WHEN CONTROL KNOBS MISLEAD: DIAGNOSING AND CALIBRATING REFERENCE CONTROL UNDER INCREMENTAL COMMITMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Inference-time reference controls let users change generated images, videos, and motions, but selecting a desired intermediate effect can remain difficult. We study this problem in samplers with incremental commitment, where tokens or temporal blocks become fixed before generation ends. If target commitments cluster within part of sampling, most output change can occur within a short interval of reference-duration settings. We examine when calibration can improve effect selection without retraining and what prevents it from doing so. To separate exposure timing from contextual prediction, we analyze an independent-token model. Under fixed commitment times and a hard duration gate, its normalized expected response equals the fraction of target tokens committed before reference release. This result explains how timing can concentrate a response, but does not imply that intermediate effects are absent. We therefore test inverse-response calibration and derive a conditional bound that separates fitting error from task shift and seed variation. We also bound sensitivity to endpoint-distance perturbations: a small zero-to-maximum-control gain can make normalized request errors unstable. We compare five generators and evaluate frozen standard isotonic calibration across seven controller configurations on three masked models, using twelve held-out tasks and three seeds per model. Commitment-order and overwrite ablations distinguish timing from the influence of reference input on predictions. Confidence-duration hit rates increase from 16.0% to 40.1% for Meissonic and from 20.1% to 44.1% for Show-o. MoMask shows no mean error reduction, including on valid confidence-duration requests. Its partial latent-code control leaves decoded-motion dependencies unmatched; matching the controlled codes therefore supplies no general endpoint guarantee. Our results show that correcting parameter selection requires checking both transfer to new tasks and the stability of the effect scale. A concentrated response alone cannot determine whether calibration will succeed.

1 INTRODUCTION

Generative models produce images, videos, and human motions from text and other inputs. Users can edit these outputs through inference-time guidance, reference conditioning, and intervention schedules (Ho & Salimans, 2022; Hertz et al., 2023; Ye et al., 2023). The scalar knobs exposed by these methods specify intervention strength or duration, whereas users seek a particular amount of change in the output. A knob may change the result substantially and still make an intermediate effect difficult to select.

We study effect selection under *incremental commitment*: a sampler fixes part of its state while generating the rest. Meissonic and Show-o decode masked image tokens, MoMask generates base motion tokens, and Self-Forcing generates causal video blocks (Bai et al., 2025; Xie et al., 2025; Guo et al., 2024; Huang et al., 2025). In our controlled loops, committed values remain fixed. Order changes which target positions receive reference input before commitment, as well as the context available to later predictions; the same duration can consequently produce different effects.

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.

For several tested commitment orders, many settings produce little change in reference distance, while a short parameter interval produces most of the response. To determine whether such a control can be corrected, we need to distinguish intermediate effects that are absent from those that are reachable but difficult to select. We also need a reliable scale for measuring the requested change. Inverse calibration can adjust parameter selection, but a map fitted on development tasks may transfer poorly to new tasks and seeds.

We first isolate timing with independent tokens to determine what commitment alone can explain. We then test inverse calibration across seven controller configurations on held-out requests, using transfer and endpoint bounds to interpret the remaining errors. Calibration improves image requests, while MoMask's decoded-motion results lead us to examine which decoder inputs the controller actually matches. This comparison identifies a conditional risk of partial latent control; its prevalence across motion generators requires further experiments.

2 Related Work

Inference-time generative control. Diffusion and flow models support interventions along an iterative sampling trajectory (Ho et al., 2020; Lipman et al., 2023; Esser et al., 2024). Cross-attention editing, feature injection, image-prompt adapters, and shared-attention style transfer modify attention, intermediate features, or conditioning signals during generation (Hertz et al., 2023; Tumanyan et al., 2023; Ye et al., 2023; Hertz et al., 2024).

Calibrated concept sliders. Continuous sliders expose a scalar parameter for varying a generative effect. FreeSliders and Measured Sliders calibrate this parameter through different control mechanisms. FreeSliders (Ezra et al., 2025) builds concept directions from frozen-model predictions without additional learned weights. Its automatic saturation and traversal detection selects an operating range and remaps nonlinear traversal; evaluations cover range, preservation, and smoothness across images, video, and audio.

Measured Sliders (Chen et al., 2026) learns measurement-supervised LoRA branches. It uses running-maximum inversion to select an endpoint, which is then linearly scaled by the user coordinate. It evaluates effect, drift, ordering, and preservation, using separate content for control selection and testing. Both works assess the output response and preservation of other content, extending evaluation beyond monotonicity of the output response.

Sampling order. Halton scheduling (Besnier et al., 2025) relates parallel-sampling error to conditional dependence and uses dispersed token sets to improve masked-image generation. We compare these approaches with our control protocol in Table 1, listing their interfaces, learned adaptations, evaluation criteria, and application scope. Their tasks and metrics differ, so the table compares protocols rather than numerical performance.

