SPATIOTEMPORAL RESPONSE MODELING FOR DIFFUSION TRANSFORMER EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Controlling diffusion-transformer generation requires knowing where an internal edit propagates and how simultaneous edits interact. We introduce **ST-Lens**, a signed, coordinate-resolved response interface conditioned on residual block and denoising regime. It preserves spatial and video-time structure, revealing regional effects and cross-time interference. ST-Lens reduces response error by 37.15% against a frozen-path readout on 768 held-out fixed-dictionary interventions. We develop **response-coupled residual editing (RCRE)** to turn this information into coordinated temporal control: a full response Gram matrix jointly allocates a bounded residual write, and native-suffix feedback verifies its effect. Image control attains all 792 prescribed targets; Wan edits execute reversals, pauses and curved return trajectories. Response diagnostics motivate generation-policy candidates, whose final outcomes are measured with complete sampler rollouts. On ImageNet, a standard 50,000-generated-image evaluation per DiT arm against the 50,000 validation images gives 13.73% lower FID and 56.66% lower KID for scheduled guidance; a separate paired 1,000-image run gives a 1.312× sampling-and-decoding speedup. On MSR-VTT, advected sharing reduces mean FVD by 29.35% in a 1,024-pair cohort across three seeds and raises the reported three-seed means in all five VBench dimensions.

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.

1 INTRODUCTION

Consider editing a generated video so that a ball rolls back, pauses, and then resumes along a different path. A useful intervention must act at the right location and video time, yet its entry point is a hidden residual stream inside an iterative generator. A change at one token passes through later blocks, combines with other changes, and affects subsequent denoising updates. The practical question is therefore: *what output response will this internal write produce, where will it appear, and how will it interact with another write?*

Diffusion transformers provide a natural coordinate system for this question: spatial patches, video frames, network depth and denoising time (Peebles & Xie, 2023; Chen et al., 2024a; Team Wan et al., 2025). These axes play different roles. Spatial and video-time coordinates specify the desired change; depth and noise level determine where an intervention is applied. An observation that retains all four can connect a user's target with the network computation responsible for its downstream effect.

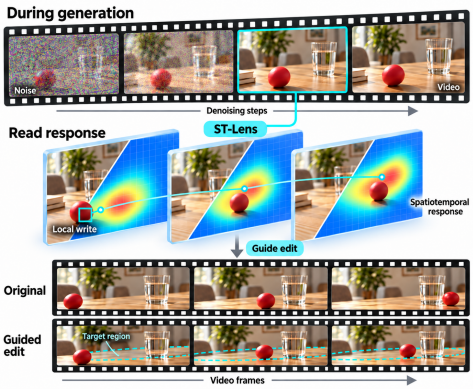

Figure 1: **Spatiotemporal insight guides editing.** Conceptual overview: ST-Lens observes the spatiotemporal response to a local probe; a downstream controller uses that observation to guide an edit.

Existing readouts illuminate intermediate content, useful feature directions and attention structure. Editing additionally requires the *signed effect of a specified change through the complete remaining network*. High directional agreement can conceal incorrect response magnitude, and pooling can erase equal-and-opposite regional effects. These are consequential information losses: a controller can identify a direction while still misjudging its strength or its non-target effects.

We introduce **ST-Lens** to make this response observable. ST-Lens estimates one signed response matrix on the editor's proposed writes. Conditional-mean calibration supplies reusable columns, complemented by online suffix probes at the current state. The output retains image coordinates or the full video grid. ST-Lens therefore exposes spatial propagation, cross-frame transport and signed reinforcement or cancellation. Its central contribution is this queryable spatiotemporal response interface, supported by an identifiability result and held-out intervention tests.

The three contributions form an observation-to-decision chain. ST-Lens estimates signed response columns; RCRE uses their Gram matrix to choose a deterministic edit; stochastic policies act through the covariance term of the same quadratic objective. Regime diagnostics motivate image guidance; ST-Lens's signed cross-time responses motivate video noise coupling. Our decision bound links response fidelity to local selection regret, and a same-state native audit tests local covariance ranking (Appendix AE). Complete sampler rollouts evaluate the resulting generation policies under the protocols in Tables 4 and 5. The displayed frames illustrate these evaluations. Figures 1 and 2 connect the response interface to its use in control.

Our contributions are:

1. **ST-Lens for observing diffusion-transformer computation.** We estimate signed residual-to-output transport across space, video time, block and noise regime. Identifiability and decision-error bounds connect the preserved information to local control and policy selection.
2. **RCRE: coordinated editing from signed responses.** RCRE optimizes the deterministic part of the response objective with a full Gram matrix, shared budget and native-suffix feedback. A scalar-gap bound characterizes possible local benefit from independent coefficients; image and video audits establish localization and command attainment in tested cases.
3. **Improved image and video generation.** ST-Lens's regime and cross-time response diagnostics motivate guidance allocation and temporal noise coupling. The common mean–covariance objective and matched local audit inform policy design; full rollouts test quality–cost and distribution–motion outcomes.

2 Related work

Reading internal representations. The logit lens directly applies the output unembedding, and the tuned lens learns affine translators (nostalgebraist, 2020; Belrose et al., 2023). Diffusion Lens decodes intermediate text-encoder representations; Diffusion Steering Lens uses steering and path patching (Toker et al., 2024; Takatsuki et al., 2025). J-Lens develops context-averaged Jacobian transport for vocabulary readout, with subsequent analysis of positional energy and affine approximation (Gurnee et al., 2026; Yan et al., 2026). ST-Lens applies mean-derivative estimation to signed native diffusion fields indexed by source/target coordinates and noise regime. The native-field, coordinate-resolved response query yields a regional Gram matrix for editing (Appendix AG). ConceptAttention and TIDE expose interpretable diffusion features (Helbling et al., 2025; Huang et al., 2026); ST-Lens measures how a specified perturbation acts after the frozen suffix.

Editing through internal structure. Pullback geometry and localized diffusion subspaces supply useful control directions (Park et al., 2023; Chen et al., 2024b); attention decomposition extracts directions from pretrained weights (Anand et al., 2026). Prompt-to-Prompt and feature injection edit generation through internal activations (Hertz et al., 2023; Tumanyan et al., 2023). DiTFlow reads cross-frame attention for motion transfer, HALO uses specialized motion and structural heads, and FlowDirector corrects editing flow (Pondaven et al., 2025; Jung et al., 2026; Li et al., 2025). RCRE instead fits the signed native-output responses of temporal residual proposals, using measured

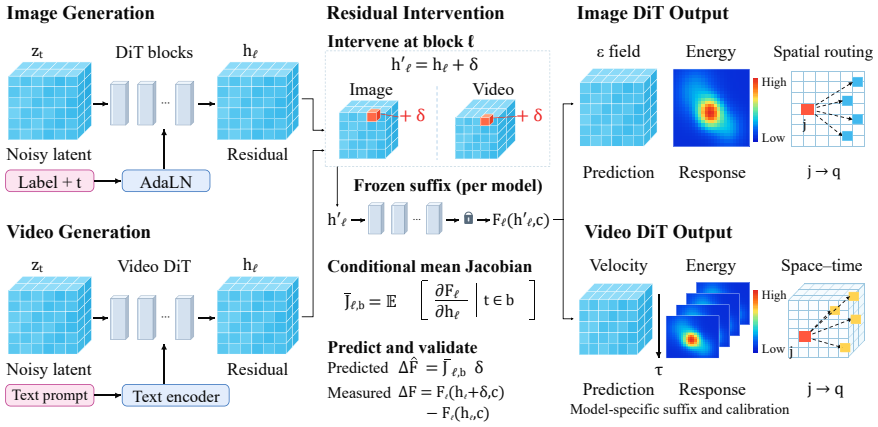

Figure 2: **From spatiotemporal insight to diffusion-transformer editing.** At block ℓ and denoising time t, a localized residual write propagates through the frozen suffix. ST-Lens estimates its coordinate-resolved response using calibrated columns and current-state column updates, then verifies the resulting write against the intervened-minus-unmodified output. The controller consumes this common response matrix (Section 4.1). Image epsilon and video velocity retain their native grids; video time τ indexes frames within a denoising state.

interference and a shared write budget. Appendix G specifies the mechanism-matched adaptations used to compare response observations under a common controller.

Generation quality and efficiency. Guidance, numerical solvers and activation reuse alter the cost and quality of diffusion sampling (Ho & Salimans, 2022; Lu et al., 2022; Ma et al., 2024). Warped noise also supports video motion control (Burgert et al., 2025). Our generation study turns regime and temporal diagnostics into guidance allocation and noise-coupling policies, evaluated with FID/KID, temporal metrics and separate VBench dimensions (Heusel et al., 2017; Huang et al., 2024). Complete rollouts measure the quality and motion outcomes of the resulting fixed policies.

3 ST-LENS: OBSERVING SPATIOTEMPORAL RESPONSE

ST-Lens estimates the signed, coordinate-resolved response to a declared set of residual writes in a frozen diffusion transformer. Calibration and online probes populate the same response matrix consumed by the editor. Spatial position and video time index its output; block and denoising regime index the operating state. Figure 2 shows this workflow.

1. Specify the residual intervention. At block ℓ and noise level t, let $h_{\ell,t} = \mathrm{vec}(H_{\ell,t})$ denote the residual state. Its N tokens retain a spatial grid for images or a spatiotemporal grid for videos, each with hidden width D. We write $h'_{\ell,t} = h_{\ell,t} + \delta$, where $\delta = aMd$: M selects source patches or voxels, d specifies a direction, and a controls amplitude. Location, direction, block and noise level jointly define the write.

2. Make the effect observable. Let $F_{\ell,t}(h, c)$ denote the remaining computation at fixed timestep and conditioning c. Suppressing t in the suffix notation, the measured change is $\Delta F = F_\ell(h + \delta, c) - F_\ell(h, c)$. The image-editing comparisons use the four-channel epsilon field; the Wan comparisons use the conditional velocity field. Each target keeps the native output coordinates, giving the controller an explicit region in which to concentrate the edit. Appendix A specifies output-head, conditioning and sampler conventions for each experiment.

3. Estimate one response matrix. For proposal dictionary $D = [d_1, \dots, d_K]$ and context x, the target response is $J_x D$. Conditional-mean calibration supplies $G = \bar{J}D$:

$$\bar{J}_{\ell,b} = \mathbb{E}_{(z_t,c)\sim\mathcal{D}|t\in b}[\partial F_\ell(h_\ell, c)/\partial h_\ell], \tag{1}$$

$$\widehat{A}_x = G + (L_x - G)M_x, \qquad \widehat{\Delta F}_x = \widehat{A}_x u, \quad \delta = Du. \tag{2}$$

Here M_x selects measured columns and $L_{x,k} = [F_x(h + sd_k) - F_x(h)]/s$ is their current-state secant response. Centered differences or JVPs provide derivative audits. $M_x = 0$ reuses calibration; $M_x = I$ refreshes every proposal, as in decoded RCRE. Intermediate masks describe partial refresh; the Wan backend also regularizes unmeasured columns (Appendix F). The estimator and controller share one response representation throughout.

4. Visualize and test the prediction. Partitioning the operator by source and target locations preserves the grid. For source directions d_i, target projectors P_q and local responses $v_i = J_F(h)d_i$, we inspect

$$\mathcal{E}_{\ell,b}(j \to q) = \|\bar{J}_{\ell,b}(q, j)\|_F^2, \tag{3}$$

$$T_{iq} = \|P_q v_i\|^2, \quad G_{ik,q} = \langle P_q v_i, P_q v_k \rangle,$$

$$\left\| \sum_i P_q v_i \right\|^2 = \sum_i T_{iq} + 2 \sum_{i<k} G_{ik,q}. \tag{4}$$

Routing energy summarizes source–target influence. Signed cross terms determine how writes combine; they enter both deterministic fitting and random-policy energy through the response Gram matrix.

We verify the assembled response against actual finite writes. For the cached component, prediction error separates into context variation, calibration error and finite-amplitude curvature:

$$\Delta F_x - \widehat{J}\delta = (J_x - \bar{J})\delta + (\bar{J} - \widehat{J})\delta + \mathcal{R}_x(\delta). \tag{5}$$

Calibration, refresh and verification. The conditional mean minimizes squared first-order prediction risk among context-independent estimates. Online refresh measures the same proposed actions at the current state; finite-suffix verification closes the loop before execution. Calibration can amortize observation cost where transfer is accurate; the Wan final-hidden audit instead requires current-state refresh. Appendix C derives the calibration risk; Appendix D bounds finite-write error.

What output coordinates identify. Let G be a dictionary-restricted response and P a pooling operator. A regional functional $q^\top G$ is determined by PG for every G exactly when $q \in \mathrm{range}(P^\top)$. Otherwise two operators can share the pooled readout while producing opposite signed regional contrasts. ST-Lens retains the coordinates needed to answer these regional queries (Appendix D).

4 Editing the generation process

4.1 Selective image and video-field control

The controller below is an evaluation backend shared across observation methods. An observer provides a response matrix or a set of directions to measure; the optimization, execution and feedback are held fixed. Let $D_s \in \mathbb{R}^{ND\times K}$ be an orthonormal source dictionary and $G = \widehat{J}D_s$ its calibrated response matrix. Coefficients u define the write $\delta = D_s u$. For a normalized target vector q supported in a declared region, target gain $g_\star > 0$ and source budget ρ, we solve

$$\min_u \ \|P_{\mathrm{out}} A u\|_2^2 \quad \text{subject to} \quad q^\top A u = g_\star, \quad \|u\|_2 \le \rho. \tag{6}$$

Here $A = \widehat{A}_x$ is the ST-Lens estimate in Eq. 2. The solver balances outside-region response and target gain within the source dictionary. When the requested linear gain exceeds the budget, it uses the bounded maximum-gain direction. The subsequent suffix evaluation determines actual target attainment, and measured correction adjusts the write to the current nonlinear response.

The shared image backend measures six calibrated directions. Its Wan counterpart starts with four, completes the unmeasured complement with a weaker calibrated estimate, and measures four additional directions when target error exceeds 10%. Every observation source receives the same completion, correction and failure policy. Appendix F gives these implementation details and distinguishes them from the response-coupled editing algorithm below.

4.2 Response-coupled residual editing

RCRE uses ST-Lens responses to coordinate temporal writes. Disjoint source regions can reinforce or cancel at shared output locations; fitting their strengths jointly targets the requested trajectory within one write budget.

At a chosen residual site, partition a proposal into video-time groups $D = [d_1, \dots, d_K]$. Measure each signed response $a_k = [F(h + \epsilon d_k) - F(h)]/\epsilon$ and form $A = [a_1, \dots, a_K]$. For a desired conditional-velocity change r, the local fitting objective is

$$\min_{0 \le b_k \le 2} \|Ab - r\|_2^2 + \lambda \|b\|_2^2 \quad \text{s.t.} \quad \mathrm{RMS}(Db) \le \rho\, \mathrm{RMS}(h). \tag{7}$$

We use $\rho = .025$, with the animal-budget extension in Appendix Y. For $C = A^\top A$, the off-diagonal entries C_{ij} capture reinforcement and cancellation between writes at different video times. A diagonal fit discards them; a scalar fit forces every group to use the same strength. RCRE retains the full signed response in a constrained fit, followed by native-suffix acceptance. The unchanged head, CFG and scheduler then generate the next latent state.

We use four temporal groups at block 28 (Appendix Q). A terminal-head inverse supplies the proposal and an explicit reference supplies r. Refreshing all four columns gives $M_x = I$ in Eq. 2. Appendix P specifies probe, budget and native-acceptance rules.

One objective, two decision types. For random coefficients b with mean μ and covariance Σ,

$$\mathbb{E}\|Ab - r\|^2 = \|A\mu - r\|^2 + \mathrm{tr}(C\Sigma). \tag{8}$$

RCRE chooses μ with $\Sigma = 0$; stochastic policies choose covariance. The off-diagonal terms of C distinguish equal-marginal temporal policies, providing ST-Lens's response-based reason to compare static and motion-aligned sharing. Appendix AE proves the identity, a decision-error bound and a scalar-gap criterion. The response must match the intervention site: initial-noise policies require full sampler transport. The same-state audit ranks eight residual covariance policies correctly at all six tested states (Appendix AG). Complete rollouts then select the reported video setting using the final generation metrics.

