Constraint-Aware Receding-Horizon Classifier-Free Guidance for Attribute-Conditioned Flow Matching

Mu Chen

Abstract

Fixed-scale classifier-free guidance (CFG) is an open-loop policy for a state- and time-varying generation process. We start from two structural facts. First, once conditional and unconditional velocities are computed, all scalar CFG actions lie on one affine line. Second, a straight Flow Matching path maps each candidate velocity to an analytic clean-endpoint estimate. Together these facts turn scale choice into a cheap finite control problem. We instantiate this view as Constraint-Aware Receding-Horizon CFG (CRH-CFG): a training-free controller that repeatedly forecasts candidate endpoints and selects the smallest temporally smooth action satisfying evaluator-based target and protected-attribute constraints. The frozen generator and 50-step Heun solver remain unchanged, and every method uses 100 guided-field evaluations (200 raw U-Net forwards) per trajectory. On three CelebA attributes with 1,000 paired seeds each, CRH-CFG changes mean target success from 0.9840 to 0.9880 with 1.80% mean wall-time overhead. This $+0.4$-point change is descriptive because per-seed outcome transitions were not retained. Common-reference preservation does not improve, and KID to the inherited full-marginal reference increases for all targets but cannot identify conditional quality. The evidence therefore supports low-overhead adaptive correction, not a joint improvement in adherence, preservation, and quality.

1. Introduction

Flow Matching learns continuous transport by regressing a time-dependent velocity field along a prescribed probability path [10]. Classifier-free guidance (CFG) sharpens conditional generation by affinely combining conditional and unconditional predictions [7, 17]. As with classifier-guided diffusion [4], stronger guidance can improve semantic alignment while sacrificing fidelity or diversity.

The usual fixed scale is structurally mismatched to the process it controls. It applies one open-loop intervention to every initial noise and every stage, even though a sample may already satisfy the condition or may become difficult only later. Prior work on restricted intervals, condition annealing, early suppression, and online feedback confirms that guidance demand is neither uniform in time nor across samples [5, 8, 12, 13].

Our starting point is the field structure itself. A conditional/unconditional pair exposes only one scalar control direction, $g_s = v_c - v_u$, but it generates every ordinary CFG candidate by affine recombination. For a straight path, each candidate also yields an analytic prediction of its clean endpoint. Thus, one expensive field pair can support many cheap, evaluator-scored actions. The same equations expose the limitation: no selector can reach a desired effect outside that one-dimensional action set.

CRH-CFG turns this geometry into feedback. At selected states, it predicts a finite bank of endpoints, filters candidates using target and protected- attribute constraints, and chooses the least intervention with a temporal smoothness penalty. An early reliability gate and a finite fallback make the policy operational, while holding the selected scale across each Heun pair keeps generator work exactly matched.

Our contributions are threefold: (1) a first-principles control view linking the CFG affine line to straight-path endpoint prediction; (2) a vectorized, training-free controller with explicit target, preservation, and minimum- intervention terms; and (3) a paired, test-isolated evaluation that reports both the small adherence point estimate and the observed evidence boundary, rather than treating proxy feasibility as a quality guarantee.

2. Related Work

Flow Matching and classifier-free guidance. Flow Matching trains continuous normalizing flows by vector-field regression along prescribed probability paths [10]. Guided Flows transfers classifier and classifier-free guidance to flow-based generation [17]. Standard CFG uses affine extrapolation of conditional and unconditional predictions [7]. CRH-CFG retains this frozen field and changes only the scalar selection policy.

When and how much to guide. Limited-interval guidance [8], CFG-Zero* [5], and condition annealing [13] modify when or how strongly a condition acts; rescaling is also

coupled to the noise schedule [9]. Online feedback selects per-step scales with auxiliary evaluators [12]. CRH-CFG is closest to the latter, but its action set is exactly the ordinary CFG affine line, its look-ahead comes from an analytic flow endpoint, and its selector explicitly penalizes intervention while monitoring non-target attributes.

Guidance geometry and cost. Strong extrapolation can leave learned trajectories. CFG++ uses a manifold-motivated correction [3], Rectified-CFG++ adapts this view to flow models [14], and manifold projection directly targets Flow Matching scale sensitivity [2]. Other work changes training to remove the two-branch inference cost [15]. We make no manifold guarantee and do not retrain: the original scalar direction is preserved, so reachability is one-dimensional and evaluator overhead is separated from generator NFE.

