

Carnegie Mellon University

Constraint-Aware Receding-Horizon CFG for Attribute-Conditioned Flow Matching

Mu Chen | CMU 10-799 Diffusion & Flow Matching | Controllability Track

One affine guidance axis + one analytic endpoint map turns fixed CFG into constrained feedback.

1-D
CFG control authority

x0
analytic endpoint forecast

+0
extra U-Net calls for selection

First principles: derive the controller

1 Field structure
All scalar CFG actions reuse $g_s = v_c - v_u$.

2 Endpoint geometry
A straight path maps each action to a clean-endpoint forecast.

3 Control principle
Satisfy target + protected-attribute proxies with the least intervention, then re-observe.

Fixed w is open-loop. CRH-CFG asks the finite constrained question again at each observed state.

CRH-CFG: field reuse -> endpoint reasoning -> constrained action

$$v_w = v_u + w(v_c - v_u) \quad \hat{x}_0(w) = x_s + (1 - s)v_w$$

1. Reuse field
Frozen conditional + unconditional pair

>

2. Forecast
Affine scale bank -> analytic endpoints

>

3. Score
Frozen F_ctrl: margin + protected drift

>

4. Constrain
Feasible set + minimum intervention

>

5. Execute
Gate / fallback -> held Heun pair -> replan

$$A_s = \{w: \text{target margin} \geq m(s), \text{protected drift} \leq \epsilon(s)\};$$
 choose the feasible w nearest 1 and w_{prev} . F_{eval} stays outside the loop and measures only final outputs.

Execution: early gate | decide every 5/50 steps | hold w* across each Heun pair | matched U-Net work

Evidence: adherence at matched generator work

Attribute	Static w=2	Full CRH-CFG
Smiling	98.5	98.7
Bangs	98.0	98.4
Blond Hair	98.7	99.3

Observed successes (out of 1,000):
Smiling 985 -> 987 | Bangs 980 -> 984 | Blond 987 -> 993

Claim -> evidence -> boundary
Mean success: .9840 -> .9880 (+0.4 pp). Cost: +1.80%, same U-Net work.
Common-w=1 preservation does not improve; KID_marg rises for all targets but cannot identify conditional quality. No per-seed transitions: descriptive only.

Evidence boundary: success, boundary, and failure cases

Columns: static w=2, interval CFG, target-only, full CRH-CFG. Rows are selected by frozen rules: maximum-margin success, nearest evaluator boundary, minimum-margin failure.

Reachability
One scalar direction cannot make every constraint feasible.

Proxy activity
The selected preservation budget never changes the finite-set argmin.

Measurement
Endpoint scores from F_ctrl do not guarantee final outcomes under F_eval.

Adaptation concentrates on difficult trajectories

At the last saved decision (s=0.9):

Smiling mean w 2.017 | fallback 5.7%
Bangs mean w 2.085 | fallback 17.6%
Blond mean w 2.186 | fallback 1.6%

Median and IQR remain w=2 at every saved decision. CRH-CFG mainly corrects a minority of hard samples.

What the ablations and failures teach us

Reliability gate is essential
Removing it drops reduced-protocol success from 98.83% to 91.54%.

Preservation budget is non-binding
Adding the preservation term changes no reported reduced-controller value.

Extra scale saturates
 w_{max} 2 -> 3 adds +0.26 pp; 3 -> 4 adds 0.00 pp.

Scalar authority is limited
Selected failures reach high fallback or w=4 without crossing the evaluator boundary.

Supported: low-overhead adaptive correction. Not supported: joint adherence-preservation-quality improvement.

Reproducibility & compute

3 targets x 1,000 paired seeds | frozen checkpoint/split | 19 tests
HW4: 3.824 measured L40S GPU-h | HW3 training: 2.688 estimated |
qualified total: 6.512 GPU-h

Selected references

Lipman et al., Flow Matching, 2023 | Ho & Salimans, CFG, 2022 | Guided Flows, 2023
Limited Guidance Interval / CADs, 2024 | Online feedback, 2025 | CFG++ / Rectified-CFG++, 2024-25

Resources & disclosure

Built independently from the course starter repository and cited papers. OpenAI Codex assisted implementation, analysis, writing, and poster production; all displayed numbers were checked against retained artifacts.