5 Experiments

Experiments test response fidelity, control, RCRE and generation. Tables retain recorded sample counts; frames illustrate outputs. Comparisons match checkpoint, conditioning, budget and target, with separate calibration/test contexts. The 48-condition Wan audit nests conditions within four independent videos (Appendix A).

5.1 ST-Lens reveals actionable spatiotemporal structure

We compare observers on eight fixed dictionary axes, three dictionaries and 32 held-out DiT contexts: 768 finite interventions at block 21, $t = 510$. Every readout predicts the same measured suffix response to the same write, independently of controller selection. This measures transfer to new contexts within the declared dictionary. The cached-column audit fixes $M_x = 0$; the editing experiments evaluate the assembled matrix with online refresh. Direct-head and frozen-path readouts use the same response target.

ST-Lens has the highest cosine and lowest relative error among these reusable readouts, reducing error by 37.15% over frozen-path (Table 1). The small cosine gap conceals a material difference in the observed field: directional agreement alone does not recover its magnitude. Coordinates add another distinction: spatial pooling loses almost the entire response despite using the same calibrated information before pooling.

Table 1: **ST-Lens observation fidelity.** All observers predict the same 768 finite interventions on 32 held-out DiT contexts.

Observer	Cosine ↑	Relative error ↓
Direct-head (Logit-Lens style)	0.99303	0.81288
Frozen-path (DSL inspired)	0.99575	0.09931
ST-Lens	**0.99807**	**0.06241**
Pooled ST-Lens (coordinate ablation)	0.01161	0.99992

Block 21, $t = 510$; three dictionaries and eight axes per context. ST-Lens reduces error by 37.15% relative to frozen-path; paired 95% CI: [34.61%, 39.88%].

Observing effects that cancel in an aggregate. In a separate paired-sign test, opposite regional effects retain only 4.5×10^{-5} of their energy in the pooled response. ST-Lens predicts their full signed fields with mean cosine 0.99805 and relative error 0.06271. This realizes the identifiability distinction in Section 3: the regional response remains observable when its pooled summary nearly vanishes. Protocols and all readouts are in Appendix D.

Conditioning preserves the operating regime. An independent 32-context factorial fixes the solver and calibration count. At block 21/$t = 510$, conditional response error is 0.0626, versus 0.0981 for frozen-path, 0.1388 for depth sharing and 0.4205 for time sharing. Full-grid regime comparisons appear in Appendices H–I. This experiment isolates the observation supplied to an unchanged downstream operation.

Figure 3 locates this dependence across three blocks and five noise levels in two backbones. At block 7/$t = 100$, cosine is .501 for DiT and .460 for PixArt; at block 21/$t = 500$, it reaches .998 and .985. This map identifies regimes for reusable observation and those requiring current-state refinement.

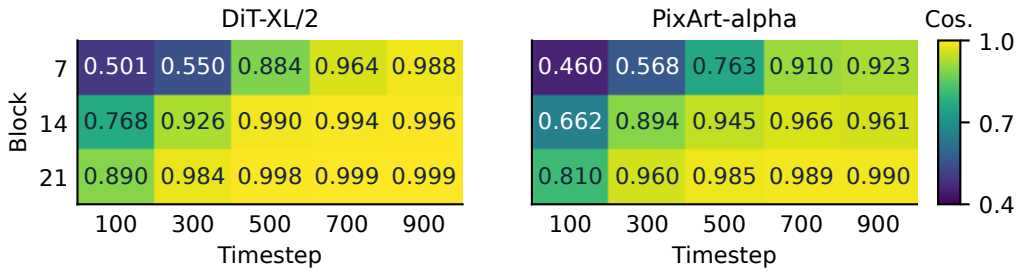

Figure 3: **ST-Lens fidelity depends on depth and denoising time.** Held-out intervention cosine under separate calibration for each cell. DiT uses 128 calibration and 128 test ImageNet contexts; PixArt uses 16+16 fixed-caption synthetic latents. Larger t means noisier states. Appendix M gives the protocol and mode-recovery audit.

5.2 Selective Image and Video-Field Control

Selective image response. The same rank-six controller compares ST-Lens with spectral, weight-space and random observation sources. ST-Lens attains all 792 targets and reduces mean outside-region response in all 36 matched comparisons, by 7.37–95.43%; 35 have positive adjusted bootstrap lower bounds. Table 2 gives the backbone-level results; Figure 5 resolves all 36 comparisons. On 32 new PixArt trajectories with disjoint prompts, frozen calibration improves all 18 matched means with adjusted support and matches Full8's 99.65% attainment with 9.021 versus 11.021 calls.

Video-field control. On 48 fresh Wan conditions, adaptive ST-Lens gives 3.363% target error against 3.600% for adaptive spectral and 3.686% for adaptive QKV observations; all three attain 41/48 targets. Its realized cost is 8.750 calls, compared with 11.875 for Full8 at 2.679% error.

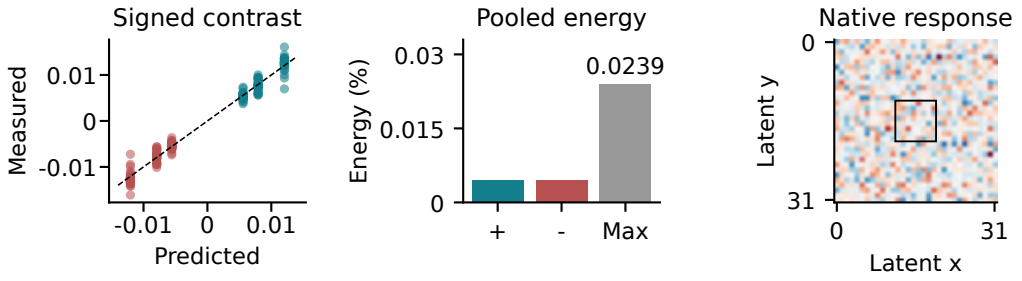

Figure 4: **Regional effects survive an almost vanishing pooled readout.** Opposite writes produce opposite regional contrasts, with pooled/full energy approximately 4.5×10^{-5}. Right: native channel 0, blue–red limits ± 0.02565. Protocol and per-write scores: Appendix D.

Table 2: **Image editing: target attainment, selectivity and cost.** Rank-six methods share the source budget, target and feedback controller.

	DiT-XL/2 (576 controls)			PixArt-α (216 controls)		
Method	Success ↑	Leakage ↓	Calls ↓	Success ↑	Leakage ↓	Calls ↓
Pullback	98.96	1.327	9.080	99.54	1.353	9.037
QKV	98.26	1.489	9.135	96.30	1.864	9.222
Random	99.65	1.496	9.080	**100.00**	1.266	9.009
ST-Lens	**100.00**	**1.018**	**9.007**	**100.00**	**1.021**	**9.000**
Full8 reference	100.00	1.000	11.007	100.00	1.000	11.000

Success: %, within 10% target tolerance. Leakage: mean outside L2 / Full8 mean, retaining failed targets. Calls include probes and verification. Bold: best at matched rank; Full8 uses eight directions.

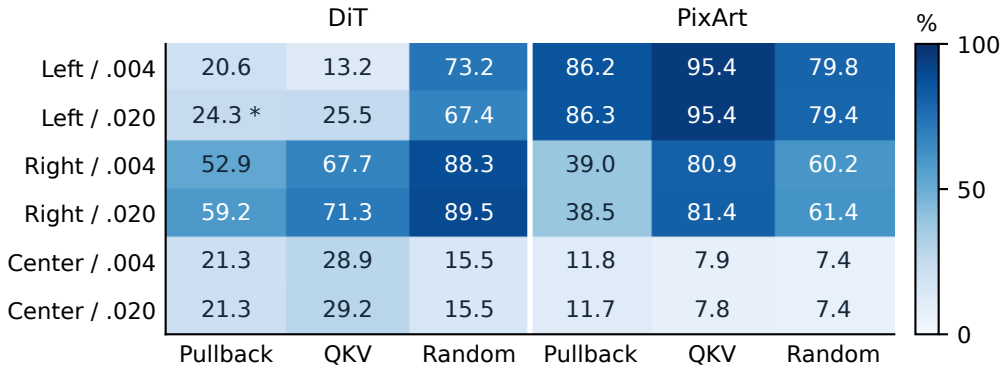

Figure 5: Full matched-rank image comparisons. Cells show outside-response reduction (%) against each baseline; * marks the adjusted interval including zero.

The complete eleven-method comparison and task breakdown are in Appendix N. A separate 240-condition final-hidden audit gives cached-column error .9408 and refreshed-column error .0724. The video editor therefore refreshes its response columns at the current state.

5.3 RCRE: COUPLING MECHANISM AND TEMPORAL COMMANDS

The controlled comparison shares ST-Lens response columns, proposals and native-suffix acceptance. Scalar fitting retains the cross terms but ties $b = c\mathbf{1}$; diagonal fitting drops cross terms; Joint RCRE retains them and frees the temporal coefficients. These test two different choices: response

coupling and allocation freedom. Joint RCRE lowers mean native target error by 5.46% versus diagonal fitting at the same 132 suffix calls and 51 full forwards. Scalar already uses the measured coupling, explaining its strong accuracy on these commands. Appendix Figure 12 shows native Wan trajectories under a common 2.5% residual RMS cap.

Two further prompt conditions request right-to-left and receding motion, while common input references specify three unequal reversals or a loop, pause and slower retrace. Table 3 reports complete-command attainment using maximum position, direction, travel and event-time errors. Joint RCRE completes both commands with exact event timing and a two-case mean maximum-position error of 1.273 px, versus 1.283 px for Scalar. This 0.010 px difference is below the prespecified 0.75 px decoded-improvement margin; Scalar also completes both commands. These examples demonstrate temporal command execution with comparable decoded accuracy for Joint and Scalar. The diagonal comparison isolates response cross terms in native field fitting; decoded registration is similar. Protected response trades target accuracy for outside-region control, exceeding the 4 px threshold in the receding-prompt case. Vehicle reversal and animal hold–traverse–retrace extend these commands to natural scenes (Appendix Y); Joint achieves 0.32 px and 0.00 px registration RMSE, respectively. Appendix U adds temporal-minimax and transition-response fits.

Table 3: **Video editing accuracy under explicit spatiotemporal commands.** Editors share prompt, seed, reference and fitting budget; no-write uses the same base sampler.

	Command error ↓					
Editor	Position (px)	Direction (°)	Travel (%)	Event (frames)	Tracked frames	Pass
A: repeated reversals; right-to-left prompt						
Vanilla (no write)	147.058	178.85	93.92	–	49/49	F
Scalar fit	1.136	**1.43**	**0.47**	0	49/49	P
Joint RCRE	**1.104**	**1.43**	**0.47**	0	49/49	P
Protected response	2.275	1.53	1.07	0	49/49	P
B: loop, hold and retrace; receding prompt						
Vanilla (no write)	–	–	–	–	48/49	U
Scalar fit	**1.430**	**1.73**	**0.32**	0	49/49	P
Joint RCRE	1.443	**1.73**	0.33	0	49/49	P
Protected response	4.048	3.09	0.91	0	49/49	F

Position/direction: maximum error; travel: maximum phase-relative error. Pass requires position ≤ 4 px, direction $\leq 10^\circ$, travel $\leq 10\%$, event error ≤ 1 frame and valid tracking throughout. Command-specific clauses: Appendix S. P/F/U: pass/fail/unverifiable. Bold: best among shown methods per task; all six methods: Table 20.

5.4 Image and video generation performance

Image quality and sampling cost. CFG writes $(w_t - 1)(\epsilon_c - \epsilon_\varnothing)$ to the native field. Its measured effect rises at low noise, motivating the manually set 1/4/6 schedule with thresholds 500/300. Table 4 shows a 13.73% FID reduction, 56.66% KID reduction and 1.312× speedup. The schedule uses 50 conditional and 25 unconditional branches versus 50+50. Appendix AC gives the paired gallery and diagnostics.

Table 4: **ImageNet generation performance relative to vanilla DiT.** FID/KID use 50,000 generated images per arm against ImageNet validation; DDIM-50, DiT-XL/2-256.

Policy	FID ↓	KID ↓ ($\times 10^3$)	Time ↓ (s)	Speedup ↑ ($\times$)
Vanilla (CFG 4)	59.9978	12.6638	1241.66	1.000
Scheduled CFG 1/4/6	**51.7615**	**5.4881**	**946.50**	**1.312**
Reduction vs. vanilla	13.73%	56.66%	23.77%	–

FID/KID: 50,000 generated samples per arm, 50,000 ImageNet validation references; KID is shown $\times 10^3$. Time: a separate 1,000-image paired run on RTX 4070, including sampling and VAE decoding.

Video distribution, consistency and retained motion. ST-Lens's signed cross-frame responses motivate equal-marginal static and motion-advected noise candidates. Complete-rollout comparisons select their coefficients, including .10. Vanilla Wan is the uncoupled reference. MSR-VTT's captioned real clips support paired FVD, and VBench custom-input scores the generated videos (Xu et al., 2016; Huang et al., 2024). Across three random seeds, advected sharing (.10) reduces mean FVD by 29.35% in the 1,024-pair cohort and retains 90.17% of vanilla flow magnitude on the reported three-seed summary. Static sharing (.05) reduces mean FVD by 6.31% in a disjoint 4,096-pair cohort. Three-seed VBench means favor advected sharing on all five dimensions (Table 5). The dynamic-degree change is small (.3750 to .3769), while the other four dimensions rise by .0074–.0198. FVD, flow magnitude and VBench dynamic degree summarize different properties: lower distribution distance coexists here with 9.83% less mean flow and a nearly unchanged dynamic score. Appendix A discusses their evaluation units and seed-level variation.

Table 5: **MSR-VTT video generation relative to vanilla Wan.** Scores are summarized over three seeds. FVD cohorts are paired and disjoint; VBench evaluates the 5,000-video MSR-VTT setting.

Policy	FVD ↓	Reduction (%) ↑	Retained flow (%)
Development cohort: 1,024 paired captions/seeds			
Vanilla Wan	4551.09	0.00	**100.00**
Static sharing (.10)	**3154.99**	**30.68**	43.06
Advected sharing (.10)	3215.19	29.35	90.17
Independent cohort: 4,096 paired captions/seeds			
Vanilla Wan	2389.21	0.00	**100.00**
Static sharing (.01)	2454.12	−2.72	–
Static sharing (.05)	**2238.36**	**6.31**	81.55

VBench dimensions: 5,000-video MSR-VTT setting

Policy	Subject ↑	Backgr. ↑	Dynamic ↑	Aesth. ↑	Imaging ↑
Vanilla Wan	.9638	.9390	.3750	.5717	.7466
Advected (.10)	**.9752**	**.9588**	**.3769**	**.5791**	**.7567**

Wan2.1-T2V-1.3B; 17 frames, 832×480, 30 UniPC steps. FVD uses Kinetics-400 I3D 400-D output; reductions compare three-seed means within each cohort. Retained flow compares motion magnitude with paired vanilla; VBench uses the 0–1 scale. Bold: best per column; –: unavailable.

6 Conclusion

ST-Lens supplies coordinate-resolved signed responses for local editing and motivates temporal covariance candidates. RCRE executes the tested commands; Joint and response-based Scalar fits have comparable decoded accuracy on these examples. Wan editing benefits from current-state response refresh. Complete rollouts show quality and motion trade-offs for the selected generation policies under the reported protocols.

REFERENCES

Tharun Anand, Mohammad Hassan Vali, Arno Solin, Green Rosh, and B H Pawan Prasad. Self-attention decomposition for training free diffusion editing. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026.

Atilim Gunes Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research*, 18(153):1–43, 2018.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting Latent Predictions from Transformers with the Tuned Lens. *arXiv preprint arXiv:2303.08112v1*, 2023.

Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, Michael Ryoo, Paul Debevec, and Ning Yu. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024a.

Siyi Chen, Huijie Zhang, Minzhe Guo, Yifu Lu, Peng Wang, and Qing Qu. Exploring low-dimensional subspaces in diffusion models for controllable image editing. In *Advances in Neural Information Processing Systems*, 2024b.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. *Transformer Circuits Thread*, 2026.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53 (2):217–288, 2011.