3. From Field Structure to Feedback

Straight-path endpoint geometry. Let $s \in [0,1]$ increase from Gaussian noise x_1 to clean data x_0 along the straight path

$$x_s = (1-s)x_1 + sx_0, \qquad \hat{x}_0(v) = x_s + (1-s)v. \quad (1)$$

For the ideal bridge, $v = x_0 - x_1$ makes the second relation exact. For a learned field it is an analytic endpoint prediction, not a guarantee. The paper/native clock conversion and derivation are given in Appendix A.

One affine control axis. For condition c and null condition $\varnothing$, define $v_c = v_\theta(x_s, s, c)$, $v_u = v_\theta(x_s, s, \varnothing)$, and $g_s = v_c - v_u$. Ordinary CFG satisfies

$$v_w = v_u + w(v_c - v_u) = v_c + (w-1)g_s. \quad (2)$$

With $a_s = w_s - 1$, the sampler is control-affine,

$$\frac{dY_s}{ds} = v_c(Y_s, s) + g_s(Y_s)a_s, \quad (3)$$

and Eq. (1) maps its complete action line to

$$\hat{x}_0(w) = \hat{x}_0(1) + (1-s)(w-1)g_s. \quad (4)$$

Thus, one conditional/unconditional pair defines every scalar action and its predicted endpoint; it also fixes the local reachability boundary.

Least intervention under constraints. Because $w = 1$ is the ordinary conditional field, we prefer actions near one and near the previous decision, subject to a target margin and a protected- attribute budget. Re-observing the state after integration and solving this finite problem again yields receding-horizon feedback without multi-step generator rollouts.

4. Method: Constraint-Aware Receding-Horizon CFG

4.1. Candidate Endpoint Bank

At each decision, the candidate set is $\mathcal{W} = \{0.5, 1.0, 1.5, 2.0, 3.0, 4.0\}$. Conditional and unconditional fields are evaluated once, after which candidate velocities and endpoints are constructed in parallel:

$$v_{w_m} = v_u + w_m(v_c - v_u), \quad w_m \in \mathcal{W}, \quad (5)$$

$$\hat{x}_0(w_m) = x_s + (1-s)v_{w_m}. \quad (6)$$

The frozen controller evaluator scores all endpoints in one batch. For target a, desired sign y_a, and protected set P_a, we use

$$M_a(w,s) = y_a F_{\mathrm{ctrl},a}(\hat{x}_0(w)), \quad (7)$$

$$D_{\mathrm{nt}}(w,s) = \frac{1}{|P_a|}\sum_{j\in P_a} \left|F_{\mathrm{ctrl},j}(\hat{x}_0(w)) - F_{\mathrm{ctrl},j}(\hat{x}_0(1))\right|. \quad (8)$$

The second quantity is a local attribute-logit proxy relative to $w = 1$, not an identity or perceptual-similarity guarantee.

4.2. Constraint-Aware Minimum Intervention

The feasible action set is

$$\mathcal{A}_s = \{w \in \mathcal{W} : M_a(w,s) \geq m(s), \quad D_{\mathrm{nt}}(w,s) \leq \epsilon(s)\}. \quad (9)$$

If $\mathcal{A}_s$ is nonempty, the controller selects

$$w_s^\star = \arg\min_{w\in\mathcal{A}_s} \rho(w-1)^2 + \eta(w - w_{\mathrm{prev}})^2. \quad (10)$$

This is the central design rule: satisfy the online proxies, then use the least change from ordinary conditional generation and the previous action. If the feasible set is empty, a finite hinge-penalty objective returns the least-violating candidate and records the violated constraints. Full schedules, normalization, and the fallback are in Appendix B.

4.3. Receding-Horizon Execution

Validation selected the default scale $w = 2$ for $s < \tau_{\mathrm{on}} = 0.1$, where predicted endpoints are least reliable. Thereafter, the controller is evaluated every five Heun macro-steps and the previous scale is held between decisions. The chosen $w_s^\star$ is held across the Heun predictor–corrector pair; only then is x_{s+h} observed and the problem replanned. Static CFG and CRH-CFG therefore both execute 100 guided fields, or 200 conditional/ unconditional U-Net forwards, per trajectory. CRH-CFG adds ten batched evaluator decisions, but no generator trajectories. The exact Heun update, algorithm, and complexity are in Appendix B.

