1.004\times$ and $1.172\times$, with the difference arising from caching and correction overhead. TWiG and SG each use three evaluations per step, or 150 in total. TWiG costs $1.380\times$ CFG on Flux2-Klein and $1.400\times$ on Z-Image, while SG costs $1.467\times$ and $1.737\times$. PAG has similar cost to TWiG at $1.359\times$ and $1.383\times$; SSG and ISG cost $1.451/1.514\times$ and $1.452/1.557\times$, respectively, reflecting hook or dropout overhead on top of the third pass. The 75-step CFG baseline also uses 150 evaluations and costs $1.480\times$ and $1.542\times$. NLG costs $1.212\times$ and $1.199\times$ despite using two evaluations per denoising step because its 10 pre-denoising refinement iterations raise the total to 120 evaluations. Raising NLG to 150 evaluations would move it beyond its recommended operating point, where performance degrades, so we use its published configuration.

I QUANTITATIVE ABLATIONS WITH CONFIDENCE INTERVALS

This appendix repeats the ablations of Table 2 with 95% confidence intervals ($1.96\,\sigma/\sqrt{N}$, $N = 5{,}000$ COCO prompts per configuration, identical prompt and seed set across configurations, Flux2-Klein at 256×256 with 50 steps). Bold marks the paper's configuration.

Table 6: Diffusers CFG scale $s \times \eta$ grid with 95% CIs: Image Reward (top) and CLIP (bottom).

IR	η=0 (CFG)	η=0.5	η=1	η=1.5	η=2
s=3	0.695 ± 0.0218	0.713 ± 0.0312	0.721 ± 0.0218	0.725 ± 0.0229	0.721 ± 0.0218
s=4.5	0.876 ± 0.0233	0.880 ± 0.0218	0.879 ± 0.0216	$\mathbf{0.882 \pm 0.0215}$	0.878 ± 0.0220
s=6	0.830 ± 0.0230	0.835 ± 0.0215	0.838 ± 0.0218	0.844 ± 0.0220	0.839 ± 0.0218
CLIP	η=0 (CFG)	η=0.5	η=1	η=1.5	η=2
s=3	26.65 ± 0.1009	26.72 ± 0.1014	26.75 ± 0.1011	26.80 ± 0.1013	26.58 ± 0.1014
s=4.5	26.93 ± 0.1022	26.99 ± 0.1013	27.00 ± 0.1005	$\mathbf{27.03 \pm 0.1011}$	26.91 ± 0.1009
s=6	27.06 ± 0.1011	27.06 ± 0.1011	27.11 ± 0.1016	27.13 ± 0.1010	27.03 ± 0.1013

Table 7: $\beta_{\min}/\beta_{\max}$ grid with 95% CIs: Image Reward (top) and CLIP (bottom).

IR	$\beta_{\max}$=1	$\beta_{\max}$=1.2	$\beta_{\max}$=1.4	$\beta_{\max}$=1.6	$\beta_{\max}$=1.8
$\beta_{\min}$=1	0.876 ± 0.0233	0.879 ± 0.0216	0.865 ± 0.0230	0.802 ± 0.0229	0.777 ± 0.0230
$\beta_{\min}$=0.9	0.877 ± 0.0219	0.880 ± 0.0218	0.867 ± 0.0219	0.785 ± 0.0218	0.750 ± 0.0219
$\beta_{\min}$=0.8	0.879 ± 0.0216	$\mathbf{0.882 \pm 0.0215}$	0.870 ± 0.0215	0.744 ± 0.0215	0.728 ± 0.0215
$\beta_{\min}$=0.7	0.848 ± 0.0219	0.866 ± 0.0311	0.851 ± 0.0219	0.712 ± 0.0219	0.702 ± 0.0222
$\beta_{\min}$=0.6	0.809 ± 0.0312	0.796 ± 0.0308	0.725 ± 0.0229	0.688 ± 0.0219	0.678 ± 0.0220
CLIP	$\beta_{\max}$=1	$\beta_{\max}$=1.2	$\beta_{\max}$=1.4	$\beta_{\max}$=1.6	$\beta_{\max}$=1.8
$\beta_{\min}$=1	26.93 ± 0.1022	26.95 ± 0.1014	26.64 ± 0.1009	26.60 ± 0.1016	26.02 ± 0.1018
$\beta_{\min}$=0.9	26.98 ± 0.1016	26.99 ± 0.1013	26.65 ± 0.1009	26.57 ± 0.1010	26.04 ± 0.1013
$\beta_{\min}$=0.8	27.00 ± 0.1005	$\mathbf{27.03 \pm 0.1011}$	26.61 ± 0.1011	26.57 ± 0.1010	26.03 ± 0.1009
$\beta_{\min}$=0.7	26.97 ± 0.1016	26.54 ± 0.1010	26.03 ± 0.1009	25.95 ± 0.1014	25.87 ± 0.1010
$\beta_{\min}$=0.6	26.88 ± 0.1009	26.13 ± 0.1010	25.99 ± 0.1009	25.91 ± 0.1016	25.84 ± 0.1010