Method	Interface	Learned weights	Evaluation	Scope
FreeSliders (2025)	Score contrast; nonlinear remapping	None	Range, preservation, smoothness	Diffusion: image, video, audio
Measured Sliders (2026)	LoRA coefficient; endpoint calibration	LoRA	Effect, drift, ordering, preservation	Image attributes
Halton scheduler (2025)	Token selection order	None	Generation quality, sampling cost	Masked images
Ours	Reference duration; response inversion	None	Request error, hits, failures, anchor validity	Masked images and motion

Table 1: Protocol comparison with related controllers. Learned weights are additional model adaptations; calibration still requires development generations. Evaluation lists the measured properties, not a matched numerical ranking.

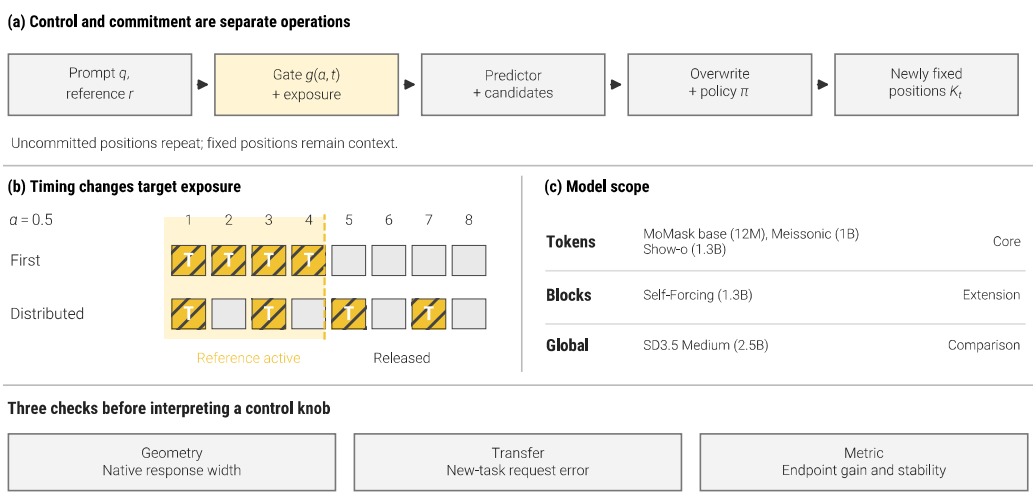


Figure 1: **Reference control under incremental commitment.** (a) Exposure, prediction, optional overwrite, and policy π determine new commitments. (b) Schematic: the same reference window covers different target commitments (hatched gold tokens). (c) Token/block commitment differs from global denoising. The bottom row separates geometry, transfer, and metric diagnostics.

3 Control Process and Evaluation

We describe an editing task by its source prompt q, reference r, seed s, and spatial or temporal target T. The field $M_i \in [0, 1]$ specifies control at position i. During generation, K_t contains the positions fixed after iteration t and U_t the remaining positions, with $K_t \subseteq K_{t+1}$. Once a value enters K_t, the control loop cannot revise it, but later predictions can still use it as context. Reference control must therefore act while the relevant positions remain uncommitted.

At progress p_t, duration control uses exposure probability $g_{it} = h(\alpha M_i, p_t)$, so increasing α keeps the reference active later into generation. Amplitude control uses $g_{it} = \alpha M_i h(1, p_t)$, varying exposure probability over a fixed window. Exposed uncommitted positions supply reference input to the predictor, which produces candidate values. Optional overwrite replaces exposed candidates by their reference values before policy π selects new commitments (Figure 1). The same duration can consequently cover different target commitments under different policies. Appendix A specifies the gate and random coupling. We denote the decoded output by $y_{\pi,\alpha}(q, s, r)$; α specifies the intervention applied during sampling rather than a percentage of visual style.

We test calibration in Meissonic (1B), Show-o (1.3B), and MoMask's base-token stage (12M) (Bai et al., 2025; Xie et al., 2025; Guo et al., 2024). We include Self-Forcing (1.3B) for causal video blocks (Huang et al., 2025) and SD3.5 Medium (2.5B) for global continuous generation (Stability AI, 2024). Global diffusion and flow update the full latent state without this hard token boundary (Ho et al., 2020; Lipman et al., 2023; Esser et al., 2024). Reported sizes exclude auxiliary encoders and codecs. Our commitment analysis concerns the studied loops, not every inference mode of each architecture.

We measure the output with mean absolute reference distance $L_T(y, r)$ inside T, in decoded pixels for images and joint coordinates for motion. To locate changes along each policy's parameter axis, we use its endpoint gain $G_\pi = L_T(y_{\pi,0}, r) - L_T(y_{\pi,1}, r)$:

$$z_\pi(\alpha) = \frac{L_T(y_{\pi,0}, r) - L_T(y_{\pi,\alpha}, r)}{G_\pi}, \qquad W_{10\text{–}90} = \alpha_{90} - \alpha_{10}. \tag{1}$$

A small $W_{10\text{–}90}$ places the middle 80% of the endpoint-normalized change within a short α interval; it does not count reachable output levels. We find crossings by linear interpolation on the sampled curve's cumulative maximum, without assuming the original response is monotone. Width is undefined when $G_\pi \leq 10^{-6}$.