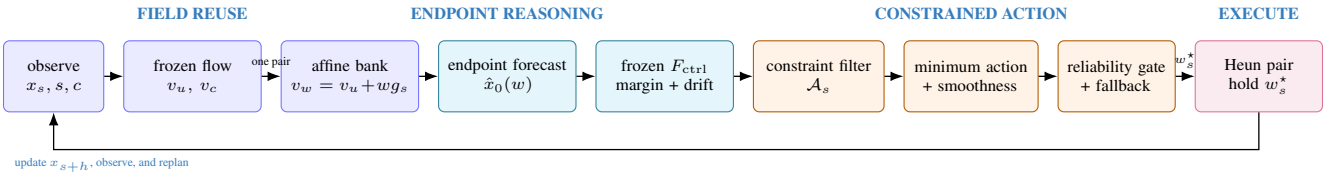

Figure 1. CRH-CFG follows the derivation from left to right: reuse one frozen field pair, forecast the affine candidate endpoints, select a feasible minimum intervention, execute it, and re-observe. Candidate expansion and endpoint scoring add no generator calls. F_{eval} is outside this loop and is used only for final measurement.

5. Experimental Protocol

Frozen generator and data. We freeze the 18.5M-parameter, 64×64 attribute-conditioned U-Net and its 50,000-iteration EMA checkpoint. The inherited CelebA subset contains 63,715 images [11]; a deterministic split isolates controller training, validation, and final test data. We evaluate positive Smiling, Bangs, and Blond Hair requests. Architecture and split counts are in Appendix C.

Methods and paired protocol. The five methods are static $w = 2$, validation-selected interval CFG, target-only feedback, target-plus-preservation feedback, and full CRH-CFG. Every comparison shares checkpoint, EMA weights, conditions, initial noises, and the 50-step Heun solver. Each target uses the same 1,000 seeds across methods; ablations use the first 256. Controller schedules and penalties are frozen on validation data and listed in Appendix C.

Evaluators and metrics. F_{ctrl} selects actions; a separately fine-tuned F_{eval} measures final target success and protected-logit drift. Both use ImageNet-pretrained ResNet-18 features [6], so independent heads and fine-tuning do not eliminate correlated representation error. CelebA annotations are also imperfect [16]. A common, same-seed $w = 1$ reference makes preservation comparisons meaningful. KID [1] uses the inherited full CelebA marginal; it measures closeness to that marginal, not conditional-target quality. Evaluator accuracy, KID sampling details, and metric definitions are in the appendix.

6. Results

6.1. Main Comparison

Table 1 shows a consistent but small positive adherence point estimate. Full CRH-CFG changes the unweighted mean success from 0.9840 to 0.9880, or $+0.4$ percentage points. Target-wise counts change from 985/1,000 to 987/1,000 for Smiling, 980 to 984 for Bangs, and 987 to 993 for Blond Hair. Because the artifact retains aggregate counts but not per-seed success transitions, McNemar or paired-bootstrap uncertainty cannot be reconstructed. These values are descriptive and do not establish a statistically reliable or large practical improvement.

Closeness to the inherited full-marginal reference moves in the opposite direction. Relative to static $w = 2$, full CRH-CFG changes KID_{marg} by $+0.00001$, $+0.00013$, and $+0.00211$ for Smiling, Bangs, and Blond Hair. The first two changes are small relative to the reported subset standard deviations, which are not confidence intervals for method differences. Since the reference mixes positive and negative examples, these increases do not identify a decline in conditional quality. The main hypothesis is nevertheless unsupported because common-reference preservation does not improve; conditional quality remains unresolved under this KID reference.

The frozen-feature diversity results reinforce the same trade-off (Table 2). Target-only control is more diverse for every attribute because it selects substantially lower scales. Full CRH-CFG remains close to static CFG for Smiling and Bangs and is less diverse for Blond Hair (.2579 to .2515, a 2.46% relative decrease). Its unweighted mean changes from .3532 to .3507. This feature-space proxy has no confidence interval, so it is descriptive rather than inferential. No method is uniformly better on observed success, marginal-reference KID, protected drift, diversity, and wall time.