Table 8: Norm correction (left) and CFG-then-TWiG warm start (right) with 95% CIs.

	IR	CLIP
Off	0.828 ± 0.0220	26.92 ± 0.1010
On (TWiG)	$\mathbf{0.882 \pm 0.0215}$	$\mathbf{27.03 \pm 0.1011}$

TWiG start step	IR	CLIP
K=0 (TWiG)	$\mathbf{0.882 \pm 0.0215}$	$\mathbf{27.03 \pm 0.1011}$
K=5	0.881 ± 0.0230	27.02 ± 0.1010
K=15	0.879 ± 0.0216	26.98 ± 0.1016
K=25	0.878 ± 0.0220	26.94 ± 0.1013
K=50 (CFG)	0.876 ± 0.0233	26.93 ± 0.1022

J AN INTUITION FOR HOW THE SECOND CONTRAST CAN INCREASE DIVERSITY

TWiG involves no stochasticity: given the prompt and the seed, the sampler is fully deterministic, and the only randomness is the shared initial latent $\epsilon = z_1$, exactly as in CFG. Diversity is therefore a property of how the deterministic velocity field maps the seed distribution to the output distribution, and the two contrastive terms of Eq. (6) act on it very differently. Consider the forward ensemble $z_t \sim p_t(\cdot \mid c)$ at each t, and recall $u(z_t, t, c) = \mathbb{E}[\,\epsilon - z_0 \mid z_t, c\,]$ with unconditional counterpart $u(z_t, t) = \mathbb{E}[\,\epsilon - z_0 \mid z_t\,]$. The CFG term has a non-zero conditional mean, $\mathbb{E}[\,v^+ - v^- \mid c\,] \neq 0$ in general: its mean is the average conditional-versus-unconditional direction, a vector field that pulls every seed of the prompt toward its high-density regions. This shared drift is exactly the well-documented diversity cost of raising w (Ho & Salimans, 2021; Cideron et al., 2025). The TWiG term instead has zero mean over the ensemble. By Lemma 1, the exact matched query $z^\star_{t_\beta} = (1 - t_\beta)\, z_0 + t_\beta\, \epsilon$ is a forward-process sample at time t_β of the same pair (z_0, ϵ), so the exact probe is $v^\beta = \mathbb{E}[\,\epsilon - z_0 \mid z^\star_{t_\beta}, c\,]$, and the tower property gives

$$\mathbb{E}\big[\,v^+ - v^\beta \mid c\,\big] \;=\; \mathbb{E}[\,\epsilon - z_0 \mid c\,] - \mathbb{E}[\,\epsilon - z_0 \mid c\,] \;=\; 0 \qquad \text{for every } t. \tag{15}$$

Both branches are conditional expectations of the same target under the same joint law, conditioned on two observations of the same underlying pair. Thus, for the exact unnormalized probe under the forward ensemble, the second contrast introduces no shared mean drift and acts only through seed-dependent fluctuations. This observation offers a possible mechanism for changing diversity without the common directional pull associated with CFG, but zero mean alone does not guarantee increased dispersion. Moreover, the deployed method uses a norm-corrected proxy z_{t_β} rather than the exact probe, so Eq. (15) should be read as intuition rather than as a guarantee for the implemented sampler. The proxy deviation is controlled by Lemma 2 and the schedule clip, while the direction and magnitude of the empirical diversity change are established by the experiments.