Because this normalization gives each policy its own zero and one, it hides baseline differences that matter when a user requests the same effect from different controllers. We therefore introduce a common scale using confidence-duration endpoint distances L_0^c, L_1^c on the same task and seed:

$$e_\pi(\alpha; q, s) = \frac{L_0^c - L_T(y_{\pi,\alpha}, r)}{G_c}, \qquad G_c = L_0^c - L_1^c > 10^{-8}. \tag{2}$$

On this scale, $\beta = 0.5$ requests half the common reference-distance improvement. We leave e unclipped because another policy can start below zero or exceed one. A hit satisfies $|e - \beta| \le 0.1$, and bounded error is $\min(|e - \beta|, 1)$. Invalid anchors, missing outputs, and requests outside the frozen fit receive error 1 and hit 0. We also report raw target distances and change outside T to distinguish response concentration, request accuracy, and collateral effects.

4 From commitment to calibrated control

Under duration control, a token can receive reference input only if the reference remains active before its commitment. We isolate this timing dependence from learned context with independent target tokens and $M_i = 1$. Each unforced candidate matches its reference with probability $b < 1$, independently of overwrite. Token i commits at fixed progress τ_i, independent of α, candidate randomness, and overwrite randomness. It matches either because overwrite supplies the reference or because an unforced candidate already matches, giving

$$\mathbb{E}[\mathbb{I}\{y_i = r_i\}] = g(\alpha, \tau_i) + (1 - g(\alpha, \tau_i))b = b + (1 - b)g(\alpha, \tau_i). \tag{3}$$

Let $m(\alpha)$ be the expected reference-match fraction averaged over target tokens, with mismatch distance $1 - m(\alpha)$. A hard duration gate gives $g(\alpha, \tau_i) = \mathbb{I}\{\tau_i \le \alpha\}$; for $0 < \tau_i \le 1$, $m(0) = b$ and $m(1) = 1$. Averaging Equation 3 and normalizing this distance gives

$$z(\alpha) = \frac{m(\alpha) - b}{1 - b} = \frac{1}{|T|} \sum_{i \in T} \mathbb{I}\{\tau_i \le \alpha\} = F_T(\alpha). \tag{4}$$

Thus, even without dependence between tokens, the response counts target commitments made before reference release. When most occur within a short interval, the response also changes mainly there. A soft gate smooths each step, whereas full-window amplitude control gives $g = \alpha$ and $z(\alpha) = \alpha$ regardless of the fixed order. Comparing duration and amplitude in an independent-token sampler tests whether timing alone can produce concentration (Appendix H).

In learned generators, reference input also changes predictions and the information carried by earlier commitments. Commitment times can depend on α, and decoding changes the measured output, so Equation 4 is not an identity for neural outputs. We retain reference input while disabling overwrite to test how much effect survives through prediction. In Appendix B, we use a Gaussian example to distinguish sensitivity to the current input from reference information stored in fixed context.

Since timing can compress a response without removing its intermediate values, concentration alone does not establish a loss of reachable effects. For a continuous strictly increasing response, an inverse selects the setting that produces a requested effect β, instead of assuming $\alpha = \beta$. This remaps existing outputs, so it cannot repair a constant response or an effect outside the actual range, and a discontinuous response may still skip levels. Whether such an inverse helps in practice also depends on how well a response estimated on development tasks describes a new task.

We quantify this transfer error using the development response f_{dev}, its monotone fit $\hat{f}$, and a new task's seed mean f_q on a valid finite common scale. The frozen inverse chooses α_β with residual $r_\beta = |\hat{f}(\alpha_\beta) - \beta|$. Adding and subtracting these responses separates error due to inversion, fitting, task shift, and seed variation. If fit discrepancy is bounded by ϵ_{fit} and task discrepancy by δ_q, the triangle inequality gives

$$|e(\alpha_\beta; q, s) - \beta| \le r_\beta + \epsilon_{\text{fit}} + \delta_q + |e(\alpha_\beta; q, s) - f_q(\alpha_\beta)|. \tag{5}$$

Reducing the first two terms through calibration leaves task shift and seed variation uncontrolled. A good development fit therefore cannot establish correction on unseen tasks, which requires freezing the map before generating new-task outputs.