6.2. A Common Preservation Reference

Baseline-relative drift in Table 1 cannot establish that static CFG preserves the ordinary conditional solution because its own drift is zero by construction. Table 3 instead compares all methods with paired $w = 1$ trajectories and averages only after retaining target-wise records.

Full CRH-CFG is slightly farther from $w = 1$ than static CFG on both common reference metrics. Target-only and target-plus-preservation stay substantially closer, but lose 7.73 mean success points. The selected controller therefore exposes a target–preservation trade-off rather than a preservation win.

6.3. Component and Hyperparameter Evidence

Table 4 yields three conclusions. First, feedback without the full regularized package moves toward $w = 1$ and sacrifices adherence. Second, adding the preservation term changes no

Table 1. Main results on 1,000 paired seeds per target. All methods use the same frozen generator and 50-step Heun solver. Drift and flip are relative to the paired static $w = 2$ result; $\mathrm{KID}_{\mathrm{marg}}$ is mean $\pm$ subset standard deviation against the inherited full-marginal reference. All rows use 100 guided-field evaluations.

Target	Method	Succ. ↑	Δ pp	Drift ↓	Flip ↓	$\mathrm{KID}_{\mathrm{marg}}$ ↓	$\bar{w}$	Δ time
Smiling	Static $w = 2$	.985	+0.0	.0000	.0000	.01223±.00063	2.0000	+0.00%
	Interval	.979	-0.6	.1169	.0037	.01173±.00060	–	-0.50%
	Target only	.860	-12.5	1.5417	.0498	.00904±.00050	1.0739	+2.30%
	Target + pres.	.860	-12.5	1.5417	.0498	.00904±.00050	1.0739	+2.54%
	Full CRH-CFG	**.987**	**+0.2**	.0025	.0003	.01223±.00063	2.0067	+2.70%
Bangs	Static $w = 2$	.980	+0.0	.0000	.0000	.04110±.00162	2.0000	+0.00%
	Interval	.972	-0.8	.2707	.0093	.04087±.00161	–	-0.07%
	Target only	.888	-9.2	3.1288	.1258	.02379±.00126	1.1952	+1.61%
	Target + pres.	.888	-9.2	3.1288	.1258	.02379±.00126	1.1952	+1.97%
	Full CRH-CFG	**.984**	**+0.4**	.0476	.0018	.04123±.00162	2.0480	+1.58%
Blond	Static $w = 2$	.987	+0.0	.0000	.0000	.06872±.00225	2.0000	+0.00%
	Interval	.986	-0.1	.3156	.0095	.06711±.00221	–	+0.07%
	Target only	.972	-1.5	1.7170	.0555	.05884±.00206	1.2571	+1.47%
	Target + pres.	.972	-1.5	1.7170	.0555	.05884±.00206	1.2571	+1.29%
	Full CRH-CFG	**.993**	**+0.6**	.2858	.0070	.07083±.00229	2.1689	+1.12%

Figure 2. Paired Blond Hair generations for the first 16 frozen seeds. Rows are static $w = 2$, interval CFG, target-only control, target-plus-preservation, and full CRH-CFG. The two reduced controllers are visually and numerically identical, while the full controller remains close to static CFG. This grid is seed-ordered and not selected by image quality.

Table 2. Mean cosine distance between paired random samples in frozen F_{eval} backbone features. Higher denotes greater feature diversity. Target-only and target-plus-preservation are identical.

Target	Static	Interval	Target	Full
Smiling	.3729	.3750	.4383	.3724
Bangs	.4290	.4344	.4858	.4281
Blond	.2579	.2637	.2885	.2515

Table 3. Unweighted mean preservation diagnostic against common, same-seed $w = 1$ outputs. Lower drift and flip imply closer protected attributes, but can coincide with lower target success.

Method	Success ↑	Drift ↓	Flip ↓
Static $w = 2$	.9840	2.7770	.0963
Interval	.9790	2.7596	.0973
Target / target+pres.	.9067	1.1067	.0318
Full CRH-CFG	.9880	2.8433	.0980

reported value, both with and without minimum intervention. The selected budget is therefore non-binding in these variants. Third, removing the early reliability gate drops mean success from 0.9883 to 0.9154; this comparison validates the full package, but does not isolate smoothness from every simultaneous design change.