K ANATOMY OF THE TIME-WARPED PROBE: AWAY VS. TOWARD

A contrastive guidance signal can be useful when it is informative yet degraded: informative, so that the contrast $v^+ - v^\beta$ follows a meaningful direction rather than an arbitrary perturbation; degraded, so that moving away from it can favor the better prediction. This is the premise of autoguidance, where a weaker network supplies the degraded prediction. TWiG instead constructs a structured self-contrast by querying the same conditional model at an approximately matched latent-time point. This appendix tests this interpretation and the sign of the resulting update on 100 generated Flux2-Klein trajectories with v_t^+ and the local trajectory oracle v_t^{oracle} cached at every step.

The probe is informative but degraded. We evaluate the probe at two query points: the exact matched forward sample $z^\star_{t_\beta}$ (teacher-forced; velocity $v_t^{\beta,\mathrm{clean}}$), and the estimated query z_{t_β} used during sampling (velocity $v_t^{\beta,\mathrm{est}}$). Table 9 reports three quantities: $A = \cos(v_t^{\beta,\mathrm{est}}, v_t^+)$, how close the probe is to the conditional prediction; $B = \cos(v_t^{\beta,\mathrm{clean}}, v_t^{\mathrm{oracle}}) - \cos(v_t^+, v_t^{\mathrm{oracle}})$, the probe's alignment advantage over v^+ when the query is exact; and C, the same advantage with the estimated query. The probe remains informative: $A \geq 0.86$ everywhere and exceeds 0.95 for $t \geq 0.25$. The exact probe is slightly better aligned with the oracle ($B > 0$), whereas the deployed cached proxy is slightly worse aligned ($C \leq 0$). Thus, the useful away direction is an empirical property of the deployed approximation, not a consequence of the matched-signal identity alone.

Table 9: Character of the time-warped probe on real trajectories (100 trajectories; mean±95% CI). A: cosine of the deployed probe to v^+. B/C: the probe's alignment advantage over v^+ against the trajectory oracle, with the exact/estimated query.

t	A	B	C
0.05 (data end)	0.861	$+0.039 \pm 0.008$	-0.075 ± 0.003
0.15	0.923	$+0.053 \pm 0.004$	-0.029 ± 0.003
0.25	0.959	$+0.045 \pm 0.002$	-0.013 ± 0.002
0.35	0.977	$+0.034 \pm 0.002$	-0.006 ± 0.001
0.50 ($t^\star$)	1.000	0.000	0.000
0.65	0.996	$+0.008 \pm 0.001$	-0.006 ± 0.000
0.80	0.997	$+0.008 \pm 0.001$	-0.003 ± 0.000
0.90 (noise end)	0.996	$+0.013 \pm 0.002$	-0.005 ± 0.001

Away improves the update; toward degrades it. On the same trajectories, we measure how each sign of the TWiG term changes the alignment of the executed update with the oracle,

$$D(\eta) \;=\; \cos\big(v^+ + \eta\,(v^+ - v^\beta),\; v_t^{\mathrm{oracle}}\big) \;-\; \cos\big(v^+,\; v_t^{\mathrm{oracle}}\big),$$

where $D(1.5)$ is TWiG's actual update (away from the probe) and $D(-1.5)$ flips the sign (toward the probe). Table 10 shows that away improves alignment and toward degrades it at every reported step, particularly at the early, high-noise steps. The effect vanishes identically at $t^\star$, where $\beta = 1$ and the probe coincides with v^+, and increases away from $t^\star$, consistent with the proposed interpretation.