Alec Helbling, Tuna Han Salih Meral, Benjamin Hoover, Pinar Yanardag, and Duen Horng Chau. ConceptAttention: Diffusion transformers learn highly interpretable features. In *International Conference on Machine Learning (ICML)*, volume 267, pp. 22946–22963, 2025.

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.

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)*, 2017.

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

Victor Shea-Jay Huang, Le Zhuo, Yi Xin, Zhaokai Wang, Fu-Yun Wang, Yuchi Wang, Renrui Zhang, Peng Gao, and Hongsheng Li. TIDE: Temporal-Aware Sparse Autoencoders for Interpretable Diffusion Transformers in Image Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(1):435–443, 2026. doi: 10.1609/aaai.v40i1.37006.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Sunyoung Jung, Jiwoo Park, Yoonseok Choi, Kyobin Choo, Ming-Hsuan Yang, and Seong Jae Hwang. Controlling motion transfer in diffusion transformers via attention heads. *arXiv preprint arXiv:2607.11081*, 2026.

Guangzhao Li, Yanming Yang, Chenxi Song, and Chi Zhang. FlowDirector: Training-free flow steering for precise text-to-video editing. *arXiv preprint arXiv:2506.05046*, 2025.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems*, volume 35, pp. 5775–5787, 2022. doi: 10.52202/068431-0418.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. DeepCache: Accelerating Diffusion Models for Free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022. doi: 10.52202/068431-1262.

nostalgebraist. Interpreting GPT: the logit lens. LessWrong, 2020.

Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of Riemannian geometry. In *Advances in Neural Information Processing Systems*, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

Alexander Pondaven, Aliaksandr Siarohin, Sergey Tulyakov, Philip Torr, and Fabio Pizzati. Video motion transfer with diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Ryota Takatsuki, Sonia Joseph, Ippei Fujisawa, and Ryota Kanai. Decoding vision transformers: the diffusion steering lens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 4858–4863, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Michael Toker, Hadas Orgad, Mor Ventura, Dana Arad, and Yonatan Belinkov. Diffusion Lens: Interpreting Text Encoders in Text-to-Image Pipelines. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9713–9728. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.524.

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.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5288–5296, 2016.

Shi-Qi Yan, Kai-Xuan Ding, Chao-Hong Tan, Qian Chen, Wen Wang, Xiangang Li, and Zhen-Hua Ling. Short horizons and sparse concepts: a mathematical view of the readout in the J-lens. *arXiv preprint arXiv:2608.25347*, 2026.

AI USE STATEMENT

Large language models (LLMs) were used to assist manuscript drafting, language refinement and organization, as well as research tasks involving experimental design and/or data analysis. The authors ground scientific claims in the mathematical definitions, measured intervention responses and experimental evaluations. LLM assistance supported the presentation and analysis of these results. The authors retain responsibility for checking LLM-assisted text, design choices and analyses against the underlying records, and for the accuracy, originality and integrity of the complete manuscript.

ETHICS STATEMENT

This work studies computational interventions in frozen diffusion transformers, using ImageNet and COCO inputs, synthetic latents, custom prompts and the selected captioned-video cohort described in the experimental protocols. It does not involve recruiting human participants or conducting animal experiments. Existing images and videos should not be interpreted as consented identity data, and the reported response maps do not establish privacy protection, fairness or removal of biases inherited from the data and pretrained models. Generative interventions can facilitate deceptive synthetic media, including misattributed events and nonconsensual depictions. The experiments include no identity-specific safeguards, abuse filtering, provenance marking, watermarking or access control; none should be inferred from the control metrics. Deployment would require a separate risk assessment and safeguards such as provenance disclosure and restrictions on sensitive uses, followed by evaluation of their effectiveness. Data and checkpoint reuse must respect applicable licenses, access conditions and privacy constraints. Evaluations use existing datasets and declared constructed editing targets.

REPRODUCIBILITY STATEMENT

The method in Section 3 specifies the conditional mean suffix Jacobian, calibration and intervention-response measurements. Appendix A documents model and evaluation conventions; Appendices F and P give the shared controller and RCRE algorithms. Independent image and video control cohorts are described in Appendices K and N. Reproduction must preserve each model's conditioning, output parameterization, precision, sampler, seeds, source budget and timing boundaries.

The DiT quality comparison uses 50,000 generated images per arm and 50,000 ImageNet validation references for FID/KID. A separate paired 1,000-image run measures sampling and decoding time. The video table separates the 1,024- and 4,096-pair FVD comparisons from the 5,000-video MSR-VTT VBench setting; Wan-generation values are three-seed means. These numeric records and displayed examples have distinct roles; Appendix A specifies their evaluation units. No uncertainty intervals accompany these generation metrics, and the package does not contain the pretrained checkpoints or every saved input. Reuse of code, data and pretrained assets is subject to their respective licenses.

A READER GUIDE AND EXPERIMENTAL PROTOCOLS

ST-Lens: Supplementary Material

The supplement follows the three contributions in the main paper. It supplies proofs, complete comparisons and supporting visual evidence. Main figures retain the central observation and editing examples; supplementary figures resolve parameter dependence, additional methods and individual cases.

Table 6: Evidence map. Sample counts describe the unit used in each experiment.

Contribution	Supporting evidence
ST-Lens observable	Calibration, mean-action and pooling proofs; numerical fidelity; depth/noise and held-out response comparisons.
Generation-process editing	Shared controller; full image and Wan comparisons; native residual algorithm; all motion variants and command audits.
Generation performance	Image guidance protocol and paired gallery; condition diagnostics and MSR-VTT evaluation protocol.

Model and output conventions. DiT-XL/2-256 and PixArt-α-512 use four-channel epsilon fields in the matched control experiments. Wan2.1-T2V-1.3B controls conditional velocity. Residual depth identifies a completed transformer block. Denoising indices identify sampler iterations, while τ or frame f identifies time within one video. Historical routing diagrams explicitly indicate any inclusion of learned-variance channels or single-step x_0 decoding.

Full evaluation and detailed records. The image and video studies have distinct evaluation units. The DiT quality comparison uses 50,000 generated images per arm against 50,000 ImageNet validation references; timing uses a separate 1,000-image paired run. Video generation uses the FVD and VBench cohorts stated below; editing uses declared prompts and seeds. Displayed frames are selected views of complete rollouts and serve as qualitative examples.

Independent units in the detailed comparisons. The fixed-axis observer comparison has 32 held-out DiT contexts, three source dictionaries and eight axes: 768 intervention measurements. Image control has 576 DiT and 216 PixArt conditions per method. Fresh PixArt confirmation uses 32 disjoint-prompt trajectories. Fresh Wan control has four videos and 48 nested conditions. The decoded ball studies use a constructed scene and individually declared seeds. Image-generation quality uses 50,000 generated images per arm and 50,000 ImageNet validation references; runtime uses a separate paired 1,000-image run. The main-text video table reports two separate FVD cohorts of 1,024 and 4,096 pairs and the 5,000-video VBench setting, with Wan-generation values averaged across three seeds.

Why MSR-VTT and how VBench is used. MSR-VTT provides human-described real clips across varied content (Xu et al., 2016). Generating from its captions keeps prompts matched between vanilla Wan and each covariance policy, while the corresponding real clips provide a dataset-matched FVD reference. Because these captions are outside VBench's official prompt suite, we evaluate the generated clips with VBench's custom-input mode on subject consistency, background consistency, dynamic degree, aesthetic quality and imaging quality (Huang et al., 2024). These are within-study scores for matched generated videos, not official VBench prompt-suite leaderboard scores.

Interpreting video-quality and motion summaries. The 1,024-pair FVD comparison and the 5,000-video VBench comparison have different evaluation units. All reported Wan-generation values are means across three random seeds, including the unchanged vanilla baselines. FVD compares distributions in the declared I3D feature space; retained flow compares mean motion magnitude with vanilla within the FVD cohort. The latter is 90.17% for advected sharing (.10), or 9.83% below vanilla. VBench dynamic degree is a separate video-level score. Its reported advected-minus-vanilla difference is only .0019, whereas the other four VBench dimension differences range from .0074 to

.0198. Together, these metrics describe improved distribution match with near-baseline motion magnitude; the small dynamic-degree difference is best read descriptively. The two FVD cohorts also have different mean vanilla scores (4551.09 and 2389.21); reductions compare each policy with its matching cohort's mean vanilla, not pooled absolute FVD values. Three-seed averaging addresses observed run-to-run fluctuation; individual seed scores and uncertainty intervals are not part of the reported summary.

Matched computation. Within a comparison, all methods share the checkpoint, source space, target, budget and numerical backend. Online suffix cost includes baseline, probes, candidate verification, correction and failure feedback. Reported suffix-call comparisons omit model loading, shared prefix extraction, offline calibration and decoding. These components are not normalized into an end-to-end cost against Full8; reported call savings should be read as online suffix savings only. Every reported mean retains failed targets.

B Calibration and numerical fidelity

For calibration contexts x_i at a fixed block/noise regime, estimate $\widehat{J}D = n^{-1}\sum_i J_{x_i}D$ on a declared source dictionary D. Centered finite differences provide one implementation:

$$J_x d \simeq \{F_x(h+sd) - F_x(h-sd)\}/(2s).$$

Exact JVPs independently check the derivative actions. For a randomized operator sketch, set $Y = n^{-1}\sum_i J_{x_i}\Omega$, $Q = \mathrm{orth}(Y)$ and $B = n^{-1}\sum_i Q^\top J_{x_i}$. The approximation is QB, computed with JVPs and VJPs (Halko et al., 2011; Baydin et al., 2018).

Table 7: Numerical fidelity of measured response columns. Dispersion labels follow the recorded protocol.

Test	Result
64-fold FD step range	Relative error .054843–.054855 on four calibration/four test contexts.
FD versus exact JVP	512 direction–context pairs: cosine $.9984 \pm .0001$, error $.0571 \pm .0016$.
Averaged eight-probe action	Cosine .99997 against the exact mean action.

All three audits assess the response columns used by ST-Lens. Calibration averages these columns across contexts; refresh measures the columns at the current state. Both supply the same native-output coordinates to the solver. RCRE denotes the downstream coefficient optimization, not the mean estimator.

Step-size control. If the third directional derivative is bounded by M, centered differences have error at most $Ms^2/6$. Endpoint error bounded by e_{eval} adds e_{eval}/s. Across r columns, a Frobenius bound gives $\sqrt{r}(Ms^2/6 + e_{\mathrm{eval}}/s)$. This balance motivates the recorded step halving and autodiff checks. Spatial DiT tests give maximum differences .103% under step halving and .043% between batched and single-context evaluation.

Amortization. Spatial DiT mean calibration takes .6201 s, a local eight-direction operator .0770 s and a cached query .0272 ms on RTX 4070: operator-only break-even is 8.05 contexts. PixArt block-21 calibration takes 3.4415 s, local measurement .4303 s and a query .0400 ms. These boundaries isolate observation cost.

C Why a conditional mean predicts reusable action

Fix a checkpoint, block, output grid and noise regime. For a deterministic write δ, let $Y_X = J_X\delta$, $\mu = \mathbb{E}Y_X$ and $V = \mathbb{E}\|Y_X - \mu\|^2$, with finite second moments. Expanding a square gives

$$\mathbb{E}\|Y_X - v\|^2 = V + \|\mu - v\|^2. \tag{9}$$

The cross term vanishes because $\mathbb{E}(Y_X - \mu) = 0$. Therefore $\bar{J}\delta$ is the unique optimal context-independent response under squared first-order loss. The statement concerns the response to a fixed write, extended columnwise to the declared proposal dictionary.

For n independent calibration contexts, independent of the test context, $\widehat{\mu} = n^{-1}\sum_i Y_{X_i}$ has variance V/n. Consequently

$$\mathbb{E}\|Y_X - \widehat{\mu}\|^2 = (1 + n^{-1})V. \tag{10}$$

Calibration reduces estimation variance; the context-dependent action retains the irreducible term V. Repeated tokens or probes within one context do not increase n in this identity.

Conditioning on denoising regime. For regime B, the total-variance identity is

$$\mathbb{E}\|Y - \mathbb{E}Y\|^2 = \mathbb{E}\|Y - \mathbb{E}[Y \mid B]\|^2 \tag{11}$$

$$+ \mathbb{E}\|\mathbb{E}[Y \mid B] - \mathbb{E}Y\|^2. \tag{12}$$

Conditioning removes the between-regime mean variation from population prediction risk. Finite calibration introduces a bias–variance choice: finer bins can reduce within-bin variation while providing fewer independent observations in each bin. Depth and noise remain explicit indices in ST-Lens.

Transfer between context distributions. Let calibration follow P and a new test context follow Q, with compatible operator coordinates. Then independence and the same expansion yield

$$\mathbb{E}\|J_Q\delta - \widehat{J}_P\delta\|^2 = V_{Q,\delta} + V_{P,\delta}/n + \|(\bar{J}_Q - \bar{J}_P)\delta\|^2. \tag{13}$$

The final term measures shift of the mean action. This gives an operational reason for a current-state probe when a new video differs from calibration. No Gaussian assumption is needed for these derivative-action identities. An affine predictor of absolute outputs additionally includes its calibrated intercept and is evaluated as a separate approximation.

D FINITE-RESPONSE ERROR AND COORDINATE IDENTIFIABILITY

Finite-response bound. If F is differentiable along $h + s\delta$ and its Jacobian is L-Lipschitz on that segment, the integral remainder gives

$$F(h + \delta) - F(h) = J_h\delta + \mathcal{R}, \qquad \|\mathcal{R}\| \le L\|\delta\|^2/2.$$

For an estimated conditional mean, adding and subtracting J_h and $\bar{J}$ separates context variation, calibration error and curvature. With an orthonormal dictionary U and $\delta = Ua$,

$$\|\Delta F - \widetilde{A}a\| \le (\|(J_h - \bar{J})U\| + \|\bar{J}U - \widetilde{A}\|)\|a\| + L\|a\|^2/2.$$

The norms and Lipschitz assumption define the scope of this local statement.

Regional identifiability under pooling. For a dictionary response $G \in \mathbb{R}^{m\times r}$ and pooling matrix P, $q^\top G$ is determined by PG for every G exactly when $q \in \mathrm{range}(P^\top)$. If $q = P^\top u$, then $q^\top G = u^\top PG$. Otherwise set $z = (I - P^\dagger P)q \neq 0$. The pair $G_\pm = G_0 \pm zv^\top$ satisfies

$$PG_+ = PG_-, \qquad q^\top(G_+ - G_-) = 2\|z\|^2 v^\top.$$

Thus a pooled readout leaves a nonzero regional ambiguity. The proof applies to spatial or temporal pooling in the corresponding output coordinates.

Constructing opposite regional writes. Let $B = PG$, $a = G^\top q$ and $N = I - B^\dagger B$. By Cauchy–Schwarz,

$$\max_{Bc=0,\|c\|\le 1} a^\top c = \|Na\|.$$

When $Na \neq 0$, $c_* = Na/\|Na\|$ attains the bound; $\pm c_*$ give opposite regional responses with equal zero pooled prediction. The DiT test uses three eight-direction dictionaries, eight calibration contexts, 32 held-out contexts and a .4% hidden-RMS write at block $21/t = 510$. The target contrasts a center 8×8 square with the remaining grid.

Table 8: Measured pooled-null writes, 96 responses per row.

Write	Contrast	Correct sign	Pooled energy%	Cosine
Null +	.008507	96/96	.004509	.998055
Null −	−.008526	96/96	.004435	.998049

E PROTECTED RESPONSE AND LOCAL REACHABILITY

For whitened residual coordinates $D^\top D = I$, set $f(b) = F(h + Db) - F(h)$. Let T and P select target and protected output coordinates, each with norm at most one, and let A be the signed response. In the exact linear model, let N orthonormally span $\ker(PA)$ and set $H = TAN$. A protected edit $TAb = g$, $PAb = 0$, $\|b\| \le \rho$ exists iff

$$(I - HH^\dagger)g = 0, \qquad \|H^\dagger g\| \le \rho.$$

Every protected coefficient is $b = Nc$, so the target is $Hc = g$ and the minimum-norm solution is $b_* = NH^\dagger g$. The criterion follows from orthogonality of the pseudoinverse solution and $\ker H$.