The preregistered $w_{\max}$ sweep is consistent with limited extra control authority. Increasing $w_{\max}$ from 2 to 3 changes success from 0.9857 to 0.9883 while increasing drift from 0.0004 to 0.0775 and reduced KID from 0.04152 to 0.04215. Raising it to 4 leaves success at 0.9883. Sweeping the endpoint parameter $\epsilon(1) \in \{0.25, 0.5, 1.0\}$, corresponding to actual last-decision budgets $\epsilon(.9) \in \{.425, .650, 1.100\}$, produces identical aggregate values. This further confirms that the preservation constraint does not determine the chosen

Table 4. Component ablation on 256 seeds per target, reported as unweighted target means. KID uses the reduced 100-image/50-subset full-marginal-reference protocol and is not comparable with Table 1.

Variant	Succ.	Drift	Flip	KID
Static $w = 2$	.9857	.0000	.0000	.04152
Target feedback	.8503	2.9184	.1104	.02595
Target + pres.	.8503	2.9184	.1104	.02595
Target + min. int.	.9063	2.0954	.0749	.03070
Target + pres. + min.	.9063	2.0954	.0749	.03070
Full package	.9883	.0833	.0036	.04224
Full, no gate	.9154	1.8887	.0677	.03480
Full, fixed threshold	.9870	.5143	.0169	.04476

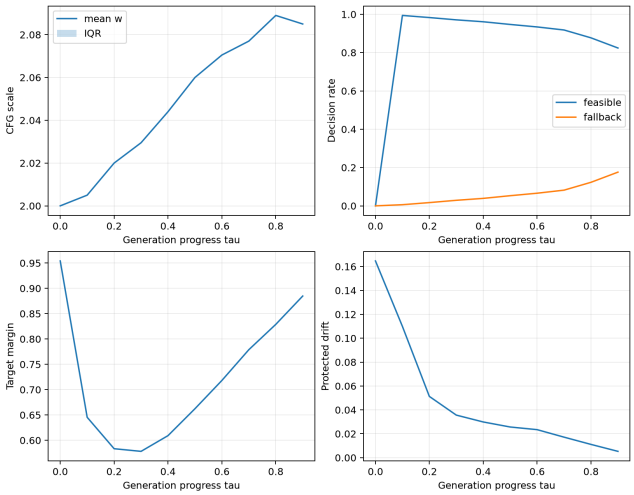

Figure 3. Full-controller dynamics for Bangs over 1,000 trajectories. The median and interquartile scale remain 2, while the fallback rate rises late in sampling. This is the target with the highest final fallback rate.

actions.

6.4. Controller Dynamics and Failures

The controller is adaptive for a minority of difficult samples, but the population policy remains close to static CFG. At $s = 0.9$, mean scales are 2.0170, 2.0850, and 2.1855 for Smiling, Bangs, and Blond Hair; the median and interquartile scale remain 2 at every saved decision. Final fallback rates are 0.057, 0.176, and 0.016, with feasible rates 0.943, 0.824, and 0.984. Bangs is thus the least frequently feasible target, whereas Blond Hair receives the strongest late intervention and exhibits both the largest positive adherence point estimate and the largest increase in marginal-reference KID.

The rule-selected failures clarify the method boundary. For Smiling seed 410577, the static and full signed margins are -10.291 and -10.139, and the controller falls back at 30% of decisions despite reaching $w = 3$. For Bangs seed 410550, fallback reaches 80%, the maximum selected scale

Table 5. Controller state at the final saved decision $s = 0.9$. Although mean scale differs by target, all scale medians and interquartile ranges equal 2.

Target	Mean w	Feasible	Fallback
Smiling	2.0170	.943	.057
Bangs	2.0850	.824	.176
Blond Hair	2.1855	.984	.016

remains 2, and the final margin worsens from -6.201 to -7.777. For Blond Hair seed 410327, the controller reaches $w = 4$ and improves the margin from -7.205 to -5.250 without crossing the decision boundary. These cases show insufficient scalar control authority and imperfect transfer from predicted endpoints scored by F_{ctrl} to final outcomes scored by F_{eval}.

6.5. Cost

Full CRH-CFG matches generator work exactly and increases mean wall time by 1.80% across targets. Its candidate tensors and evaluator batches raise peak memory from 644,830,208 to 673,739,776 bytes, an increase of 27.6 MiB. The measured HW4 study, including calibration, main experiments, ablations, and analysis, consumed 3.824 L40S GPU-hours. The retained HW3 timestamps imply an additional 2.6875 GPU-hours for generator training, explicitly an estimate rather than a billing record.