The estimation error is a systematic signal, not noise. For the cached estimate used by TWiG, Lemma 2 gives

$$\big\|z_{t_\beta} - z^\star_{t_\beta}\big\| \;=\; |\beta - 1|\,\|\hat{\epsilon}_{t_p} - \epsilon\|.$$

Table 10: Effect of the sign of the TWiG term on the executed update's alignment with the trajectory oracle (mean±95% CI over 100 trajectories). Rows are the sampler's discretization steps.

t (noise→data)	$D(1.5)$ (away, ours)	$D(-1.5)$ (toward)
0.90 (noise)	$\mathbf{+0.024 \pm 0.001}$	-0.033 ± 0.002
0.79	$\mathbf{+0.017 \pm 0.001}$	-0.027 ± 0.001
0.62	$\mathbf{+0.007 \pm 0.001}$	-0.025 ± 0.001
0.50 ($= t^\star$)	0 (identically; $\beta = 1$)	0
0.31	$\mathbf{+0.021 \pm 0.002}$	-0.109 ± 0.006
0.17 (data)	$\mathbf{+0.097 \pm 0.007}$	-0.244 ± 0.011

The displacement therefore reflects the combined approximation and one-step cache error. This magnitude identity makes no claim about which direction is beneficial. The schedule clip bounds its prefactor, and at $\beta = 1$ the same estimation error produces no displacement. To test whether the resulting contrast is robust to perturbations at a representative model-error scale, we recompute the probe with $\hat{\epsilon}$ perturbed by random vectors whose norm is set, per step, to $(1-t)\,\|v^+ - v_t^{\mathrm{oracle}}\|$, and measure how much of the contrast $\delta = v^+ - v^\beta$ survives averaging over the replicas,

$$\mathrm{sysfrac} \;=\; \frac{\left\| \frac{1}{4}(\delta + \delta_1 + \delta_2 + \delta_3) \right\|^2}{\frac{1}{4}\left(\|\delta\|^2 + \|\delta_1\|^2 + \|\delta_2\|^2 + \|\delta_3\|^2\right)}.$$

A perturbation-invariant signal gives 1, while independent equal-scale perturbations give about $1/4$ for an average of four. We measure $\mathrm{sysfrac} = 0.964$, indicating that the contrast is largely preserved under perturbations at this scale. This statistic diagnoses structure, not which sign is beneficial; that question is tested directly by Table 10. Consistently, widening the clip or raising η injects more of the signal, while quality peaks at moderate settings and falls beyond them (Appendix I).

L Compositional benchmarks: GenEval and GenEval2

This appendix reports the per-task breakdowns behind the GenEval/2 columns of Table 1. Setup: both backbones at 512×512 (the GenEval detector is unreliable at 256×256), with the same hyperparameters as Table 1. GenEval (Ghosh et al., 2023): 553 prompts $\times$ 4 seeds (2,212 images per method), official aggregation. GenEval2 (Kamath et al., 2025): 800 prompts per method, scored with soft-TIFA (Hu et al., 2023) using a Qwen3-VL-8B judge (Bai et al., 2025). TWiG has the highest overall GenEval point estimate on both backbones and the highest GenEval2 point estimate on Flux2-Klein; it ties SG-prev on Z-Image GenEval2. The per-task results show that no method leads every category consistently. Together with Table 1, these results suggest that TWiG can improve quality and diversity without sacrificing compositional alignment.

Table 11: GenEval overall score from 0 to 1 and per-task accuracy in percent on Flux2-Klein (top) and Z-Image (bottom). **Bold** = best, underline = second best per column, computed within each backbone.