The request error also depends on the reliability of its effect scale. Nearly equal endpoint distances in Equation 2 give a small G_c, so slight distance changes can produce large changes in e. Writing the numerator as $A = eG_c$, perturbing each of the three distances by at most ε changes both A and G_c by at most 2ε. With changes δ_A, δ_G, the identity $\widetilde{e} - e = (\delta_A - e\delta_G)/(G_c + \delta_G)$ gives, for $G_c > 2\varepsilon$,

$$|\widetilde{e} - e| \leq \frac{2\varepsilon(1 + |e|)}{G_c - 2\varepsilon}. \tag{6}$$

Even a positive gain may therefore define an unstable scale, while a zero or negative gain does not define the intended increasing scale at all. Parameter inversion does not repair this denominator, so we evaluate endpoint gains alongside request errors and retain invalid requests. Our observed gains do not certify a perturbation margin; we use them to distinguish scale failures from timing and transfer errors.

5 Testing Calibrated Control

We test the timing prediction in Equation 4 by comparing confidence, random, target-first, and target-last at a fixed reveal budget. BALANCED distributes target commitments according to $\lfloor |T| C_t / N \rfloor$, where C_t is cumulative reveal count. Within these quotas, CONFLICT subtracts 0.5 from confidence when an exposed reference disagrees with the sampled candidate; this coefficient was fixed before testing. Confidence with amplitude control tests the alternative parameterization. All seven configurations retain 64 masked control iterations for Meissonic and 18 for Show-o and MoMask, excluding residual generation, decoding, references, and development calibration. Matching masked iterations does not match total calls or wall time.

Because parameter selection may correct a concentrated response, we fit one response map per configuration using the same development protocol. We evaluate 33 equally spaced α values on four development tasks at seed 0, normalize by Equation 2, average equally across valid tasks, and fit standard isotonic regression (de Leeuw et al., 2009). Its piecewise-linear inverse selects $\alpha_\pi(\beta) = \hat{f}_\pi^{-1}(\beta)$, using plateau midpoints for exact ties and abstaining outside the fit range.

After fitting, we freeze each model–configuration map and generate new tasks and seeds to evaluate the transfer error in Equation 5. Raw control uses $\alpha = \beta$, with generator weights fixed throughout. Test anchors establish evaluation units and never choose α or adapt the map.

We diagnose residual errors by comparing raw/mapped accuracy, fitted ranges, and held-out $\alpha = 0$ baselines for policy-dependent shifts. We check endpoint gains separately because unstable normalization can resemble transfer failure. These diagnoses use recorded outputs and do not change the selected controllers.

6 Experiments and Diagnosis

We first change commitment order to measure its effect on the response, then test whether calibration improves requested effects on unseen tasks. We specify a central image box or a temporal motion interval so that response width and request accuracy concern the same target. Image prompts change object color; motion prompts change action extent or pose. Meissonic and Show-o share prompts but generate separate references. We compare five generators and test frozen calibration on Meissonic, Show-o, and MoMask with four development and twelve disjoint test tasks per model, seeds 3, 4, and 5, and requests $\beta = 0.1, \ldots, 0.9$.

We retain all 13,608 test requests, including 1,000 invalid-anchor slots and 108 abstentions, because dropping them would reward controllers with restricted operating ranges. Nominal 95% intervals use 50,000 task-bootstrap resamples after within-task aggregation, without multiple-comparison adjustment. Appendix D records development, ablation, and endpoint jobs separately.

6.1 Does commitment timing explain response concentration?

Equation 4 predicts that moving target commitments changes where duration control takes effect. We therefore compare target-first and target-last while retaining the duration gate and reveal budget. Measuring endpoint gain alongside width lets us distinguish a concentrated response from reduced

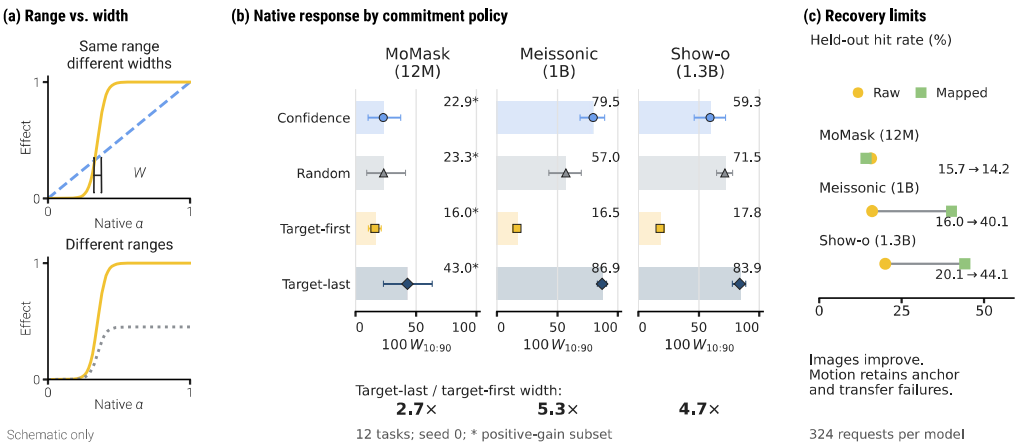


Figure 2: **Geometry, evidence, and correction.** (a) Schematic: range and native width describe different properties. (b) Measured $100W_{10–90}$, four orders, 65-point sweeps, twelve tasks, seed 0; whiskers are nominal 95% task bootstrap. Asterisks mark 10–11 valid MoMask widths. (c) Confidence-duration hits over all 324 held-out requests; gold/green denote raw/mapped control. Images improve, motion does not. Model order follows raw hit rate for display, not capability ranking.