The artifact audit checks exact seed coverage, sample and trajectory counts, finite metric values, frozen checkpoint hashes, the deterministic split hash, and agreement between aggregate CSV rows and paper tables. Nineteen automated tests cover time conversion, endpoint prediction, candidate vectorization, fallback behavior, Heun scale holding, paired-seed reproducibility, split isolation, checkpoint freezing, and matched generator NFE. These checks do not validate the semantic proxy, but they separate implementation failures from the empirical limitations observed above.

7. Discussion and Limitations

What is supported. The full policy is a conservative correction around $w = 2$, not a broadly new sampler. Smoothness and the early gate prevent the collapse toward $w = 1$ seen in reduced variants, giving a small positive adherence point estimate at matched generator work. The non-binding preservation term explains why the full policy inherits static CFG's preservation behavior; the study therefore supports low-overhead adaptive correction, not the intended joint benefit.

Structural and proxy limits. Equation (3) gives only one local control direction. If every point on the affine line violates a constraint, replanning cannot create new authority.

(a) Smiling (b) Bangs (c) Blond Hair

Figure 4. Rule-selected paired cases. Within each panel, columns are static $w = 2$, interval CFG, target-only control, and full CRH-CFG; rows are the maximum-margin success, the case closest to the independent evaluator boundary, and the minimum-margin failure. Selection uses frozen statistics rather than visual preference.

Endpoint estimates are also least reliable near noise, and both online constraints depend on F_{ctrl}. The separate evaluator reduces direct reuse but shares architecture and pretrained features. Useful extensions are uncertainty-aware evaluator ensembles and independently validated control directions; both require new matched-compute baselines.

Evidence limits. Across 3,000 trajectories, static CFG records 2,952 successes and CRH-CFG records 2,964. Because per-seed outcome transitions were not retained, this net difference cannot support a paired significance test. KID subset deviations are not confidence intervals, and the marginal reference cannot identify conditional quality. The results describe one checkpoint, dataset, resolution, solver, and three positive requests rather than generalization across models or data regimes.

Scope and responsible use. Protected-logit drift measures selected attributes, not identity, perceptual similarity, fairness, or consent. CelebA labels are noisy and correlated [16]; appearance labels can encode social assumptions. The method should not be interpreted as reliable real-person editing or sensitive-attribute control.

8. Conclusion

Starting from the CFG affine control axis and the straight-path endpoint map, we derived CRH-CFG as finite, minimum-intervention feedback for a frozen flow. It preserves solver and generator evaluation counts and adds 1.80% mean wall time. Across three paired 1,000-sample targets, mean success changes by a descriptive $+0.4$ points, while common-reference preservation does not improve and conditional quality remains unresolved by marginal-reference KID. The method is therefore a low-overhead adaptive correction with an explicit one-dimensional reachability boundary, not evidence for a universal joint win.

References

[1] Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. 3

[2] Jian-Feng Cai, Haixia Liu, Zhengyi Su, and Chao Wang. Improving classifier-free guidance of flow matching via manifold projection. *arXiv preprint arXiv:2601.21892*, 2026. 2

[3] Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. CFG++: Manifold-constrained classifier free guidance for diffusion models. *arXiv preprint arXiv:2406.08070*, 2024. 2

[4] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, 2021. 1

[5] Weichen Fan, Amber Yijia Zheng, Raymond A. Yeh, and Ziwei Liu. CFG-Zero*: Improved classifier-free guidance for flow matching models. *arXiv preprint arXiv:2503.18886*, 2025. 1

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3

[7] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1

[8] 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. *arXiv preprint arXiv:2404.07724*, 2024. 1

[9] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. *arXiv preprint arXiv:2305.08891*, 2023. 2

[10] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. 1

[11] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision*, 2015. 3

[12] Pinelopi Papalampidi, Olivia Wiles, Ira Ktena, Aleksandar Shtedritski, Emanuele Bugliarello, Ivana Kajić, Isabela Albuquerque, and Aida Nematzadeh. Dynamic classifier-free diffusion guidance via online feedback. *arXiv preprint arXiv:2509.16131*, 2025. 1, 2

[13] Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. CADS: Unleashing the diversity of diffusion models through condition-annealed sampling. In *International Conference on Learning Representations*, 2024. 1