Assume $\|Df(0) - A\| \le \varepsilon$ and Df is L-Lipschitz on the radius-ρ ball. The integral remainder implies

$$\|Pf(b)\| \le \|PAb\| + \varepsilon\|b\| + \tfrac{L}{2}\|b\|^2. \tag{14}$$

This yields a convex protected fit:

$$\min_{\|b\| \le \rho} \|TAb - g\|^2 + \lambda\|b\|^2 \quad \text{s.t.} \quad \|PAb\| + \varepsilon\|b\| + \tfrac{L}{2}\|b\|^2 \le \tau. \tag{15}$$

To prove the bound, expand $f(b)$ about zero, bound the derivative mismatch by $\varepsilon\|b\|$ and integrate $Lt\|b\|^2$ over $t \in [0, 1]$. The certificate applies to the current native field under these assumptions. CFG multiplies its change when the unconditional output is fixed.

Numerical verification. One hundred linear systems with eight source coordinates, a rank-three protected map and four target coordinates verify feasibility and minimum norm. Maximum target/protected residuals are 2.17×10^{-14} and 1.51×10^{-14}. Unreachable targets are included explicitly. One hundred nonlinear systems use $f(b) = (A + E)b + v(b^\top Qb)/2$, $\|E\| = .02$, $\|v\| = 1$ and $Q = Q^\top$, $\|Q\| = .15$. With $\rho = .5$, $\tau = .12$ and $\lambda = .01$, all fits satisfy the bound; maximum protected response is .116171 and constraint excess is below 9×10^{-12}.

The generated-video protection variant uses a soft outside-response penalty. The exact constants above belong to the controlled mathematical validation; uniform Wan curvature/error constants remain uncalibrated.

F SHARED CONTROLLER IMPLEMENTATION

The observer supplies response information while targets, coefficient solver, norm budget and feedback are shared within each comparison. An orthonormal source dictionary D_s maps coefficient u to a residual write. The native response estimate A is used to minimize outside-region energy subject to prescribed signed target gain and a source norm bound (Equation 6). Infeasible gains use the bounded maximum-gain direction, followed by actual suffix evaluation.

Rank-six image observation. The calibrated subspace contains the mean control direction, its target gradient, and leading variation modes of normalized calibration-context controls after removing those two directions. Six one-sided differences measure the action $Y \simeq J_X D_s U$ on each new state. The common solver then performs measured gain correction, budget clipping and failure feedback. All methods receive the same target information and correction procedure.

Adaptive video observation. Four initial directions come from a calibration ensemble of control and target-gradient vectors. For $U^\top U = I$, completion uses

$$A_\gamma = YU^\top + \gamma G(I - UU^\top), \qquad \gamma = .25.$$

The measured directions replace their cached action; the remaining complement retains a weaker estimate. If measured target error exceeds 10%, four remaining orthogonal directions are measured and the common solver is repeated. Spectral and QKV baselines share this trigger.

Endpoints and pairing. Target error is $e = |q^\top \Delta F - g_*|/g_*$. Success is $e \leq .1$; outside response is $L = \|P_{\text{out}}\Delta F\|$. Unconditional means retain all failures. Paired ordering first compares success, then outside response when both succeed, then target error when both fail. Image resampling averages dictionary repeats within context before paired context bootstrap. Multiplicity is adjusted over the declared 36- or 18-comparison family.

The full eight-direction reference spends additional probes. Realized online counts include verification and feedback, so a failed condition can cost more than an immediate success. Offline calibration and shared prefix execution are separate from these suffix-call counts.

G BASELINE MECHANISMS AND COMPARISON DESIGN

Table 9: Operations evaluated on the common residual-to-native-field task.

Method	Mechanism and information supplied
Direct head	Apply the final head to the intermediate residual, following direct intermediate readout.
Frozen path	Hold later additive residual contributions fixed and propagate the write through the head.
Conditional ST-Lens	Average signed suffix derivative actions within the declared block/noise regime.
Spatial pooling	Pool and broadcast coordinates of the same mean operator; isolates coordinate loss.
Pullback style	Leading or target-aligned modes of the dictionary-restricted response Jacobian.
QKV directions	Eigenvectors of the input-coordinate weight Gram restricted to the common dictionary.
Random / Full8	Random reduced subspace / measurement of all eight common directions.

Direct-head and frozen-path rows adapt the readout principles of the logit lens and DSL (nostalgebraist, 2020; Takatsuki et al., 2025). Pullback-style and localized nulling use the geometric mechanisms of Park et al. (2023); Chen et al. (2024b). They operate on the same residual source and native target as ST-Lens.

For $J = U\Sigma V^\top$, the rank-r spectral target direction is $c_r \propto V_r V_r^\top J^\top q$. QKV uses $W_Q^\top W_Q + W_K^\top W_K + W_V^\top W_V$ in PyTorch column-input coordinates, restricted to the source dictionary (Anand et al., 2026). Its signs and target alignment use response calibration. This gives both direction families access to the requested target.

Projected J-Lens retains summed-target/averaged-source transport for the specified final-hidden response task (Gurnee et al., 2026). The vocabulary readout of the original language-model method is outside this comparison's output space. The evaluated rows are precisely these mechanism adaptations. RCRE's scalar, diagonal, minimax and protected variants are internal design ablations, with their objectives specified below.

Related feature interventions include ROME's weight editing (Meng et al., 2022) and temporal sparse feature extraction in TIDE (Huang et al., 2026). Their goals clarify the distinction between a feature coordinate, a parameter update and the downstream response observed here.

H COMPLETE CONDITIONING ABLATION AT BLOCK 7

Thirty-two new DiT contexts have classes and seeds disjoint from calibration and earlier confirmation cohorts. Each reusable operator uses eight calibration observations. Balanced assignments share observations across time, depth or both; the spatial ablation pools and broadcasts output positions. All rows use one constrained solver and one measured gain correction. Four fixed mixture directions test prediction; three target regions test control.

Table 10: Complete block-specific factorial. Success averages the three regional targets and three dictionaries at 10% target tolerance.

Block/time	Budget	Operator	Prediction error	Success%
7/102	0.004	Conditional	0.8734	42.36
7/102	0.004	Time-shared	0.8533	42.71
7/102	0.004	Depth-shared	0.8522	42.71
7/102	0.004	Global-shared	0.8473	42.36
7/102	0.004	Spatial-pooled	1.0001	28.82
7/102	0.004	DSL-frozen-path	0.8509	43.75
7/102	0.004	Local-J	0.0517	100.00
7/102	0.020	Conditional	0.8752	32.64
7/102	0.020	Time-shared	0.8552	32.64
7/102	0.020	Depth-shared	0.8542	33.68
7/102	0.020	Global-shared	0.8492	25.00
7/102	0.020	Spatial-pooled	1.0001	21.53
7/102	0.020	DSL-frozen-path	0.8528	37.15
7/102	0.020	Local-J	0.2547	96.88
7/510	0.004	Conditional	0.4070	47.57
7/510	0.004	Time-shared	0.7747	48.96
7/510	0.004	Depth-shared	0.3934	54.17
7/510	0.004	Global-shared	0.7031	47.22
7/510	0.004	Spatial-pooled	1.0001	29.17
7/510	0.004	DSL-frozen-path	0.3836	53.82
7/510	0.004	Local-J	0.0291	100.00
7/510	0.020	Conditional	0.4147	38.54
7/510	0.020	Time-shared	0.7764	39.24
7/510	0.020	Depth-shared	0.4017	39.24
7/510	0.020	Global-shared	0.7056	37.50
7/510	0.020	Spatial-pooled	1.0001	18.06
7/510	0.020	DSL-frozen-path	0.3922	44.44
7/510	0.020	Local-J	0.1442	90.97

The two amplitudes test finite-write sensitivity. Local derivatives are remeasured on each context, while all cached rows use their fixed calibration. The shallow and late tables jointly show the operating-regime dependence.

I COMPLETE CONDITIONING ABLATION AT BLOCK 21

Thirty-two new DiT contexts have classes and seeds disjoint from calibration and earlier confirmation cohorts. Each reusable operator uses eight calibration observations. Balanced assignments share observations across time, depth or both; the spatial ablation pools and broadcasts output positions. All rows use one constrained solver and one measured gain correction. Four fixed mixture directions test prediction; three target regions test control.

Table 11: Complete block-specific factorial. Success averages the three regional targets and three dictionaries at 10% target tolerance.

Block/time	Budget	Operator	Prediction error	Success%
21/102	0.004	Conditional	0.4803	69.10
21/102	0.004	Time-shared	0.4718	72.22
21/102	0.004	Depth-shared	0.5918	63.89
21/102	0.004	Global-shared	0.5487	62.85
21/102	0.004	Spatial-pooled	1.0000	25.35
21/102	0.004	DSL-frozen-path	0.5339	60.76
21/102	0.004	Local-J	0.0174	96.88
21/102	0.020	Conditional	0.4860	65.62
21/102	0.020	Time-shared	0.4752	68.40
21/102	0.020	Depth-shared	0.5977	56.25
21/102	0.020	Global-shared	0.5525	61.11
21/102	0.020	Spatial-pooled	1.0000	25.35
21/102	0.020	DSL-frozen-path	0.5350	53.47
21/102	0.020	Local-J	0.0875	92.36
21/510	0.004	Conditional	0.0626	98.26
21/510	0.004	Time-shared	0.4205	97.22
21/510	0.004	Depth-shared	0.1388	92.01
21/510	0.004	Global-shared	0.6401	91.32
21/510	0.004	Spatial-pooled	0.9999	38.89
21/510	0.004	DSL-frozen-path	0.0981	97.92
21/510	0.004	Local-J	0.0022	100.00
21/510	0.020	Conditional	0.0632	97.57
21/510	0.020	Time-shared	0.4207	95.83
21/510	0.020	Depth-shared	0.1391	90.28
21/510	0.020	Global-shared	0.6403	86.11
21/510	0.020	Spatial-pooled	0.9999	38.19
21/510	0.020	DSL-frozen-path	0.0984	95.83
21/510	0.020	Local-J	0.0114	99.31

The two amplitudes test finite-write sensitivity. Local derivatives are remeasured on each context, while all cached rows use their fixed calibration. The shallow and late tables jointly show the operating-regime dependence.

J IMAGE CONTROL ACROSS REGIONS AND BUDGETS

The matched-rank study compares three reduced-rank competitors on two backbones, three target regions and two source budgets: 36 paired comparisons. Each mean retains every target failure. ST-Lens lowers all 36 outside-response means, with adjusted positive lower bounds in 35. The aggregate 792 target conditions comprise 576 DiT and 216 PixArt controls.

Table 12: All 36 matched-rank comparisons underlying Figure 5. Reduction is percent; bounds are in native outside L2 units. The intervals use paired context resampling and a 36-comparison adjustment.

Model	Region / budget	Baseline	Reduction	Lower	Upper
DiT	left / 0.004	Pullback	20.65	0.00044	0.00384
DiT	left / 0.004	QKV	13.18	0.00030	0.00183
DiT	left / 0.004	Random	73.19	0.00350	0.03023
DiT	left / 0.020	Pullback	24.30	-0.00936	0.03676
DiT	left / 0.020	QKV	25.48	0.00110	0.04024
DiT	left / 0.020	Random	67.35	0.01529	0.13818
DiT	right / 0.004	Pullback	52.87	0.00096	0.01788
DiT	right / 0.004	QKV	67.71	0.00194	0.02366
DiT	right / 0.004	Random	88.31	0.02466	0.05158
DiT	right / 0.020	Pullback	59.15	0.00633	0.10140
DiT	right / 0.020	QKV	71.31	0.01566	0.13366
DiT	right / 0.020	Random	89.46	0.13814	0.28794
DiT	center / 0.004	Pullback	21.34	0.03944	0.05457
DiT	center / 0.004	QKV	28.87	0.06184	0.07860
DiT	center / 0.004	Random	15.45	0.02646	0.03731
DiT	center / 0.020	Pullback	21.35	0.19800	0.27328
DiT	center / 0.020	QKV	29.18	0.31086	0.40612
DiT	center / 0.020	Random	15.53	0.13417	0.18654
PixArt	left / 0.004	Pullback	86.23	0.00293	0.05373
PixArt	left / 0.004	QKV	95.43	0.04725	0.08196
PixArt	left / 0.004	Random	79.85	0.00528	0.02212
PixArt	left / 0.020	Pullback	86.32	0.01397	0.25916
PixArt	left / 0.020	QKV	95.43	0.23535	0.39795
PixArt	left / 0.020	Random	79.41	0.02512	0.10570
PixArt	right / 0.004	Pullback	38.96	0.00029	0.01078
PixArt	right / 0.004	QKV	80.93	0.00059	0.04171
PixArt	right / 0.004	Random	60.16	0.00201	0.01059
PixArt	right / 0.020	Pullback	38.48	0.00142	0.05116
PixArt	right / 0.020	QKV	81.38	0.00274	0.20793
PixArt	right / 0.020	Random	61.43	0.00967	0.05545
PixArt	center / 0.004	Pullback	11.84	0.00792	0.02075
PixArt	center / 0.004	QKV	7.86	0.00461	0.01359
PixArt	center / 0.004	Random	7.44	0.00559	0.01113
PixArt	center / 0.020	Pullback	11.74	0.03913	0.10175
PixArt	center / 0.020	QKV	7.85	0.02281	0.06709
PixArt	center / 0.020	Random	7.37	0.02734	0.05463

Differences are baseline minus ST-Lens outside L2. Three dictionary seeds are averaged within each context before 50,000 paired-context resamples. The percentile tails are $0.05/72$ and $1 - 0.05/72$. The cohorts contain 32 DiT and 12 archived PixArt contexts; the new-prompt cohort follows.

K Frozen confirmation on new image prompts

Thirty-two newly generated PixArt trajectories use prompts disjoint from the original cohort. Eight calibration contexts, three dictionaries, rank six, solver and feedback remain fixed. All five controller variants are evaluated on the same new trajectories. Eighteen matched-rank comparisons have lower mean outside response under ST-Lens and positive adjusted lower bounds.

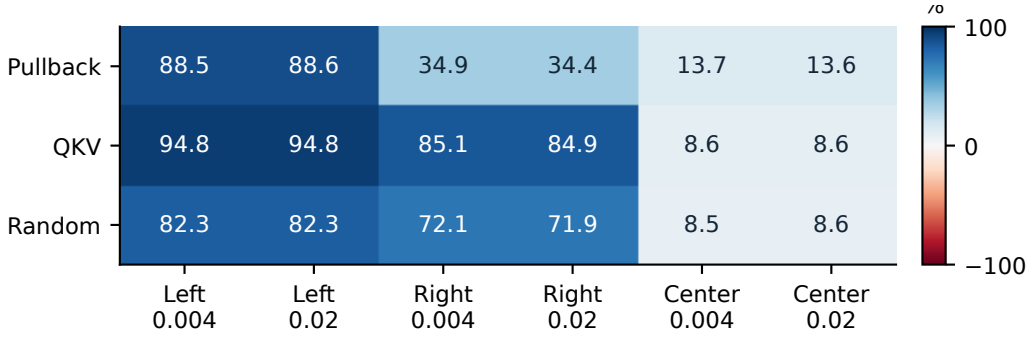

Figure 6: Every region/budget/baseline comparison in the frozen new-prompt PixArt cohort. Calibration and controller parameters are fixed before this evaluation.

ST-Lens and Full8 both attain 99.65% of prescribed targets, with 9.021 and 11.021 mean suffix evaluations. ST-Lens uses 18.15% fewer online calls; its outside response is .02–2.38% higher across the six conditions. This cohort tests the same accuracy–measurement trade-off on new prompts. The context is the resampling unit; source-dictionary repeats are averaged within each context before uncertainty analysis.

L Spatial response visualizations

The following views connect the native-field response to its spatial organization. A source patch can influence other output patches after the frozen suffix. The depth and architecture alter this propagation.