Method	Overall	Single obj.	Two obj.	Counting	Colors	Position	Color attr.
Flux2-Klein-4B-Base							
CFG	0.779	99.1	89.1	**77.5**	89.9	49.5	62.3
CFG-75	**0.782**	99.4	87.4	76.9	**91.5**	**50.8**	63.3
PAG	0.723	98.1	85.1	75.6	82.5	48.3	44.5
NLG	0.773	**99.7**	87.4	71.9	89.6	50.5	64.5
ISG	0.762	98.3	88.7	74.5	85.4	48.3	63.5
SSG	0.749	99.1	84.9	76.3	85.4	48.8	55.0
SG	0.759	99.1	88.1	71.6	89.9	47.5	59.0
SG-prev	0.781	99.4	88.9	**77.5**	90.4	49.0	63.3
APG	0.613	97.8	65.4	49.4	79.3	37.3	38.8
EP-CFG	0.772	99.4	**90.2**	71.6	88.6	48.8	**64.8**
TWiG	**0.782**	99.4	89.1	76.3	90.2	50.3	63.8
Z-Image							
CFG	0.693	98.8	86.9	60.9	83.5	33.8	51.8
CFG-75	0.701	98.8	86.4	**63.8**	82.7	**34.5**	54.8
PAG	0.465	82.8	55.1	46.3	55.6	21.8	17.5
NLG	0.689	97.8	86.9	62.5	81.1	32.5	52.8
ISG	0.674	98.1	81.6	62.5	82.2	30.3	49.5
SSG	0.676	98.4	85.9	60.9	81.1	32.5	47.0
SG	0.674	96.6	83.6	59.1	81.9	30.8	52.5
SG-prev	0.682	96.9	82.6	61.6	81.7	33.8	52.5
APG	0.631	95.9	77.5	50.3	77.9	28.5	48.3
EP-CFG	0.698	**99.1**	86.4	61.6	84.6	33.8	53.8
TWiG	**0.704**	98.1	**87.4**	63.4	**85.1**	31.5	**56.8**

Table 12: GenEval2 overall score from 0 to 1 and per-atom-type accuracy in percent (soft-TIFA, Qwen3-VL-8B judge) on Flux2-Klein (top) and Z-Image (bottom). **Bold** = best, underline = second best per column, computed within each backbone.

Method	Overall	Object	Attribute	Count	Position	Verb
Flux2-Klein-4B-Base						
CFG	0.672	82.9	74.3	47.7	58.6	16.6
CFG-75	0.666	82.2	74.2	47.3	57.5	16.0
PAG	0.624	77.8	67.1	44.3	53.1	13.3
NLG	0.672	82.7	74.0	**48.3**	58.7	15.9
ISG	0.629	79.9	70.3	44.8	52.3	10.2
SSG	0.631	78.3	69.2	45.3	53.2	10.6
SG	0.661	81.9	73.4	45.8	58.9	13.7
SG-prev	0.672	82.7	73.7	48.0	58.9	**19.0**
APG	0.575	71.6	68.6	35.7	52.3	9.4
EP-CFG	0.664	82.3	74.2	46.8	57.1	14.1
TWiG	**0.676**	**83.4**	**75.1**	47.8	**60.1**	18.1
Z-Image						
CFG	0.752	91.7	85.0	58.8	58.4	23.3
CFG-75	0.755	91.7	84.6	59.7	57.9	24.6
PAG	0.600	71.9	61.8	49.6	48.9	22.5
NLG	0.762	**92.6**	85.7	60.1	59.9	27.9
ISG	0.742	89.0	82.5	59.2	56.7	27.7
SSG	0.749	90.8	83.9	59.2	58.6	29.4
SG	0.746	91.9	84.5	57.8	56.3	21.5
SG-prev	**0.764**	92.0	**85.9**	60.3	**61.3**	29.9
APG	0.707	87.5	81.8	52.2	56.6	19.8
EP-CFG	0.758	91.8	85.1	60.1	59.0	23.7
TWiG	**0.764**	92.2	85.8	**60.4**	59.8	**31.3**

M EXTENDED QUALITATIVE COMPARISON

Figure 5 compares TWiG against the autoguidance baselines on a fixed prompt and seed set: TWiG produces sharper textural detail, more faithful prompt adherence, and more diverse compositional structure than the baselines.