[14] Shreshth Saini, Shashank Gupta, and Alan C. Bovik. Rectified-CFG++ for flow based models. *arXiv preprint arXiv:2510.07631*, 2025. 2

[15] Zhicong Tang, Jianmin Bao, Dong Chen, and Baining Guo. Diffusion models without classifier-free guidance. *arXiv preprint arXiv:2502.12154*, 2025. 2

[16] Haiyu Wu, Grace Bezold, Manuel Günther, Terrance Boult, Michael C. King, and Kevin W. Bowyer. Consistency and accuracy of CelebA attribute values. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, 2023. 3, 6

[17] Qinqing Zheng, Matt Le, Neta Shaul, Yaron Lipman, Aditya Grover, and Ricky T. Q. Chen. Guided flows for generative modeling and decision making. *arXiv preprint arXiv:2311.13443*, 2023. 1

A. Time Conventions and Endpoint Derivation

The paper-time convention runs from data to noise,

$$x_t = (1-t)x_0 + tx_1, \qquad t = 0 \text{ data}, \quad t = 1 \text{ noise}. \tag{11}$$

Its ideal bridge velocity is $u_t = x_1 - x_0$, so

$$x_t - tu_t = (1-t)x_0 + tx_1 - t(x_1 - x_0) = x_0. \tag{12}$$

The implementation uses the native generation clock $s = 1 - t$ and the reverse velocity $v_s = -u_t = x_0 - x_1$. Substitution gives

$$\hat{x}_0 = x_s + (1-s)v_s = x_t - tu_t. \tag{13}$$

This equality is exact for the ideal straight bridge. With a learned velocity, the same expression is a one-step endpoint forecast.

For CFG, $v_w - v_1 = (w-1)(v_c - v_u) = (w-1)g_s$. The native endpoint map then implies

$$\hat{x}_0(w) - \hat{x}_0(1) = (1-s)(v_w - v_1) \tag{14}$$

$$= (1-s)(w-1)g_s, \tag{15}$$

which yields Eq. (4). This derivation explains both the computational reuse and the one-dimensional reachability limit; it does not assert that evaluator scores vary monotonically with w.

B. Controller and Solver Details

Schedules and normalization. Each controller-evaluator margin is divided by its validation-only standard deviation, preserving its zero decision boundary. The target threshold and protected budget are

$$m(s) = m_{\text{early}} + (m_{\text{late}} - m_{\text{early}})s^{\gamma}, \tag{16}$$

$$\epsilon(s) = \epsilon_{\text{early}} + (\epsilon_{\text{late}} - \epsilon_{\text{early}})s. \tag{17}$$

Decisions occur at $s \in \{0, .1, \ldots, .9\}$. Therefore, the "late" values parameterize the schedule endpoint at $s = 1$ rather than an observed decision. For the frozen configuration, the final decision uses $m(.9) = .567$ and $\epsilon(.9) = .650$.

Fallback. When $\mathcal{A}_s$ is empty, CRH-CFG minimizes the finite objective

$$J_s(w) = \alpha\,[m(s) - M_a(w,s)]_+ + \lambda\,[D_{\text{nt}}(w,s) - \epsilon(s)]_+ + \rho(w-1)^2 + \eta(w - w_{\text{prev}})^2, \qquad w \in \mathcal{W}. \tag{18}$$

Because $\mathcal{W}$ is finite and nonempty, the fallback always returns a candidate. The implementation separately logs target, preservation, and joint violations.

Heun integration. With $h = 1/50$, predictor and corrector share the selected action:

$$\tilde{x}_{s+h} = x_s + h\, v_{w_s^\star}(x_s, s), \tag{19}$$

$$x_{s+h} = x_s + \frac{h}{2}\left[v_{w_s^\star}(x_s, s) + v_{w_s^\star}(\tilde{x}_{s+h}, s+h)\right]. \tag{20}$$

Holding $w_s^\star$ avoids changing the numerical protocol inside a macro-step.