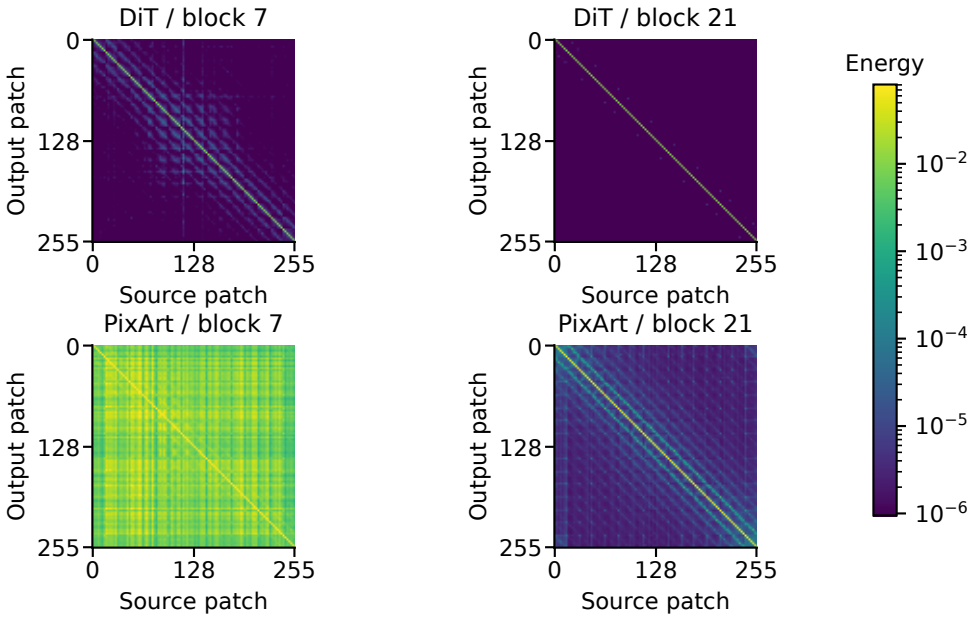

Figure 7: Context-averaged local Jacobian-block energy at the recorded depths. This statistic averages squared local responses and includes the exporter's native output channels. Off-diagonal fractions are 4.9/0.3% for DiT and 80.4/1.6% for PixArt (blocks 7/21).

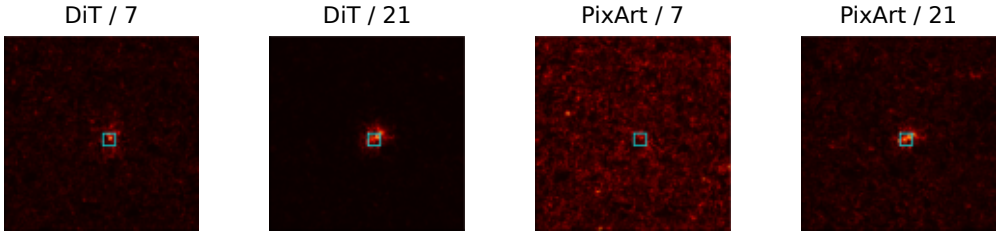

Figure 8: A center-patch intervention decoded into a single-step x_0 estimate at $t = 100$. Cyan marks the source region. The four localization fractions are 2.25/10.19% for DiT and .68/2.96% for PixArt, at blocks 7/21.

Native energy and decoded localization use their respective output coordinates.

M Depth and denoising dependence

The depth–noise sweep calibrates each model/regime separately. DiT cells use 128 calibration and 128 held-out real ImageNet contexts. PixArt cells use 16+16 synthetic latents with the fixed caption "a photo of a cat". Figure 3 in the main paper shows the full sweep on a shared cosine scale. It connects conditional observation to the choice of block and noise regime; this section supplies the cohort specification and the complementary dictionary-mode audit.

At block 21/$t = 500$, cosine is .998 for DiT and .985 for PixArt; at $t = 900$ it is .999 and .990. Earlier low-noise states give .501 and .460 at block 7/$t = 100$. PixArt block 14 decreases from .966 to .961 in the last two bins. The overall regime dependence motivates explicit block/noise conditioning and current-state refinement in difficult states.

An independent disjoint-class dictionary-mode audit recovers the leading direction with cosine .9997 and the first eight modes with absolute cosine at least .997. These mode results concern the tested source dictionary and complement the columnwise numerical checks in Appendix B.

N Complete fresh Wan control comparison

Four new trajectories depict a spinning dancer and a tracked train, with seeds 9711/9712 for each prompt. Generation uses 49 frames at 416×240, 50 steps, CFG 6 and shift 8. At completed block 21, sampler indices 25/45, six masks yield 48 paired conditions and 528 method–condition evaluations. The dictionary crosses four quadrants with two temporal halves. Source budget is .02 of calibrated hidden scale, with target coefficient .85.

Table 13: Frozen control on four new Wan videos (48 conditions per method). Errors and P90 are percentages; L/Full8 is the paired geometric mean outside-response ratio. Calls are complete online suffix evaluations. Pullback and QKV are common-dictionary mechanism adaptations (Park et al., 2023; Anand et al., 2026). All low-budget variants allow at most 13 calls; Full8 allows 17. Bold method names identify ours; bold numerical values mark the best low-budget values (including ties), with the Full8 reference separated.

Method	$e \leq 10\%$	$e \leq 2\%$	Error ↓	P90 ↓	L/Full8 ↓	Calls ↓
Pullback, 4	22/48	18/48	22.913	64.274	1.2979	10.250
Pullback + completion	35/48	28/48	13.665	43.721	1.1665	8.625
Pullback + adaptive	**41/48**	33/48	3.600	12.765	1.1304	8.625
QKV, 4	27/48	22/48	16.678	46.355	1.2688	9.625
QKV + completion	32/48	25/48	14.238	44.176	1.2344	9.000
QKV + adaptive	**41/48**	33/48	3.686	12.766	1.1628	9.000
Random, 4	30/48	26/48	14.625	45.000	1.3232	9.250
Random + completion	36/48	28/48	13.705	48.556	1.1676	**8.500**
ST-Lens, regularized 4	35/48	29/48	9.505	30.298	1.1474	8.708
ST-Lens, adaptive 4→8	**41/48**	**35/48**	**3.363**	**12.598**	**1.1017**	8.750
Full8 reference	41/48	39/48	2.679	12.742	1.0000	11.875

ST-Lens has lower task-mean error than both adaptive comparators on moving tube, alternating frames, signed dual patch and temporal reversal. Occluded zigzag and expanding disk give mixed orderings. All seven ST failures also occur under Full8 and exhaust ST's source budget. The comparison therefore retains target feasibility alongside measurement cost and outside response. Full8 is a higher-measurement reference with a 17-call ceiling; the other methods have a common 13-call ceiling.

O When video observation requires local measurement

Six complete Wan generations use bicycle, sailboat and umbrella prompts with seeds 8721/8722; four calibration trajectories from three cat prompts remain fixed. At block 21 and indices 25/45, five regions, two channel directions and two amplitudes (.004/.02 hidden RMS) produce 240 conditions. All methods predict the same final-hidden residual response on 5,070 tokens.

Table 14: Mean relative response error on new Wan trajectories; lower is better. Each region contains 48 correlated conditions. Local FD spends two online suffix calls per direction; the calibrated methods spend none for prediction. The cached ST columns and hybrid require 80 calibration suffix calls per phase versus 16 for projected J-Lens, using the same four calibration trajectories.

Region	ST	J projection	ST+J	Frozen	Local FD
Global	0.9278	0.8584	0.8726	1.1756	0.0862
Left	0.9426	0.8990	0.8892	1.2055	0.0635
Right	0.9308	0.9036	0.8871	1.1983	0.0637
Early	0.9457	0.9140	0.8820	1.1817	0.0810
Late	0.9570	0.9125	0.8948	1.1698	0.0676

Mean errors are .9408 for cached ST, .8975 for projected J-Lens, .8851 for the fixed half/half hybrid and .0724 for current-state central differences. Against projected J-Lens, ST improves on 11/240 conditions and the hybrid improves on 181/240. Each local derivative uses two online suffix calls per direction. Calibration uses 80 calls per phase for ST/hybrid and 16 for the projected readout. Cached calibrated columns transfer poorly to these held-out Wan trajectories. The decoded editor therefore refreshes all four proposal columns before optimization. Calibration and current-state probe costs are reported separately.

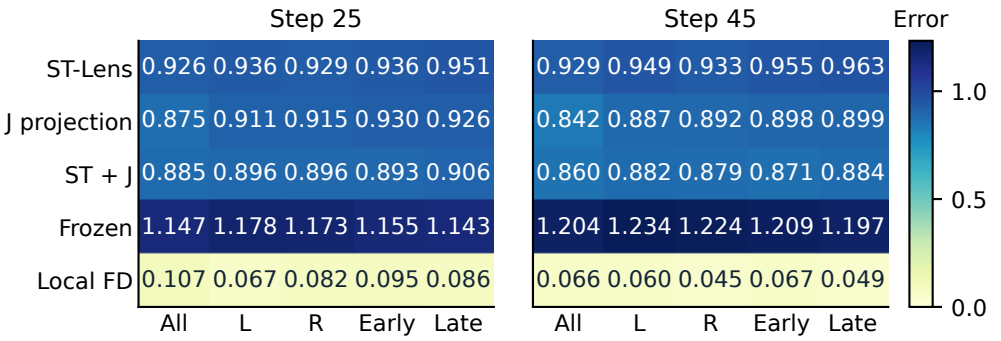

Figure 9: Source-region results separated by the two sampled phases, with identical color scales and all amplitudes, directions, prompts and seeds retained.

P Response-coupled residual editing algorithm

At completed block 28, a terminal-head inverse supplies a proposal $d = h^*_{30} - h_{30}$. Four disjoint masks partition its 13 latent-time positions into $[0, 3)$, $[3, 6)$, $[6, 9)$ and $[9, 13)$, defining $d_k = M_k d$. Each temporal position has 240 tokens at the controlled resolution. For conditional/unconditional velocities c, u and CFG $s = 6$, set

$$v = u + 6(c - u), \qquad v^* = v + \alpha[(x_\sigma - z^*)/\sigma - v], \tag{16}$$

$$c^* = u + (v^* - u)/6, \qquad r = \mathrm{vec}(c^* - c), \tag{17}$$

where the input reference supplies z^* and $\alpha = .9 \min(1, (i-2)/4)$ for active indices $i = 3, \ldots, 24$. One suffix call per group measures $a_k = \mathrm{vec}[F(h + .5d_k) - F(h)]/.5$. With $A = [a_k]$, $C = A^\top A$, $g = A^\top r$ and $s_k = \mathrm{RMS}(d_k)^2$, solve

$$\min_{0 \le b_k \le 2} \tfrac{1}{2} b^\top (C + \lambda I) b - g^\top b, \quad \sum_k s_k b_k^2 \le B^2, \tag{18}$$

where $B = .025\,\mathrm{RMS}(h)$ and $\lambda = .001\,\mathrm{tr}(C)/4$. The source masks are disjoint, so their residual energies add exactly. The response energy additionally has signed cross terms $2\sum_{i<j} b_i b_j C_{ij}$. These determine reinforcement or cancellation after the suffix.

1. Run the ordinary prefix and obtain the reference-conditioned target.
2. Split the head-inverse proposal and measure the four signed responses.
3. Solve the bounded coefficient problem using the full response Gram.
4. Evaluate full and half candidate writes with the real suffix.
5. Select the lowest actual target error among both candidates and zero write; continue with unchanged CFG, scheduler and VAE.

Twenty-two active steps use 132 suffix calls per editor. Each rollout also uses 50 full forwards plus one full-forward/suffix-equivalence audit. References are input conditions; reported RGB arrays are fresh native decoded outputs. The frozen implementation remeasures signed suffix responses at each evolving state.

Q Residual placement and denoising windows

The common motion backend tests how placement and timing affect a supplied trajectory. The full-dimensional inverse of the actual normalized native head preserves its declared zero-mean and norm

constraints. At block 28, three transported corrections are checked with suffix scales $1, .5, .25, 0$. Seed 940 is fixed across the spatial/window sweep; seed 941 across the time-law sweep. Generation uses 49 frames, 320×192, 16 fps, 25 UniPC steps, shift 8 and CFG 6. The prompt specifies a red ball moving left to right.

Table 15: Complete spatial and denoising-window sweep, seed 940. RMSE is largest-red-component trajectory error; PSNR is fidelity to the supplied reference. Write is maximum relative residual RMS.

Condition	RMSE (px)	PSNR (dB)	Write (%)	Sampler (s)
No write	127.105	15.03	0.00	41.76
Rightward, B30	0.218	40.04	2.00	42.70
Leftward, B30	0.179	40.10	2.16	42.61
Away, B30	0.170	40.93	2.01	42.36
Toward, B30	0.174	41.00	1.92	42.15
Leftward, B30 early	14.692	21.13	2.16	41.76
Leftward, B30 late	0.184	40.08	2.52	41.75
Leftward, B30 last	0.320	33.79	8.20	41.54
Leftward, B28	0.173	40.15	4.16	53.81

Table 16: Complete time-law sweep at block 28, seed 941. All controlled clips satisfy the frozen motion criteria; the no-write condition does not.

Condition	RMSE (px)	PSNR (dB)	Detected frames	Sampler (s)
No write	87.896	15.85	17/49	41.00
Return	0.200	39.07	49/49	53.97
Stop/resume	0.222	39.59	49/49	53.76
Accelerate	0.209	39.83	49/49	53.77

Early-only indices 3–10 allow trajectory drift; late indices 17–24 retain the command. Last-only correction at index 24 uses a larger write and has visible residual artifacts. These observations motivate repeated late-block measurement in RCRE. A hooked full forward and suffix agree within RMS 2.1×10^{-7} at the first intervention.

R TEMPORAL FITTING DESIGNS AND MATCHED ABLATION

All designs consume the same signed response columns and proposal groups. Scalar fitting sets $b = c\mathbf{1}$ while retaining the full Gram in its one-dimensional objective; diagonal fitting removes off-diagonal entries of C; Joint RCRE retains C and independently optimizes the coefficients. The budget-aware seed-944 comparison uses identical prompt, reference, noise, four probes and acceptance scales.

Table 17: Matched motion ablation (seed 944): four temporal probes, 2.5% residual RMS cap and 132 suffix calls per fit. Field RMS measures native target error; tracking and PSNR compare output with the shared reference.

Target	Fit	Field RMS ↓	Track px ↓	PSNR ↑	Time s ↓
Bounce	Scalar	0.04251	0.222	36.75	50.99
Bounce	Diagonal	0.04491	0.405	33.74	50.42
Bounce	Joint (RCRE)	0.04250	0.208	36.77	50.48
Stop/resume	Scalar	0.04098	0.224	37.23	50.40
Stop/resume	Diagonal	0.04339	0.284	34.05	50.44
Stop/resume	Joint (RCRE)	0.04098	0.223	37.17	50.44

Joint fitting decreases mean native target error by 5.46% against diagonal fitting. Its mean reference tracking error is .216 px versus .223 for the stronger scalar baseline, a negligible decoded difference. All six outputs contain 49 detected centers.

Coupling and coefficient freedom. With $C = A^\top A$ and $g = A^\top r$, scalar fitting minimizes $c^2\mathbf{1}^\top(C + \lambda I)\mathbf{1} - 2c\mathbf{1}^\top g$ under the same budget. Its objective contains the cross terms C_{ij}. Thus scalar is a coupled-response baseline with tied coefficients, whereas diagonal fitting is the cross-term ablation. Joint versus diagonal measures response coupling; Joint versus scalar measures the value of independent temporal strengths. This distinction explains why similar decoded Scalar and Joint outputs can coexist with a field-error benefit over the diagonal fit:

the full Gram accounts for cross-time interactions removed by the diagonal fit. Scalar fitting constrains all groups to one shared coefficient; Joint RCRE includes that family and can allocate unequal strengths when their responses require it. In the present examples, Scalar and Joint have comparable decoded tracking despite their different coefficient constraints. Native-field error measures the allocation directly, while decoded tracking measures command execution.