total reference-distance improvement. Target-first and target-last yield widths of 0.165 and 0.869 for Meissonic, and 0.178 and 0.839 for Show-o (Figure 2). Their paired differences, 0.704 [0.678, 0.727] and 0.660 [0.601, 0.706], consistently place the target-first change within a shorter parameter interval. Endpoint gains differ by 0.032 [0.001, 0.072] and 0.002 [−0.025, 0.027]. In Show-o, the gain interval includes zero while the width interval does not; in Meissonic, both change. We therefore identify concentration as a distinct consequence of order, without treating it as evidence that the attainable range has disappeared.

We inspect commitments to check which tokens the same duration covers. At $\alpha = 0.5$, target-first copies all target tokens and target-last copies none in the development runs of all three masked models. We repeat the comparison with independent tokens to remove contextual prediction: first/last duration curves remain narrow, whereas amplitude width is 0.8 for every order (Figure 7, Appendix E). Concentration persists without context coupling, and changing the gate removes its order dependence. Target-last is nevertheless wide in learned image models, so timing alone does not explain their complete response.

We retain reference exposure and disable overwrite to test whether prediction accounts for effects beyond copying. Meissonic L1 falls from 0.307 without control to 0.055 with exposure alone, compared with 0.019 with overwrite; Show-o gives 0.292, 0.103, and 0.055. Because substantial distance reduction survives removal of copying, counting copied tokens cannot fully predict decoded effects. This ablation includes same-position reconstruction and cross-position context; it shows that prediction contributes but does not separate the two components.

6.2 Can a frozen inverse improve requested effects?

Since concentration need not remove intermediate effects, we ask whether selecting different native parameters makes those effects easier to request. We freeze the development inverse and generate held-out outputs, testing transfer rather than merely inverting the curve used to fit it. Meissonic hits rise from 52 to 130 of 324 requests and Show-o hits from 65 to 143, while bounded errors fall from 0.303 to 0.160 and 0.271 to 0.137 (Table 2). Calibration increases the number of requests within tolerance and reduces mean bounded error, supporting improvement beyond the binary hit-rate summary. The paired hit-rate intervals exclude zero for both image models. Because we change only the selected parameter, these gains show that part of the native error is correctable without changing the sampler or weights. More than half the requests still miss tolerance, so the frozen global inverse remains an incomplete correction on held-out tasks and seeds (Appendix E).

Model	Bounded error $\times 100$ ↓			Hits / 324 (%) ↑		
	RAW	MAPPED	Δ [95% CI]	RAW	MAPPED	Δ [95% CI], pp
MoMask (12M)	51.9	52.0	+0.02 [-4.6, 5.2]	51 (15.7)	46 (14.2)	-1.5 [-9.9, 6.2]
Meissonic (1B)	30.3	**16.0**	-14.3 [-18.0, -9.9]	52 (16.0)	**130 (40.1)**	+24.1 [13.3, 35.8]
Show-o (1.3B)	27.1	**13.7**	-13.4 [-18.7, -6.3]	65 (20.1)	**143 (44.1)**	+24.1 [4.9, 39.5]

Table 2: **Held-out confidence-duration calibration.** All 324 requests remain (12 tasks, three seeds, nine effects); invalid slots receive error 1, hit 0. Error is scaled by 100; hit tolerance is 0.1. Numbers are absolute; green intensity denotes improvement from raw; other cells are white. Δ gives paired mapped-minus-raw means and nominal 95% task-bootstrap intervals. Bold marks improvement, not significance.

	RAW ($E \times 100$, Hit%)						MAPPED ($E \times 100$, Hit%)									**Overall**		
	MoMask (12M)		Meissonic (1B)		Show-o (1.3B)		MoMask (12M)			Meissonic (1B)			Show-o (1.3B)			MoMask (12M)	Meissonic (1B)	Show-o (1.3B)
Policy	E ↓	Hit ↑	E ↓	Hit ↑	E ↓	Hit ↑	Fit	E ↓	Hit ↑	Fit	E ↓	Hit ↑	Fit	E ↓	Hit ↑	Hit: mapped/raw $\times 100$		
Confidence	51.9	15.7	30.3	16.0	27.1	20.1	9/9	52.0	14.2	9/9	16.0	40.1	9/9	13.7	44.1	90.2	250.0	220.0
Random	57.7	11.7	36.8	9.6	20.8	25.6	8/9†	65.8	7.7	9/9	15.4	43.2	9/9	15.5	41.7	65.8	451.6	162.7
Target-first	58.6	9.0	45.9	6.2	39.9	9.0	9/9	59.2	9.9	9/9	13.9	**44.8**	9/9	14.3	45.7	110.3	725.0	510.3
Target-last	60.1	12.0	15.2	43.2	17.8	39.5	9/9	61.2	10.8	9/9	15.8	40.1	9/9	15.7	42.9	89.7	92.9	108.6
Balanced	52.9	15.4	32.1	13.3	23.6	19.4	9/9	57.0	11.1	8/9†	24.1	37.0	9/9	14.2	44.8	72.0	279.1	230.2
Conflict-aware	53.0	15.4	32.6	13.0	23.5	21.0	9/9	56.9	11.1	8/9†	25.4	37.0	9/9	12.6	**50.9**	72.0	285.7	242.6
Amplitude	50.3	18.5	38.2	9.9	19.0	28.4	9/9	50.6	**17.6**	9/9	17.4	37.7	9/9	12.5	45.1	95.0	381.2	158.7