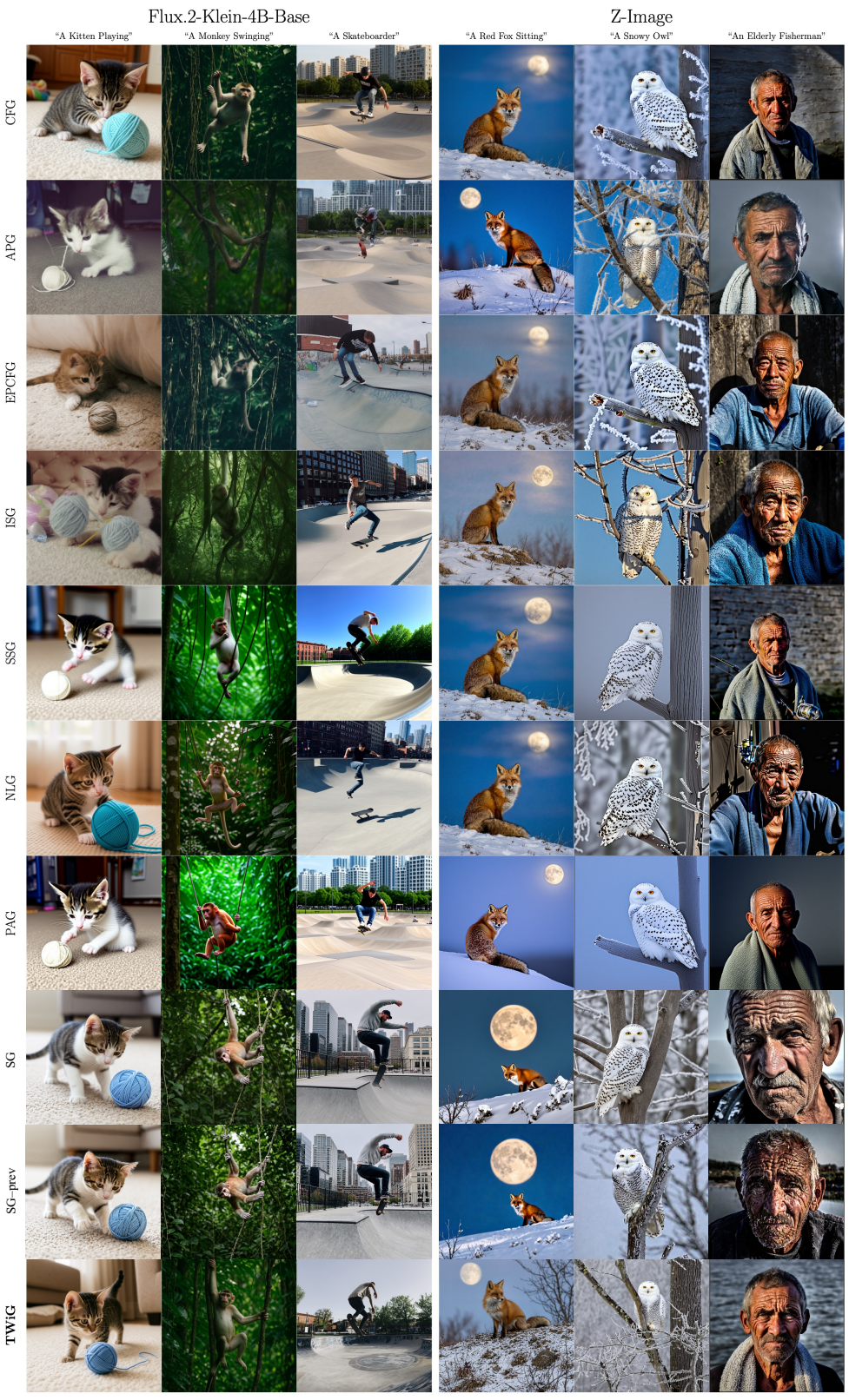


Figure 5: Qualitative comparison across prompts and backbones. Rows: prompts; columns: methods, with **TWiG** (ours) on the rightmost column. TWiG matches or exceeds the baselines on visual quality and prompt adherence at matched per-step cost.

N USER STUDY: DESIGN AND FULL OUTCOME RATES

For every method included in the user study, we generated 400 images using 100 MS-COCO prompts $\times$ 4 seeds, with prompts and seeds shared across methods. The analysis contains 19 participants and 1,949 randomized pair records; study assignments were organized around 100 comparisons. Each comparison requested one judgment for overall image quality and one for prompt alignment, giving 3,898 axis-level judgments across the two backbones. In every pair, one image was generated by TWiG and the other by a baseline method drawn uniformly at random; the prompt and seed were also drawn uniformly. Participants did not know which image came from which method and could select image A, image B, or a tie on each axis.

Table 13 reports the aggregate outcome rates for TWiG versus each baseline on both backbones. Descriptively, the share of pairs in which TWiG was preferred exceeds the share in which the baseline was preferred in every cell.