Additional response objectives. Protected response adds $\eta_P\|PAb\|^2$ outside a radius-32-pixel reference tube, with $\eta_P = .5\,\mathrm{tr}(C)/\,\mathrm{tr}(A^\top P^\top PA)$. Temporal minimax minimizes the maximum normalized output-time-group error, plus .01 of normalized total error and the common ridge term. Transition response adds $\eta_\Delta\|\Delta_t(Ab - r)\|^2$, with $\eta_\Delta = .5\,\mathrm{tr}(C)/\,\mathrm{tr}(A^\top\Delta_t^\top\Delta_t A)$. It penalizes transitions of the fitting error, preserving the desired target changes. Every variant shares actual-native-RMS acceptance.

The later seed-945 screens retain each objective, the coefficient box, budget, probe scale and call count. Their selection endpoint is command accuracy, specified in the next section. The protected soft penalty is an empirical variant of the preservation idea; the hard certificate is given separately in Appendix E.

S Edit Contract Audit and error attribution

ECA evaluates requested position and event timing from the decoded output. Let p_f^0, p_f, q_f denote unedited, edited and commanded positions. On valid frames with $\|s_f\| \geq \epsilon$, define $s_f = q_f - p_f^0$, $d_f = p_f - p_f^0$, $E = \sum_f \|s_f\|^2$, $w_f = \|s_f\|^2/E$, $a_f = d_f^\top s_f/\|s_f\|^2$ and $v_f = d_f - a_f s_f$. With $\bar{a} = \sum_f w_f a_f$, $\gamma^2 = \sum_f w_f(a_f - \bar{a})^2$, $\beta^2 = \sum_f \|v_f\|^2/E$ and $e^2 = \sum_f \|d_f - s_f\|^2/E$,

$$e^2 = (\bar{a} - 1)^2 + \gamma^2 + \beta^2. \tag{19}$$

Orthogonality $v_f^\top s_f = 0$ and weighted variance prove the identity. It separates mean under/over-editing, inconsistent gain and lateral error. Two hundred random numerical checks have maximum error 1.33×10^{-15}; 200 rigid-coordinate/scale checks have maximum error 5.55×10^{-16}.

Table 18: Required clauses in the complex motion studies. All must pass.

Clause	Frozen threshold or rule
Position	Maximum centroid error ≤ 4 px.
Direction	Maximum angle $\leq 10^\circ$ on declared moving windows; zero executed motion fails.
Travel	Maximum phase-relative path-length error $\leq 10\%$.
Events	Correct event count; each event within one frame (62.5 ms).
Hold/loop	Hold excursion and loop closure ≤ 4 px; specified loop orientation.
Observation	Exactly one valid tracked red component in each required frame.

The HSV extractor accepts area 32 to .3 of the frame. Missing or ambiguous observations are marked unverifiable and retained in the denominator. Nominal thresholds $(12, 90, 60)$ and alternatives $(8, 80, 50)$ and $(10, 120, 70)$ are fixed before scoring. A pass count describes complete videos,

not correlated frames. Size and video-quality scores are excluded from these position/time commands.

The complex evaluator passed 12 analytic fixtures before generation. Paired attribution additionally needs a valid no-write track; absolute command clauses can evaluate a valid edited output independently.

T UNEQUAL REVERSALS AND LOOP–HOLD–RETRACE COMMANDS

Seed 945 and the original left-to-right ball prompt are fixed across methods. The reversal command linearly interpolates positions $(72, 150), (240, 126), (104, 148), (224, 130), (80, 146)$ at frames $0, 10, 24, 34, 48$. The three turns occur at unequal intervals. The loop command is $q = (160-80\cos\theta, 140+18\sin\theta)$: $\theta = 2\pi f/16$ on frames 0–16, 2π on 16–28, then $2\pi-2\pi(f-28)/20$ on 28–48. It requests a forward loop, a .75 s pause and a slower reverse loop. All references, budgets and settings are shared.

Table 19: All stress-test outputs under the frozen position/time contract. Position and angle are worst-case errors; travel is the worst phase-relative path-length error. Event error is in frames (one frame is 62.5 ms). P/F/U denote complete-command pass/fail/unverifiable. Size and video-quality scores are excluded.

Method	Position px	Angle °	Travel %	Event fr.	Visible	Status
Three reversals with vertical displacement						
No write	208.15	179.50	80.0	–	49/49	F
Scalar fit	1.09	1.49	0.5	0.0	49/49	P
Joint RCRE	1.09	1.49	0.5	0.0	49/49	P
Protected response	2.33	1.66	0.9	0.0	49/49	P
Temporal minimax	1.07	1.46	0.5	0.0	49/49	P
Transition response	1.12	1.46	0.5	0.0	49/49	P
Loop, hold and slower reverse traversal						
No write	207.74	178.42	81.8	–	49/49	F
Scalar fit	1.44	1.88	0.3	0.0	49/49	P
Joint RCRE	1.45	1.88	0.3	0.0	49/49	P
Protected response	2.54	2.07	0.7	0.0	49/49	P
Temporal minimax	1.52	1.88	0.3	0.0	49/49	P
Transition response	1.48	1.77	0.3	0.0	49/49	P

All five editors satisfy both complete commands; no-write controls fail. All edited event errors are zero at frame resolution. The lowest nominal position error differs by task: minimax for reversals, scalar for the loop. Subpixel differences are evaluated alongside fixed extractor sensitivity. These image-plane paths evaluate spatial/temporal compliance in the controlled scene. Their forward/reverse orientation and unequal timing preclude a simple palindromic command that would mask reversal errors.

U TWO NEW PROMPT CONDITIONS

Table 3 in the main paper compares four representative editors. Table 20 below retains all six, including temporal-minimax and transition-response variants, with the same targets, inputs, tracking rules and error definitions. The additional rows resolve the direction/travel trade-offs behind the main comparison.

Prompt A requests "A small bright red ball moves smoothly from right to left across a light wooden tabletop. Fixed camera, gray wall, soft lighting, one ball, no people." Prompt B replaces the motion phrase with "moves away from the camera ... becoming smaller as it moves into the distance." The same dense commands from Appendix T override these motion descriptions. Prompt changes are the only change from the corresponding preceding tasks; methods within each case share seed 945 and every non-editor setting.

Table 20: All new prompt-conditioned outputs under the frozen position/time contract. Scalar fit is the prior editing baseline; the remaining nonzero-write methods are candidate designs. Position and angle are worst-case errors; travel is the worst phase-relative path-length error. Event error is in frames (one frame is 62.5 ms). P/F/U denote complete-command pass/fail/unverifiable. Size and video-quality scores are excluded.

Method	Position px	Angle °	Travel %	Event fr.	Visible	Status
Prompt A: right-to-left; commanded repeated reversals						
No write	147.06	178.85	93.9	–	49/49	F
Scalar fit	1.14	1.43	0.5	0.0	49/49	P
Joint RCRE	1.10	1.43	0.5	0.0	49/49	P
Protected response	2.28	1.53	1.1	0.0	49/49	P
Temporal minimax	1.16	1.40	0.5	0.0	49/49	P
Transition response	1.13	1.52	0.5	0.0	49/49	P
Prompt B: away from camera; commanded loop/hold/retrace						
No write	–	–	–	–	48/49	U
Scalar fit	1.43	1.73	0.3	0.0	49/49	P
Joint RCRE	1.44	1.73	0.3	0.0	49/49	P
Protected response	4.05	3.09	0.9	0.0	49/49	F
Temporal minimax	1.46	1.78	0.3	0.0	49/49	P
Transition response	1.46	1.84	0.3	0.0	49/49	P

Joint RCRE has mean maximum-position error 1.273 px; Scalar fit has 1.283, Transition 1.295, Minimax 1.309 and Protected response 3.162. Four methods pass 2/2 commands; Protected response passes 1/2. Its Prompt B maximum position error is 4.048 px. All edited event boundaries are exact. The no-write Prompt B case is unverifiable on one of 49 frames.

The selection rule first ranks complete passes, then mean maximum-position error, selecting Joint RCRE for these two displayed prompt conditions. Scalar and Joint both attain the full commands. Their .010 px mean gap is below the predefined .75 px decoded-improvement margin, so these examples establish command execution and compatibility with the scalar baseline. The equal-budget native-field ablation in Appendix R separately measures the effect of retaining response cross terms.

V Complete decoded outputs for repeated reversals

The first row below is the supplied reference. All following rows are native Wan decoded outputs under Prompt A, with the same reference, initial noise, sample count and write budget. Every method, including the no-write control, is shown in fixed order. The selected display frames include the turns.

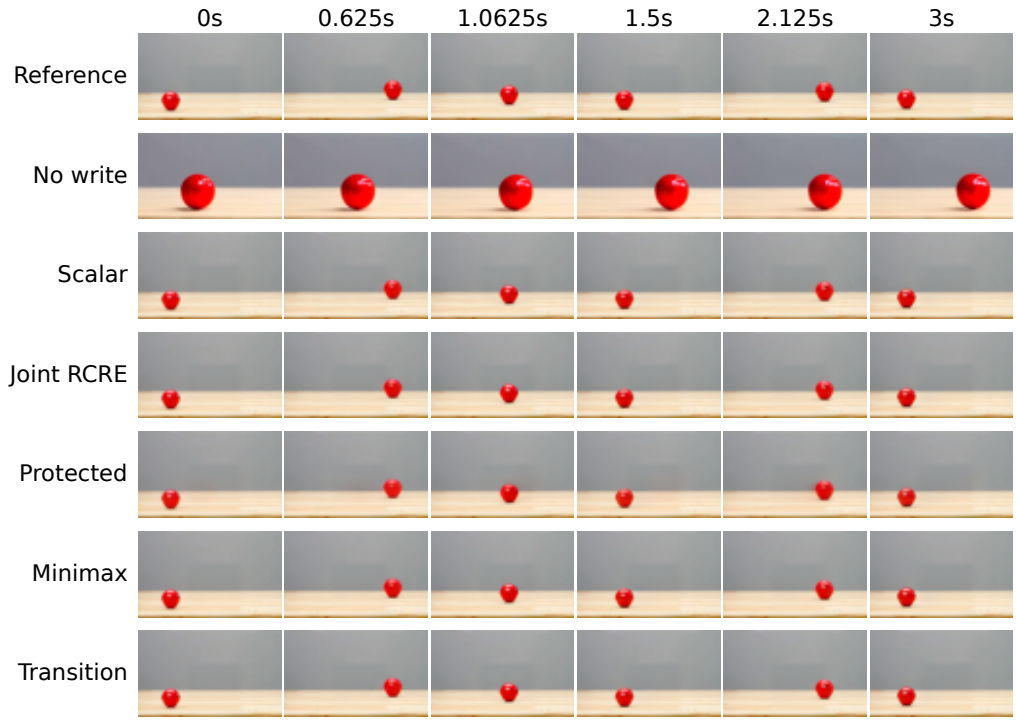

Figure 10: Repeated-reversal outputs for every editor under the right-to-left prompt. Column times use the original video clock.

The full 49-frame arrays determine the scores; the displayed frames explain the movement and preserve the corresponding output appearance. Cyan overlays in the separately delivered comparison video indicate the commanded position for viewing only. They do not enter the raw outputs or evaluator.

The near-coincident controlled trajectories illustrate why complete command success and field-level fitting loss answer distinct questions. Once a dense target is attained, small coefficient changes can produce very similar decoded motion. The no-write row establishes the effect of applying the reference-conditioned residual update under this prompt.

W Complete decoded outputs for loop and retrace

Prompt B requests receding motion, while the reference requests a forward ellipse, stationary interval and slower reversed ellipse. All controller settings match Prompt A; the target and language condition are declared for this separate case. Both the pause and the reversed traversal are scored.

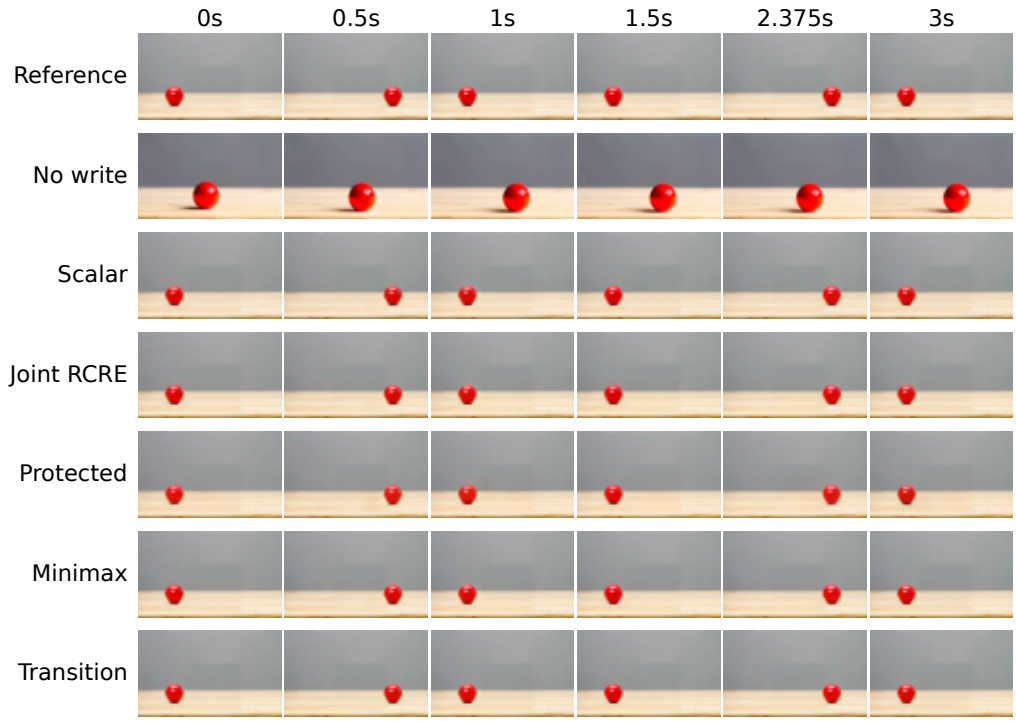

Figure 11: Loop/hold/retrace outputs under the receding-motion prompt. Reference input and all six native output rows follow the same ordering as the preceding figure.

Motion size is unrestricted by the command. The position/time evaluator therefore uses centroid movement, phase travel, hold excursion and loop orientation. Protected response has accurate event boundaries but exceeds the position tolerance; ECA retains this failure because one good clause cannot compensate for another failed clause.

The two cases test competing motion descriptions in the same scene. They provide controlled evidence about residual fitting under dense references. The recorded comparison covers this declared target and prompt set.

X FRAME-ALIGNED TRAJECTORY DIAGNOSTICS

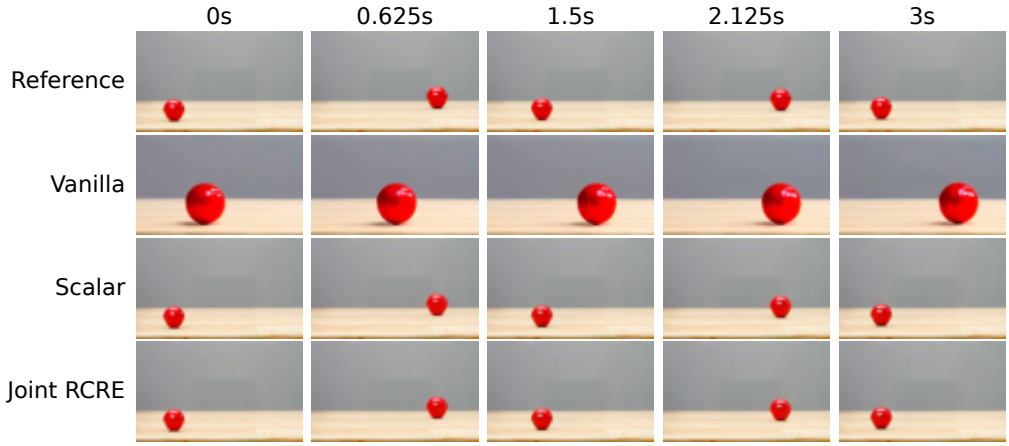

Figure 12: **Native video editing: repeated reversals.** The top row is the supplied command reference. Vanilla Wan (zero write), Scalar fit and Joint RCRE use the same prompt, initial noise and sampler. Saved RGB frames show reversal times and endpoints; Table 3 scores all 49 frames.