Table 3: **Seven controller configurations on a common scale.** Fit counts requests inside the development range; daggers mark exclusions, not physical limits. Error is scaled by 100; hits retain all 324 slots. Green intensity denotes mapped improvement from the same configuration's raw mean; other cells are white. Overall is mapped/raw hit rate $\times 100$ (100 means unchanged). Bold marks the highest mapped hit rate, not significance. All except Amplitude use duration control.

6.3 WHICH POLICY DIFFERENCES SURVIVE CALIBRATION?

The improvement for confidence motivates comparing all policies after calibration: does their raw ranking persist when we fit a parameter map for each? Mean error improves for six of seven Meissonic configurations and all seven Show-o configurations, and the largest-to-smallest error spread contracts from 0.307 to 0.116 and from 0.221 to 0.032. Meissonic target-first has higher error than confidence before calibration (0.459 versus 0.303) but lower error afterward (0.139 versus 0.160), despite its narrow native response. Raw rankings therefore mix policy behavior with parameter selection; these descriptive comparisons do not partition their causal contributions.

To explain remaining errors, we compare each fitted range with held-out baselines (Figure 3). Meissonic BALANCED and CONFLICT begin at $e(0) = 0.1569$ in development and reject $\beta = 0.1$, but their test mean baseline is -0.0221. A development exclusion therefore cannot establish that the effect is physically unreachable on new tasks. Their mapped hit rates are 37.0%, below confidence's 40.1%; distributing commitments does not ensure a transferable map. Show-o CONFLICT reaches 50.9% versus 44.1%, but its paired interval includes zero. We consequently evaluate policies through held-out accuracy and baseline transfer, rather than infer a generally superior schedule from width or the highest observed mean.

6.4 WHY DOES MOMASK REQUIRE A DIFFERENT DIAGNOSIS?

MoMask does not reproduce the image gains, so we first ask whether invalid normalization alone explains its error. Eight of 36 test task/seed anchors have nonpositive gain, contributing $72/324 = 0.2222$ to each confidence condition's bounded error. Removing those penalties leaves 252 valid requests with slot-averaged error 0.3821 before and 0.3823 after calibration. The near-unchanged

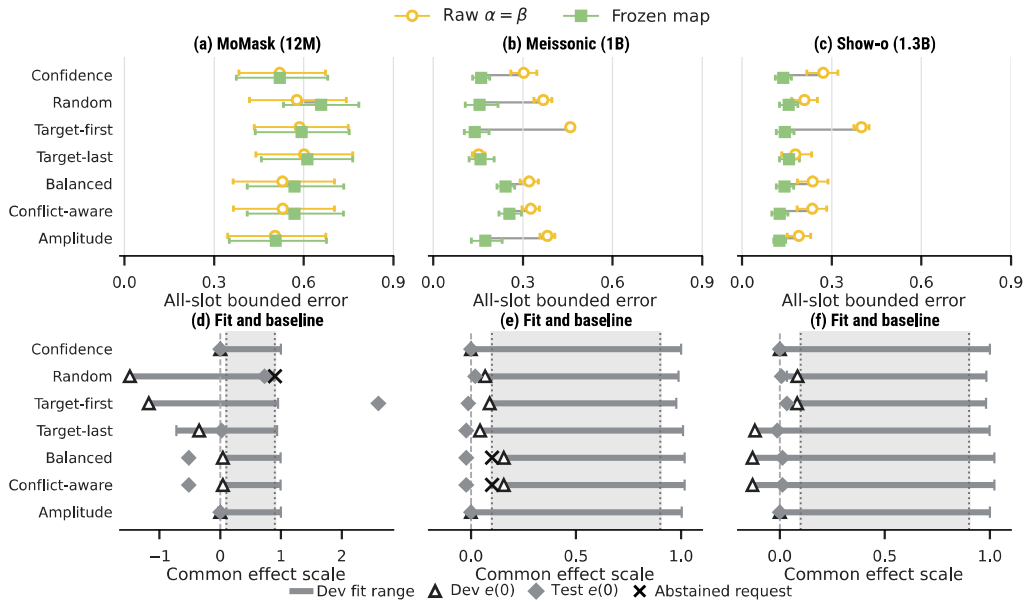


Figure 3: **Parameterization and policy transfer.** Top: raw gold circles and mapped green squares, bounded error with nominal 95% task-bootstrap intervals. Bottom: frozen fit ranges and development/test mean $e(0)$ on valid anchors, black/gray markers. Crosses denote excluded requests; shading covers $\beta = 0.1$–0.9. MoMask uses a wider bottom axis. Test endpoints never update the maps.