The clustered significance analysis reported in Section 4 treats each participant as a cluster. Let b_k, c_k, and n_k denote the TWiG-preferred, baseline-preferred, and total pair counts for participant k, respectively; ties contribute to n_k but to neither b_k nor c_k. With the participant-normalized difference $q_k = (b_k - c_k)/n_k$, the variance-corrected statistic (Durkalski et al., 2003) is

$$\chi^2 \;=\; \frac{\left(\sum_k q_k\right)^2}{\sum_k q_k^2}.$$

It is χ^2_1 under the null and allows arbitrary within-participant correlation. Applied to the study responses, it gives $\chi^2 = 5.16$ ($p = 0.023$) for quality and $\chi^2 = 9.49$ ($p = 0.002$) for prompt alignment.

Table 13: Pairwise user-study outcome rates for **TWiG vs. each baseline**, on the axes of image *Quality* and *Prompt Alignment*. Win and loss rates are percentages with descriptive 95% intervals from the binomial normal approximation; tie rates, not displayed, are $100\% - \mathrm{Win}\% - \mathrm{Loss}\%$. The formal pooled test uses participant-normalized win-loss differences and treats each participant as a cluster. Bold marks the larger point estimate within each win-loss pair.

	Flux2-Klein-4B-Base				*Z-Image*			
	Quality		Prompt Align.		Quality		Prompt Align.	
Baseline	Win↑	Loss↓	Win↑	Loss↓	Win↑	Loss↓	Win↑	Loss↓
CFG	**38.1**±9.3	25.7±8.4	**6.7**±4.8	2.9±3.2	**40.3**±8.6	33.9±8.3	**8.1**±4.8	3.2±3.1
PAG	**70.8**±8.4	17.7±7.0	**15.0**±6.6	3.5±3.4	**60.3**±8.9	19.0±7.1	**12.1**±5.9	2.6±2.9
NLG	**34.5**±8.5	25.2±7.8	**5.9**±4.2	3.4±3.2	**43.6**±8.4	27.8±7.6	**6.8**±4.3	6.0±4.0
ISG	**43.1**±8.8	27.6±7.9	**8.9**±5.0	4.9±3.8	**41.0**±9.4	30.5±8.8	**9.5**±5.6	8.6±5.4
SSG	**41.7**±8.8	32.5±8.4	**8.3**±4.9	4.2±3.6	**42.6**±9.0	33.0±8.6	**3.5**±3.3	2.6±2.9
APG	**48.3**±9.1	36.2±8.7	**7.8**±4.9	4.3±3.7	**33.3**±8.5	28.2±8.2	**9.4**±5.3	5.1±4.0
EP-CFG	**37.0**±8.1	31.1±7.8	**5.2**±3.7	3.0±2.9	**44.0**±9.3	35.8±9.0	**6.4**±4.6	4.6±3.9
SG	**37.9**±10.2	32.2±9.8	**6.9**±5.3	5.7±4.9	**39.4**±11.4	33.8±11.0	**9.9**±6.9	5.6±5.4
SG-prev	**34.3**±11.4	26.9±10.6	**4.5**±5.0	1.5±2.9	**36.5**±11.0	27.0±10.1	**8.1**±6.2	2.7±3.7

O USER STUDY SCREENS

Figures 6–8 show the three screens shown to participants of the blinded pairwise user study: the welcome screen, the instructions screen, and the per-pair question screen.