Algorithm 1 CRH-CFG with Heun Protocol B

Require: Frozen θ, state x, condition c, candidate set $\mathcal{W}$, evaluator F_{ctrl}

$w_{\text{prev}} \leftarrow 1$
for $k = 0, \ldots, 49$ **do**
  Compute (v_u, v_c) once at the current state
  if k is a controller decision **then**
    Construct and score all endpoints by Eqs. (5)–(6)
    Select $w^\star$ by Eq. (10) or Eq. (18)
  else
    $w^\star \leftarrow w_{\text{prev}}$
  end if
  Apply one Heun step, holding $w^\star$ for the corrector
  $w_{\text{prev}} \leftarrow w^\star$
end for
return x

For batch size B and $M = |\mathcal{W}|$, candidate endpoints form an $M \times B \times C \times H \times W$ tensor, clipped to $[-1, 1]$ and evaluated in one batch. Construction costs $O(MBCHW)$ tensor work plus one controller evaluator pass over MB images per decision. It does not launch an additional generator trajectory.

C. Experimental Details

The frozen conditional U-Net has base width 64, channel multipliers $[1, 2, 2, 4]$, two residual blocks per level, attention at 16×16, four attention heads, and 18,535,619 parameters. The seed-42 data permutation fixes 50,972 training, 6,371 validation, and 6,372 final-test images. Main experiments use seeds 410000–410999 for each target and method; ablations use the first 256.

Table 6. Validation-frozen controller configuration.

Component	Value
Candidate scales	$\{0.5, 1, 1.5, 2, 3, 4\}$
Target schedule	$m(0) = -.5$, $m(1) = .75$, $\gamma = 1.5$
Drift budget	$\epsilon(0) = 2.0$, $\epsilon(1) = .5$, linear
Penalties	$(\alpha, \lambda, \rho, \eta) = (1, 1, .05, .20)$
Reliability gate	$s < 0.1$, default $w = 2$
Decision period	Every 5 of 50 macro-steps

F_{eval} is independently fine-tuned from the same fixed ImageNet initialization as F_{ctrl}, using new heads and training seed 31,415. It is trained on inherited training indices, selected on validation, and opens final-test labels only in the final phase.

Table 7. Separately fine-tuned F_{eval} performance on real final- test images. These labels are not used by the controller.

Attribute	Accuracy	AUROC
Smiling	.8991	.9626
Bangs	.9510	.9771
Black Hair	.8788	.9335
Blond Hair	.9317	.9684
Brown Hair	.7960	.8786

Main KID uses 1,000 generated images, the complete 63,715-image inherited real marginal, subset size 1,000, and 100 fixed subsets. The reference positive prevalence is 44.42% for Smiling, 15.87% for Bangs, and 11.13% for Blond Hair. Ablations use subset size 100 and 50 subsets; the two protocols are not pooled.

D. Metric Definitions

Let N paired samples be indexed by i, target sign be y_a, and final independent-evaluator margin be $z_{i,a}$. Target success is

$$\text{Success}_a = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[y_a z_{i,a} \geq 0]\,. \tag{21}$$

For reference method r and protected set P_a, paired logit drift is

$$\text{Drift}_{a,r} = \frac{1}{N|P_a|}\sum_{i=1}^{N}\sum_{j\in P_a} \left|z_{i,j}^{\text{method}} - z_{i,j}^{r}\right|. \tag{22}$$

Protected flip rate counts changes in margin sign. The common-reference diagnostic uses paired ordinary conditional $w = 1$ trajectories for r.

E. Reproducibility Ledger

Table 8. Frozen experiment ledger.

Item	Frozen value
Generator	18,535,619-parameter conditional U-Net; EMA checkpoint; no HW4 updates
Integrator	50 Heun macro-steps; selected scale held per predictor–corrector pair
Main seeds	410000–410999 for each target and method
Ablation seeds	First 256 main seeds for each target
Controller	Ten decisions per trajectory; gate for $s < 0.1$
Main KID	1,000 generated; 63,715 real; subset 1,000; 100 subsets
Reduced KID	256 generated; subset 100; 50 subsets
Tests	19 implementation and protocol tests passed

F. Failure Records

For Smiling, local feedback does not reverse failure. For Bangs, all late candidates miss the target threshold. For Blond Hair, stronger intervention improves the margin but does not reach success.

Table 9. Rule-selected failures with independent-evaluator signed margins.

Target / seed	Static	Full	Max w	Fallback	Drift
Smiling / 410577	-10.291	-10.139	3	.30	.184
Bangs / 410550	-6.201	-7.777	2	.80	1.197
Blond / 410327	-7.205	-5.250	4	.10	.126