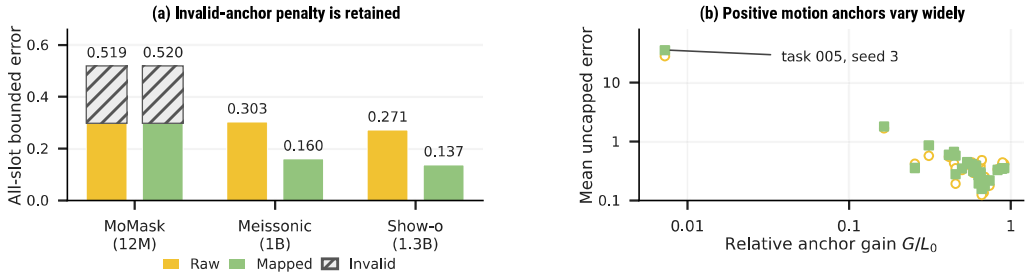


Figure 4: **Anchor validity and conditioning.** Left: gold raw/green mapped confidence error, with hatched invalid-anchor penalties. Right: all 28 valid MoMask anchors; each point averages nine uncapped errors and segments pair raw/mapped values. Dimensionless G_c/L_0^c measures relative gain. The weak anchor illustrates normalization sensitivity without isolating its architectural cause.

valid-request error rules out penalty accounting as the sole explanation; none of the seven configurations improves mean bounded error.

We then check whether positive gains provide meaningful units. The smallest MoMask gain is 0.00217, only 0.724% of its uncontrolled distance, versus minimum relative gains of 86.7% and 50.6% in the image models. At this weak motion anchor, a mapped confidence request has error 86.13 before capping. The identity $|e - \beta| = |L_T - [L_0^c - \beta G_c]|/G_c$ explains why a small decoded distance discrepancy can become a large normalized error. Positivity thus makes the ratio defined without making it stable, as Equation 6 predicts.

We trace MoMask's endpoint construction to identify dependencies that base-code control leaves unmatched. It pins base-layer codes, while residual levels still depend on the source prompt and a separate random stream. Temporal decoding mixes neighboring codes, and joint reconstruction integrates root rotation and translation. Matching the controlled codes therefore need not match the

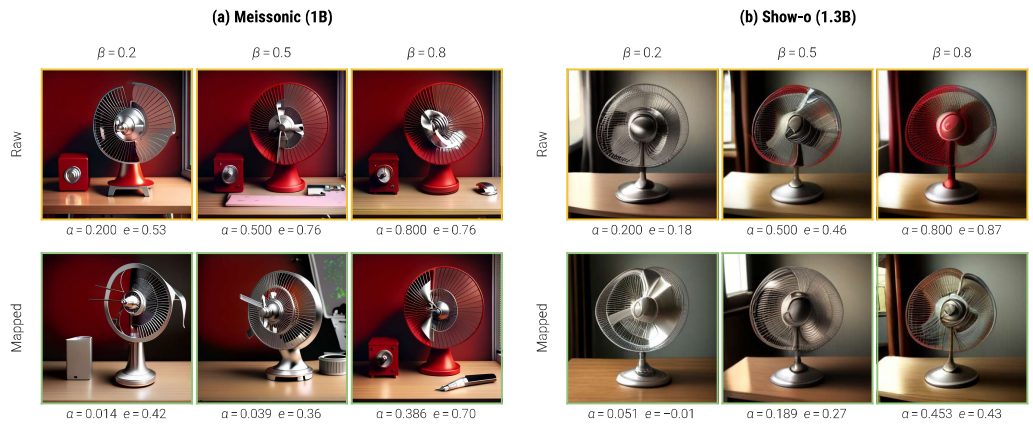


Figure 5: **Actual outputs at requested effects.** Fixed first image task, silver-to-red desk fan, seed 3, confidence-duration control. Gold raw/green mapped rows repeat $\beta = 0.2, 0.5, 0.8$ with α and measured e. Selection is fixed. Shape and background changes show why reference distance is not a pure color-strength measure.

target joints on which requests are evaluated. Other motion controllers face the same risk if they fix only part of a decoded trajectory's inputs while residual levels, neighboring latent codes, or root boundary states remain different from the reference. Appendix H shows why partial matching does not guarantee a reference endpoint. Our experiments establish the failure in MoMask; they neither test all motion generators nor isolate each decoder component's contribution. We conclude that commitment reordering or parameter inversion must be evaluated alongside the decoded endpoint and its unmatched dependencies.