The commanded and measured paths share the original frame clock. The error plots retain abrupt events and all frames; their symmetric-log ordinate has a 1 px linear region. The dashed line marks the 4 px position tolerance.

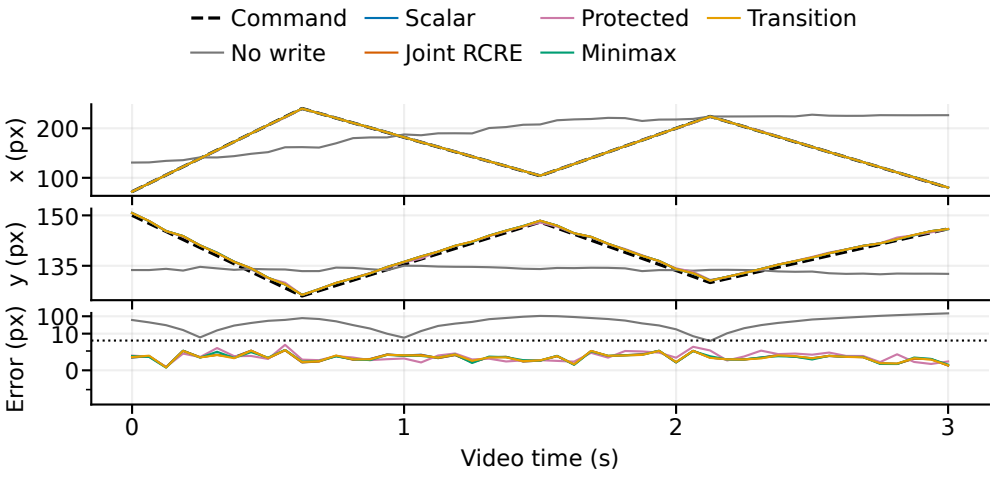

Figure 13: Prompt A: horizontal and vertical positions and command error for every method.

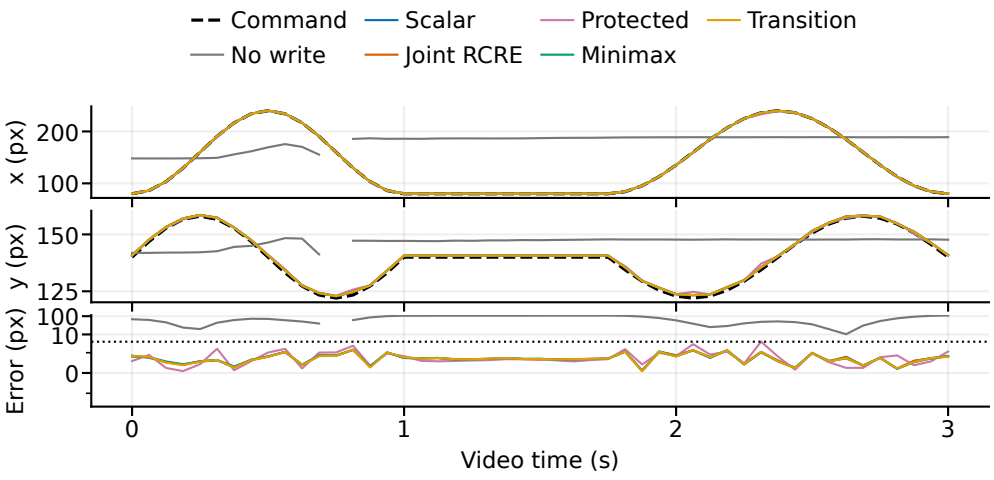

Figure 14: Prompt B: the same measurements for forward loop, hold and slower retrace.

Y NATURAL-SCENE MOTION EDITING: VEHICLES AND ANIMALS

From temporal commands to natural motion. RCRE applies the same signed-response editing rule to a rigid vehicle and an articulated animal. The vehicle command reverses the source sequence, changing image-plane travel and the order of approach and recession. The animal command holds the initial state for nine output frames, accelerates traversal of the observed motion, and retraces the latter half. These operations make direction, pace, pauses and temporal ordering explicit control variables. Dense video references specify each requested sequence; ST-Lens refreshes four temporal response columns at the current residual state, and RCRE fits their joint effect before native-suffix verification.

Paired generation and editing. For each object, we generate an original-order output and edited outputs with identical prompt, initial noise and sampler. Only the dense target changes between original-order and edited Joint runs. All displayed outputs are actual Wan sampler/VAE decodes: accepted block-28 residual writes pass through the model's remaining blocks and output head. The target-input row is shown separately. This connects a specified motion change to a measurable native response and then to a visible generated sequence.

Table 21: **Natural-scene command execution, seed 945.** All errors use all 49 output frames. Position is source-patch registration relative to the edited target; phase is whole-scene nearest-source-frame error. Original denotes the original-order Joint output scored against the edited target.

Object	Method	Cap (%)	RMSE (px)	Max (px)	Phase MAE
Vehicle	Original	2.5	85.89	135.52	4.61
	Scalar	2.5	0.35	1.41	0.00
	Joint RCRE	2.5	0.32	1.41	0.00
Animal	Original	2.5	35.60	75.96	3.33
	Scalar	2.5	20.24	91.44	0.57
	Joint RCRE	2.5	26.32	108.17	0.43
Animal	Original	10	15.41	72.99	3.24
	Scalar	10	0.00	0.00	0.00
	Joint RCRE	10	0.00	0.00	0.00

Target attainment and write budget. Vehicle reversal reduces Joint registration RMSE from 85.89 to 0.32 pixels. For the animal, increasing the common residual RMS cap from 2.5% to 10%

enables both editors to attain the requested trajectory and phase under the fixed evaluator. The 10% cap is applied equally to original-order, Scalar and Joint runs; all other settings and initial noise are preserved. The low-budget horse outputs contain a large ghost image; the higher-budget outputs resolve the intended sequence. This follow-up, chosen after the low-budget runs, illustrates how available write magnitude affects command attainment. Scalar and Joint reach the same integer-pixel registration score in the higher-budget animal case.

Inputs, metrics and reproducibility. Sources are MSR-VTT video189 (vehicle, frames 18–27) and video720 (horse, frames 4–24), stored at 3 fps, letterboxed and resampled to 49 frames. Source-only boxes and a fixed global template matcher measure registration; the pass rule is RMSE ≤ 8 px, maximum ≤ 16 px and correlation ≥ 0.6 on at least 90% of frames. Phase measures scene timing, including camera/background changes. Repeated source frames share the same phase. These metrics quantify retiming along the supplied scene path; appearance is visible in the accompanying native frame grids. The shared settings are Wan2.1-T2V-1.3B, 320×192, 49 frames, 16 fps, 25 UniPC steps, CFG 6, shift 8, block 28, active steps 3–24 and probe scale .5. Each editor uses 51 full forwards and 132 suffix calls, with equal references and noise.

Z VEHICLE REVERSAL: NATIVE GENERATED FRAMES



Figure 15: **Reversing a vehicle's observed motion.** Columns show output frames 0, 12, 24, 36 and 48. The original-order output progresses along the source sequence; the edited target reverses that order. Scalar and Joint RCRE reproduce the reversal under the same 2.5% residual RMS cap. Target input and native generated outputs are labeled separately.

The reversed command changes both position and the visible progression of vehicle scale. Joint's all-frame position RMSE is 0.32 px, with maximum error 1.41 px and zero scene-phase error. Scalar reaches 0.35 px with the same maximum and phase scores. The close results match the formal relation between the editors: Scalar retains response coupling while restricting the temporal coefficient vector to one shared value. The original-order control uses the same prompt and noise, demonstrating that the changed motion follows the edited reference supplied to the residual controller.

AA ANIMAL MOTION: HOLD, TRAVERSE AND RETRACE

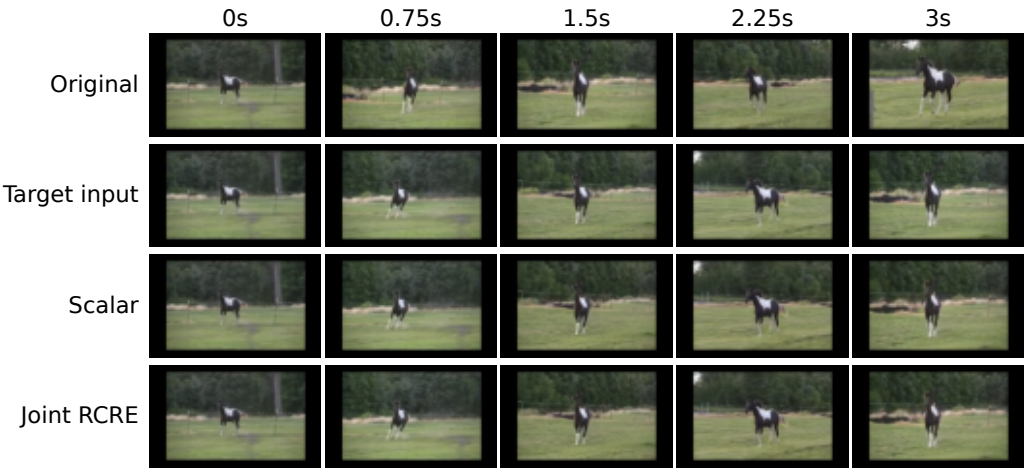

Figure 16: **Retiming articulated motion.** At a shared 10% residual RMS cap, the animal remains in its initial state, traverses the source motion faster and retraces the later segment. Rows distinguish original-order generation, target input, Scalar output and Joint RCRE output. Every output panel is an untouched native decode; labels and frame selection are for presentation.

The animal example extends the same temporal response representation from a rigid vehicle to articulated motion. Scalar and Joint both achieve zero integer-pixel patch-registration and scene-phase errors across 49 frames. Median patch correlations are 0.9780 and 0.9781, respectively; zero registration error therefore denotes aligned patches, not identical RGB images. The original-order control has 15.41 px RMSE against the edited target, making the command-induced timing change visible and measurable. The budget comparison in Table 21 records the complete progression from incomplete low-budget editing to aligned higher-budget outputs.

AB MATCHED RESPONSE ABLATION ON NATURAL MOTION

Design. An additional seed (946) tests the same vehicle and animal commands with three coefficient solvers: Diagonal removes off-diagonal response Gram terms; Scalar uses the full Gram matrix with tied coefficients; Joint uses the full Gram matrix with independently fitted temporal coefficients. The vehicle cap is 2.5% and the animal cap is 10%, fixed from the preceding runs before inspecting these six outputs. Within each object, prompt, initial noise, reference, four probes, 51 full forwards and 132 suffix calls are identical. Thus the comparison separates response coupling from coefficient freedom with equal model computation.

Table 22: **Additional-seed matched ablation.** Native RMS averages the post-acceptance target error over 22 active denoising steps. Position and phase use the fixed all-frame evaluator of Appendix Y.

Object	Solver	Native RMS	Pos. RMSE	Max (px)	Phase MAE
Vehicle	Diagonal	0.119661	0.535	1.00	0.00
	Scalar	0.119195	0.535	1.41	0.00
	Joint RCRE	0.119372	0.553	1.41	0.00
Animal	Diagonal	0.033088	0.000	0.00	0.00
	Scalar	0.032736	0.000	0.00	0.00
	Joint RCRE	0.032737	0.000	0.00	0.00

Response coupling and decoded attainment. All six outputs satisfy the position tolerance and attain zero scene-phase error. Joint reduces mean native target RMS relative to Diagonal by 0.24% for the vehicle and 1.06% for the animal. Scalar is slightly lower than Joint on native RMS in both cases; its full-Gram, tied-coefficient fit is effective for these commands. Along Joint's trajectories, the mean off-diagonal Gram fraction $\|C-\mathrm{diag}(C)\|_F/\|C\|_F$ is 6.91% and 2.98%, respectively. The cross terms are measurable, while their effect on these decoded registration scores is small. Native-response fidelity and decoded command attainment therefore provide complementary measurements of the editor.

Vehicle reversal, Joint RCRE, seed 946



Animal hold–traverse–retrace, Joint RCRE, seed 946



Figure 17: **Additional-seed native outputs.** Each row shows frames 0, 12, 24, 36 and 48. The same references and commands are retained from seed 945. Vehicle outputs retain visible texture blur and ghosting alongside accurate position and timing; the animal sequence reproduces the requested retiming at the 10% cap.

Run audit and evaluation unit. All runs pass residual-budget and suffix/full-forward equivalence checks (RMS discrepancy below 10^{-5}). Input-noise and reference hashes match across methods within each object. The six runs took 5.75 minutes including model setup and output writing; summed sampling time was 4.95 minutes on RTX 4070. They add one seed on two source clips; frames and denoising steps remain within-clip measurements. Together with seed 945, they show repeated command attainment for both natural-motion examples under fixed protocols.

AC IMAGE GENERATION POLICY AND PAIRED SAMPLES

The DiT generation-quality comparison uses 50,000 generated images per arm with DDIM-50, matched class labels and initial noise, and the ft-ema VAE. FID/KID use Inception features with all 50,000 ImageNet validation images as references. KID is the unbiased polynomial-kernel MMD estimate, displayed as $1000\times$ its value; its full-evaluation subset schedule is not specified by the available protocol record. The baseline FID of 59.9978 is specific to this checkpoint, sampler and preprocessing setup. The manually set guidance schedule is 1/4/6 at thresholds 500/300. Table 4 reports FID/KID point estimates. A separate 1,000-image paired run, seed 0, measures sampling-and-decoding time on RTX 4070. The eight pairs below illustrate output changes; they are visual examples rather than the quality-evaluation set.

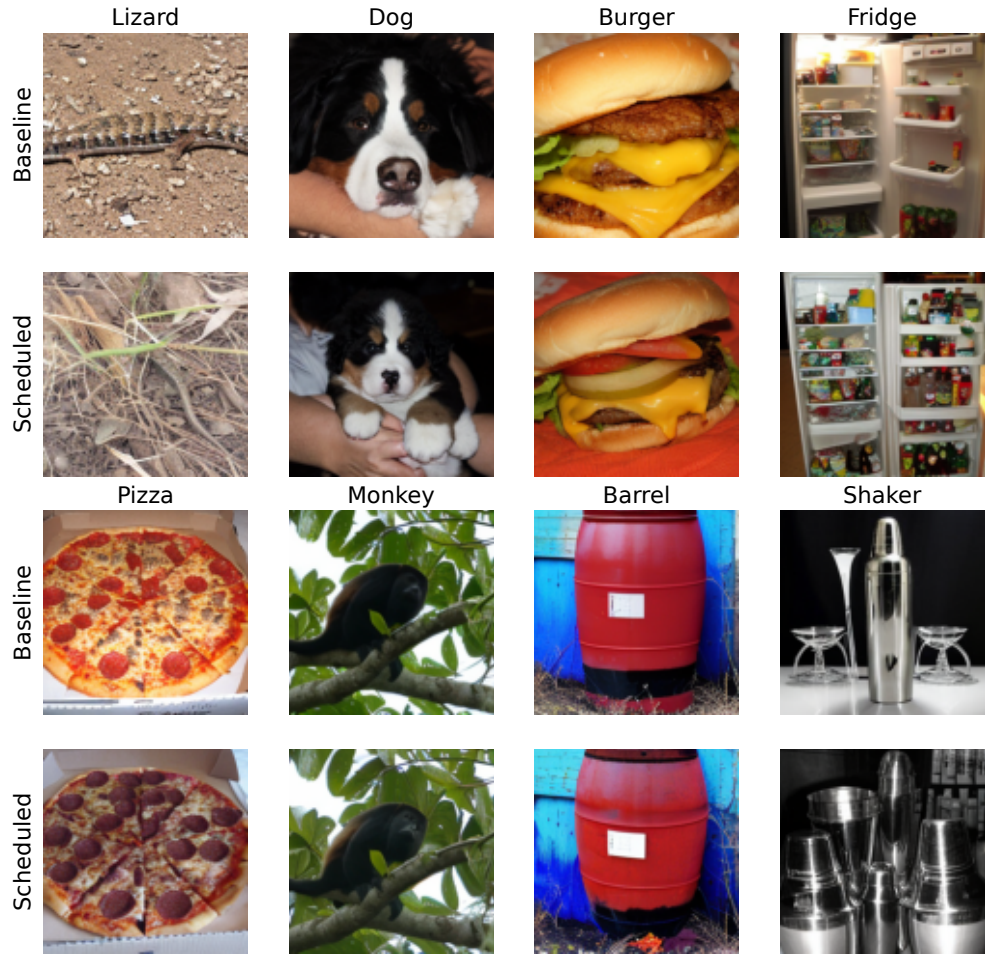

Figure 18: The first eight saved inputs, without quality-based selection. In each group constant-CFG output precedes scheduled-CFG output. Images retain their saved content.