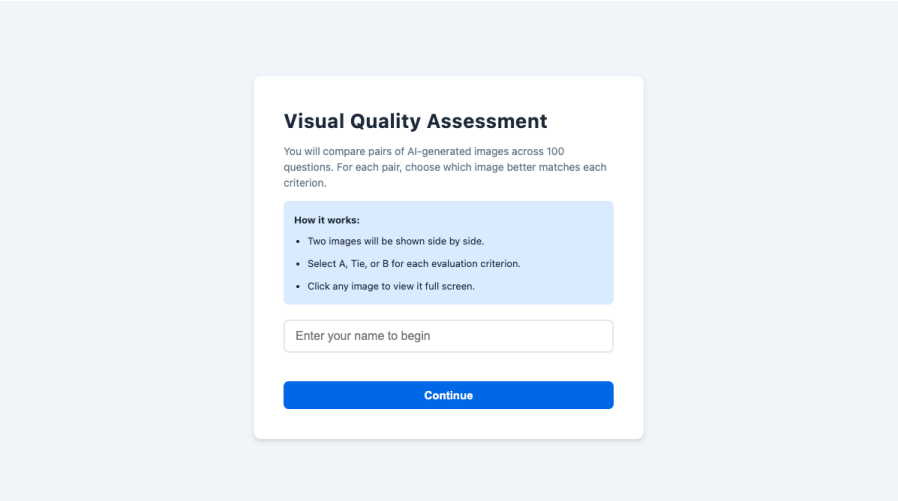


Figure 6: Welcome screen shown to user-study participants.

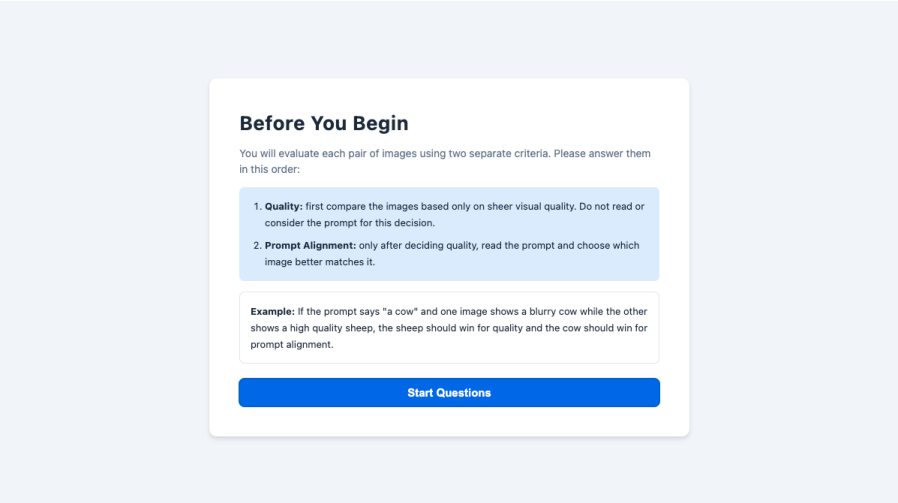


Figure 7: Instructions screen shown to user-study participants.

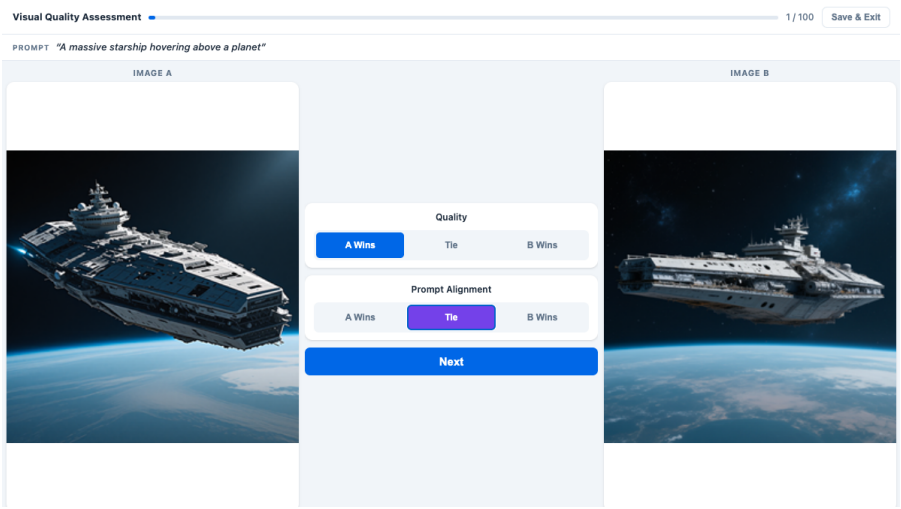


Figure 8: Per-pair question screen shown to user-study participants. Study assignments were organized around 100 such pairs, ranked on two axes (overall image quality and prompt alignment).