6.5 Generated outputs and collateral effects

We inspect decoded outputs to check what improved reference-distance accuracy means for the edit. Figure 5 fixes the first image task and seed 3: shape and background changes accompany the color change, so calibration improves the specified distance response without isolating color strength. Across confidence requests, exterior L1 changes by $+0.0020$ for Meissonic and -0.0004 for Show-o; intervals span zero. We find no resolved mean collateral change under this measure, which does not establish quality equivalence. In the appendix, we report CLIP/LPIPS and motion diagnostics and show saved video and continuous-generation outputs. Our frozen calibration results apply to the three tested models.

7 Implications and limitations

Our experiments show that parameter remapping can improve requested image effects without changing the generator. This correction depends on intermediate effects remaining reachable and the fitted response transferring to new tasks on a stable scale. When the endpoints do not define reliable units, improving the inverse alone cannot resolve the normalized error; we must also revise the scale or endpoint construction.

In MoMask, matching partial latent codes need not match decoded motion when residual levels, temporal context, or root boundary states remain different from the reference (Appendix H). The same obstruction can arise in controllers that leave these dependencies unmatched, although we have not measured its frequency in other motion models. The proposed complete-input endpoint construction also remains to be evaluated for intermediate accuracy, cost, and collateral effects.

Our curated tasks, four development prompts, three test seeds, and reference-distance scale limit generalization. A successful inverse selects effects on that scale; it does not establish smooth semantic edits or eliminate the need to validate decoded endpoints.

AI USE STATEMENT

Generative AI assistance was used for implementing and debugging experimental and analysis code, literature discovery, and checking the novelty of the paper's research idea. The empirical results reported here derive from recorded model runs and deterministic analyses of their exported outputs; illustrative diagrams are labeled separately. AI assistance is not an independent source of experimental evidence. The authors are responsible for the final accuracy, attribution, and scientific claims of the submitted work.

ETHICS STATEMENT

We study inference-time control of existing image, video, and motion generators using diagnostic prompts we wrote and model-generated references, without human-subject evaluation. Our interventions inherit limitations and potential biases of the pretrained systems. Reference-distance improvement should not be interpreted as an assurance of semantic correctness, safety, or visual quality.

REPRODUCIBILITY STATEMENT

We provide the reference controller and commitment policies in Appendix A, the prospective tasks in Appendix G, and the splits, seeds, fit rules, failure handling, and numerical provenance in Appendix D. Appendix B gives the Gaussian argument and its assumptions. We generate the numerical figures and tables from saved evaluation records using the accompanying analysis scripts. The figure manifest links each fixed qualitative example to its source job identifier. Our evaluated controller requires no additional model training.

REFERENCES

Jinbin Bai, Tian Ye, Wei Chow, Enxin Song, Qing-Guo Chen, Xiangtai Li, Zhen Dong, Lei Zhu, and Shuicheng Yan. Meissonic: Revitalizing masked generative transformers for efficient high-resolution text-to-image synthesis. In *International Conference on Learning Representations*, 2025.

Victor Besnier, Mickael Chen, David Hurych, Eduardo Valle, and Matthieu Cord. Halton scheduler for masked generative image transformer. In *International Conference on Learning Representations*, 2025.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. MaskGIT: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Yijia Chen, Boyu Wei, and Xuanhua Yin. Measured sliders: Learning continuous controls from differentiable image measurements. *arXiv preprint arXiv:2609.05234*, 2026. URL https://arxiv.org/abs/2609.05234.

Jan de Leeuw, Kurt Hornik, and Patrick Mair. Isotone optimization in R: Pool-adjacent-violators algorithm (PAVA) and active set methods. *Journal of Statistical Software*, 32(5):1–24, 2009. doi: 10.18637/jss.v032.i05.

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 *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633, 2024.

Rotem Ezra, Hedi Zisling, Nimrod Berman, Ilan Naiman, Alexey Gorkor, Liran Nochumsohn, Eliya Nachmani, and Omri Azencot. FreeSliders: Training-free, modality-agnostic concept sliders for fine-grained diffusion control in images, audio, and video. *arXiv preprint arXiv:2511.00103*, 2025. URL https://arxiv.org/abs/2511.00103.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. MoMask: Generative masked modeling of 3D human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1900–1910, 2024.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations*, 2023.

Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. MotionGPT: Human motion as a foreign language. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.14795.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Stability AI. Stable diffusion 3.5 medium: Model card. Hugging Face model release, 2024. URL https://huggingface.co/stabilityai/stable-diffusion-3.5-medium. Accessed September 25, 2026.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, 2025.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. URL https://arxiv.org/abs/2308.06721.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2M-GPT: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2301.06052.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.