The schedule changes composition as well as detail: the dog includes more of its body, and the refrigerator exposes different shelves. Other pairs change tone or object prominence. The distribution-level FID/KID result in the main paper accompanies this paired gallery. Sampling and decoding take 1241.66 s for the baseline and 946.50 s for the schedule on RTX 4070.

AD Condition diagnostics and computation allocation

Native-field guidance as an output response. At a fixed state, CFG changes the native epsilon prediction by $\Delta\epsilon_t = (w_t - 1)(\epsilon_c - \epsilon_\varnothing)$, and a change Δw_t gives the exact field response $\Delta w_t(\epsilon_c - \epsilon_\varnothing)$. This is an output-stage action in the same native field used by the residual-response objective. Measuring its regime dependence motivates where to allocate guidance computation; the thresholds are manually selected and final-sample quality is evaluated by the complete sampler rollout. This output-stage diagnostic is distinct from ST-Lens's residual-suffix Jacobian and does not establish that the same schedule would fail without ST-Lens.

Condition-swap diagnostics measure the change between class and null branches. In baseline replay, the mean relative condition effect is .188% at $t = 999$, 8.20% at $t = 143$ and 16.92% at $t = 0$. CFG adds $(w - 1)(\epsilon_c - \epsilon_\varnothing)$, so the recorded schedule removes extra extrapolation at high noise and increases it at low noise.

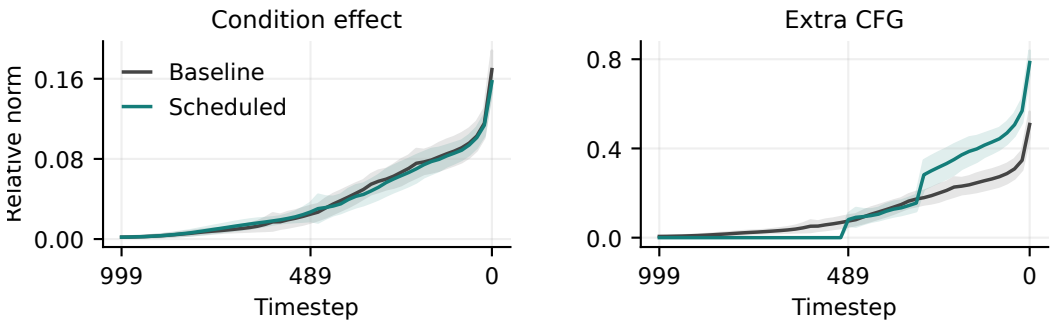

Figure 19: Class-condition effect $\|\epsilon_c - \epsilon_\varnothing\|/\|\epsilon_c\|$ (left) and extra CFG magnitude $|w-1|\|\epsilon_c - \epsilon_\varnothing\|/\|\epsilon_c\|$ (right), on eight saved inputs. Bands show one standard deviation; both use the first three epsilon channels guided by the released DiT implementation.

Table 23: Block-bypass outcomes. FID uses 5,000 matched-seed generated images; classification accuracy uses a separate matched 500-image set. These outcomes concern block bypassing, separately from the joint-policy timings.

Policy	FID ↓	Top-1 (%)	Top-5 (%)
Full model	17.26	83.4	94.6
One-block bypass	17.22	82.8	93.8
Two-block bypass	17.41	80.4	93.0
Four-block bypass	19.24	75.8	90.8

The bypass study is a separate execution policy. A four-block native-field audit gives errors .1462/.0202/.0056 at $t = 100/500/900$; the matched random subset gives .2073 at the middle setting. Calibration therefore identifies the operating regimes where removal has smaller output effect. The table reports the resulting 5,000-image FID and separate 500-image class accuracy. Guidance timing and bypass quality are kept as distinct experiments.

AE From response observation to editing and policy decisions

Proposition 1: a common decision representation. Fix a state, proposal dictionary D, output response A and target r. Regional weights can be absorbed into A and r by stacking weighted coordinate projections. Let a coefficient distribution have finite second moments, mean μ and covariance Σ. With $C = A^\top A$, $g = A^\top r$ and $H = C + \lambda I$, $\lambda \geq 0$,

$$\mathcal{L}_A(\mu, \Sigma) = \mathbb{E}[\|Ab - r\|^2 + \lambda\|b\|^2] \tag{20}$$

$$= \mu^\top H\mu - 2g^\top\mu + \|r\|^2 + \mathrm{tr}(H\Sigma). \tag{21}$$

The expected source-write energy is

$$\mathbb{E}\|Db\|^2 = \mu^\top D^\top D\mu + \mathrm{tr}(D^\top D\Sigma). \tag{22}$$

Proof. Write $b = \mu + \xi$, where $\mathbb{E}\xi = 0$ and $\mathbb{E}\xi\xi^\top = \Sigma$. The mean cross term vanishes and $\mathbb{E}\xi^\top H\xi = \mathrm{tr}(H\Sigma)$. Apply the same expansion to $D^\top D$. These are standard quadratic moment identities; their role here is to expose the statistics required by both decisions.

The three contributions as a sequence. *Observation* estimates the signed coordinate response A, hence the task-specific C and g. *RCRE* chooses a deterministic distribution, $\mu = b$ and $\Sigma = 0$, under the actual box and source-write constraints. *Stochastic policy selection* chooses Σ, or jointly chooses μ and Σ, using the same quadratic interactions. Equal marginal variances can produce different output energies because $C_{ij}\Sigma_{ji}$ retains signed cross-time coupling. This supplies a mathematical connection between the three contributions without equating their intervention sites.

Why marginal energy cannot determine the policy. Take $A_+ = [1, 1]$ and $A_- = [1, -1]$. Their column energies are identical. For $v_+ = (1, 1)^\top/\sqrt{2}$, $v_- = (1, -1)^\top/\sqrt{2}$ and $\Sigma_\pm = v_\pm v_\pm^\top$, both policies have diagonal $(1/2, 1/2)$ and equal source energy when $D^\top D = I$. Nevertheless,

$$\mathrm{tr}(A_+^\top A_+ \Sigma_+) = 2, \quad \mathrm{tr}(A_+^\top A_+ \Sigma_-) = 0,$$

and the ranking reverses for A_-. Thus marginal-energy readouts cannot identify the minimum-energy covariance for both operators. The obstruction is loss of signed cross terms, not the absence of a particular model name. Appendix D separately proves which regional queries are lost by output-coordinate pooling.

Matching the response to the action. For fixed-state CFG, changing w has the exact native-field column $a_t = \epsilon_c - \epsilon_\varnothing$, so guidance allocation is an output-stage instance of the response interface. For initial noise z and final decoded output y, the relevant derivative is $J_{y,z} D_z$. If $y = \Psi_t(s_t(z))$, with s_t containing the complete sampler state (latent, required solver history and all continuation inputs), then $J_{y,z} = J_{\Psi_t, s_t} J_{s_t, z}$. A single residual suffix omits other state paths. ST-Lens's signed cross-time responses supply a candidate-design criterion, while complete rollouts rank initial-noise settings. The six-state residual audit tests local covariance selection, not direct FVD prediction. FID, FVD and VBench remain measured endpoints, not consequences of a native-field quadratic identity. The local propositions here give no end-to-end guarantee through classifier-free guidance, multiple denoising steps, nonlinear suffix operations or VAE decoding.

AF Decision stability and when independent coefficients help

Proposition 2: observation error controls local decision regret. Consider any common family of coefficient distributions supported on $\|b\| \le R$. Let the true finite response be $f(b)$ and suppose $\|f(b) - \widehat{A}b\| \le d$ uniformly on that ball. Set $M = \|\widehat{A}\|R + \|r\|$ and $\eta = 2Md + d^2$. For the same regularizer in both objectives,

$$|\mathcal{L}_f(\pi) - \mathcal{L}_{\widehat{A}}(\pi)| \le \eta. \tag{23}$$

If $\widehat{\pi}$ minimizes the estimated objective to tolerance τ and π_* minimizes the true one over that family, then

$$\mathcal{L}_f(\widehat{\pi}) - \mathcal{L}_f(\pi_*) \le 2\eta + \tau. \tag{24}$$

Proof. For $e = f(b) - \widehat{A}b$, expand the squared residual: $2\langle \widehat{A}b - r, e\rangle + \|e\|^2$. Its absolute value is at most $2Md + d^2$, pointwise and hence in expectation. Insert the estimated objective between the two true objectives and use estimated optimality. The two approximation terms contribute 2η and the solver adds τ.

For a derivative error $\|J_h D - \widehat{A}\| \le \varepsilon$ and L-Lipschitz Jacobian on all feasible write segments,

$$d = \varepsilon R + \tfrac{1}{2} L \|D\|^2 R^2$$

follows from the finite-response remainder. Thus improving ST-Lens fidelity directly tightens a bound on both deterministic editing and bounded random policy selection. An estimated improvement greater than 2η certifies a true local improvement when these bounds hold. This is a local bound; unbounded Gaussian policies need moment-based remainder control instead.

Proposition 3: the available gain over Scalar. Let $H = A^\top A + \lambda I \succ 0$, $g = A^\top r$ and $q(b) = b^\top H b - 2g^\top b + \|r\|^2$. Let $b_s = c\mathbf{1}$ be the optimal feasible scalar fit and b_* the optimal Joint fit under the same constraints. Then

$$0 \le q(b_s) - q(b_*) \le (g - Hb_s)^\top H^{-1}(g - Hb_s) = U_s. \tag{25}$$

Proof. Scalar feasibility implies the first inequality. Completing the square gives $q(b) = q(H^{-1}g) + (b - H^{-1}g)^\top H (b - H^{-1}g)$. The unconstrained optimum lower-bounds the constrained one, giving U_s. Equality holds if $H^{-1}g$ is feasible. For an interior scalar optimum, $c = (\mathbf{1}^\top g)/(\mathbf{1}^\top H \mathbf{1})$.

A small U_s bounds the possible local benefit of independent coefficients. Larger values motivate the Joint solve; its realized predicted gain can be compared with the error bound above before native verification. Scalar retains the full Gram matrix, and the observed 0.010 px two-case decoded

difference falls below the declared 0.75 px improvement margin. The guarantee concerns the same-state local objective, whereas separately rolled-out editors visit different states. It neither orders their final images nor imposes per-step superiority after nonlinear acceptance. The existing 5.46% and natural-motion comparisons report observed rollout effects with those distinctions intact.

AG READOUT DISTINCTIONS AND A MATCHED POLICY-SELECTION AUDIT

What the interface adds to prior lenses. J-Lens estimates context-averaged effects on vocabulary likelihood, including future verbalization (Gurnee et al., 2026); ST-Lens shares the mean-derivative principle. DSL steers image-encoder representations and patches downstream submodule outputs to expose direct contributions through a diffusion decoder (Takatsuki et al., 2025). The query here instead retains the signed native epsilon/velocity at every requested space–time coordinate after the unpatched suffix, indexed by block and noise regime. Those columns yield regional C and g for the decision identities above. Original vocabulary readouts or decoded visualizations do not themselves supply these statistics. Adapting a prior lens to estimate the same native response is a valid comparator; the main tables evaluate direct-head and frozen-path adaptations without presenting them as original implementations.

Same-state covariance selection with native verification. We freeze three steps (3, 12, 24) on each of the vehicle and animal trajectories at seed 946. Four disjoint temporal proposals are rescaled so every sign combination has hidden RMS 0.25% of the current residual. For the eight sign vectors $s = (1, \pm 1, \pm 1, \pm 1)$, policy π_s applies s or $-s$ with equal probability. All policies have mean zero, covariance $ss^\top$, identical marginal variances and equal write energy.

Four individual positive probes provide the columns of A. The full-response selector minimizes $s^\top A^\top A s$ before any candidate outcome is used. The diagonal selector predicts the same $\mathrm{tr}(A^\top A)$ for all candidates. Its uniform tie-break has the exact expected energy of random selection among the eight policies. We evaluate every candidate with separate native $+s$ and $-s$ suffix calls: $E_s = (\|F(h + Ds) - F(h)\|^2 + \|F(h - Ds) - F(h)\|^2)/2$. The native oracle minimizes these measured energies.

Table 24: **Response-based local policy selection.** Full-response rank is among eight candidates by measured native energy. Reduction compares selected energy with the exact uniform random/diagonal-tie expectation. Prediction error averages relative energy error across all eight candidates.

Scene / step	Native rank	Reduction (%)	Pred. error (%)	Spearman
Vehicle / 3	1	9.08	0.187	1.00
Vehicle / 12	1	3.47	0.107	1.00
Vehicle / 24	1	25.42	0.036	1.00
Animal / 3	1	8.22	0.074	1.00
Animal / 12	1	1.90	0.065	1.00
Animal / 24	1	7.19	0.010	1.00

The signed response ranks all eight policies correctly at all six states, selecting the measured minimum with zero candidate-set regret. Identical diagonal predictions cannot express these state-dependent rankings. Each selector receives the same four column probes; an additional 16 native calls per state evaluate outcomes for all selectors. These shadow probes leave the parent decoded videos bitwise unchanged. The two runs take 2.16 minutes and each uses 51 full forwards and 192 suffix calls. This directly tests the covariance term of Eq. 21 for finite residual writes. The six states are nested within two source clips; the experiment establishes local selection accuracy on these states, while initial-noise policy quality is measured by the separate full-rollout generation study.

AH REPRODUCIBILITY AND INTERPRETATION OF THE EVIDENCE

Reproducing observation and editing. Use the frozen checkpoints and output-channel conventions in Appendix A. Rebuild calibration on its declared contexts, retain source-dictionary seeds and compare predictions with the same finite interventions. Record the complete native field, not only pooled energy. For shared-control tests, preserve target masks, gain, source budget and the feedback policy of Appendix F. Save successes and failures before aggregating paired context-level results.

The decoded editing runs use Python 3.11, PyTorch 2.7.1/CUDA 12.8, Wan2.1-T2V-1.3B and FP32 model execution with CPU/GPU offload on RTX 4070. Fixed settings are seed 945 for the complex and prompt screens, 49 frames, 320×192, 16 fps, 25 UniPC steps, shift 8 and CFG 6. Block 28, active steps 3–24, four groups, probe .5, coefficient box $[0, 2]^4$, residual cap .025 and two acceptance scales are held fixed. Protocols pin initial-noise, reference RGB/latent and executable hashes. Every run retains native RGB, MP4, latent and per-step measurement records.

Using the supplementary figures. The main paper retains the teaser in the opening introduction and the overview before the method. The depth–noise fidelity map (Figure 3), pooled-null observation, matched-rank image comparison and native editing output carry the central evidence. Supplementary image heatmaps and grids resolve regional and budget dependence; Appendix M details the main fidelity map. Complete editor frame grids and trajectories explain the command audit. The image gallery and condition profiles connect distribution-level generation scores with individual outputs. These views complement the main tables without duplicating their primary conclusions.

Evaluation units and assumptions. The mean-action result is a squared first-order-risk statement under its declared context distribution. Finite-amplitude curvature and context transfer are tested separately. The protected-response theorem assumes local error/curvature bounds; its controlled numerical validation and soft-penalty Wan test have distinct roles. The displayed motion cases provide frame-resolved command audits for their declared inputs and seeds. Generation tables specify their cohorts, metrics and fixed candidates. The observer study reports a paired fidelity interval; generation results report three-seed Wan means and ImageNet point estimates. The counts identify the numeric records, and the frame grids illustrate the resulting behavior.

Responsible use. Experiments use pretrained models, existing image/video data and constructed targets. No human participants were recruited. Data and model licenses govern reuse. Generated edits should be identified as synthetic when provenance matters. Model behavior can inherit biases from pretraining, and controllable generation can be misused to create misleading content.