STRUCTURA: INTRINSIC SCORING AND INFERENCE-TIME REFINEMENT AGAINST STRUCTURAL HALLUCINATIONS IN VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern video generators produce visually compelling clips, yet objects can develop impossible structures as they move: limbs fuse or multiply, parts detach, and geometry changes inconsistently across frames. Mitigating these spatiotemporal structural hallucinations requires feedback that identifies emerging failures and guides correction during sampling. We introduce Structura, a training-free framework that obtains this feedback from the generator's own reconstruction dynamics. We slightly re-noise each intermediate clean prediction and reconstruct it, using low-noise reconstruction as a local geometric probe. Unreliable regions tend to change more under this probe, and large changes recur at the same locations when a structural failure persists across sampling steps. Structura accumulates the magnitude of these changes to stabilize spatiotemporal localization, while retaining their signed displacement to guide stochastic local resampling. We introduce Structura Bench, a collection of hallucination-prone cases for evaluating video structure. Experiments on Dynamic-Bench and Structura Bench show improved structural and anatomical correctness across video generators. Applications to video object removal and world action models further demonstrate the use of reconstruction feedback for temporal refinement and action completion.

1 INTRODUCTION

Recent video generative models synthesize clips with striking photorealism, smooth motion, and strong prompt adherence (Team Wan et al., 2025; Yang et al., 2025), fueling growing interest in repurposing them as visual simulators and world models (NVIDIA et al., 2025). Yet as world models they remain unreliable: even state-of-the-art models frequently exhibit severe structural failures such as anatomically distorted limbs, morphing object geometries, and physical inconsistencies that emerge over time (Bansal et al., 2024; Zheng et al., 2025). We refer to this class of failures as *spatiotemporal structural hallucinations*: content that matches the statistical properties of the training distribution, yet violates the underlying physical, logical, or morphological laws of the data (Sobieski et al., 2026). Such hallucinations directly limit the reliability of generated videos for simulation, planning, and embodied decision making.

Methods for improving physical and structural fidelity obtain corrective information in several ways. Training-based approaches teach explicit motion or physics priors (Chefer et al., 2025; Zhang et al., 2025). Inference-time approaches use learned rewards or vision–language models (VLMs) to score and guide generations (Liu et al., 2025a; Kim et al., 2025; Cheng et al., 2026), or search over candidate trajectories (Oshima et al., 2025). More recent generator-native methods reuse the model's own predictions to refine uncertain regions or score physical plausibility (Jang et al., 2026; Hassan et al., 2026). Self-Refining Video Sampling performs local refinement within a fixed-noise inner loop, whereas Proprio aggregates perturbation residuals into a plausibility score for a generated latent. Neither explicitly maintains a spatial memory of recurring reconstruction changes as one solver trajectory evolves, nor separates accumulated magnitude for localization from signed displacement for steering.

Structural failures are local and may emerge only after motion begins to unfold. A useful internal signal must therefore identify where the prediction remains unstable, distinguish recurring changes

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.

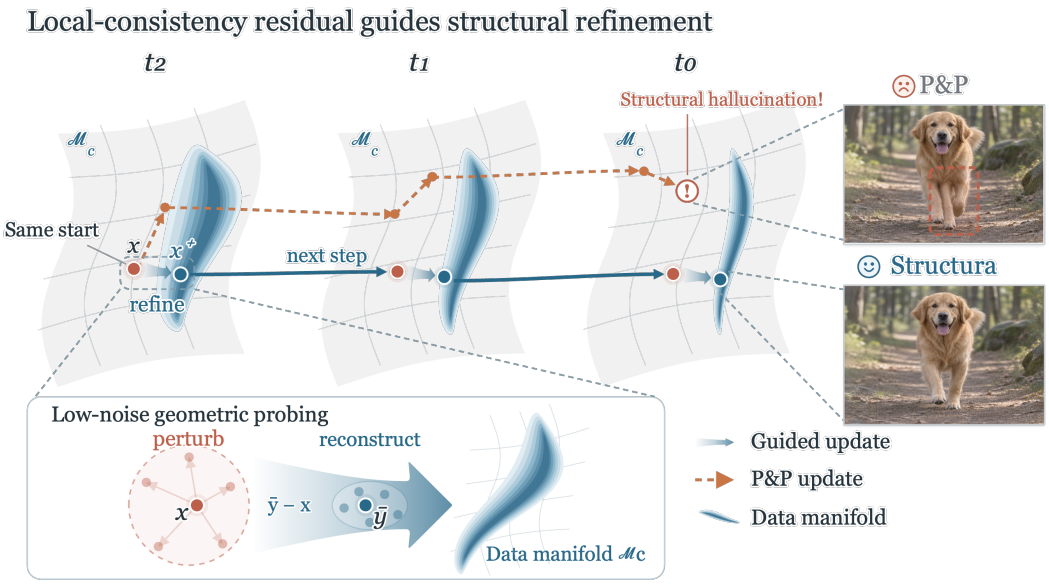


Figure 1: **Local-consistency residual guides structural refinement.** Structura perturbs and reconstructs the current clean prediction x; the mean reconstruction displacement $\bar{y} - x$ supplies a candidate refinement direction. Across sampling steps, accumulated reconstruction feedback localizes persistent disagreement and steers local re-noising, illustrated against predict-and-perturb (P&P) updates from the same start. Trajectories and local views of the data manifold $\mathcal{M}_c$ are schematic. The lower inset illustrates low-noise probing; the right panels compare P&P and Structura outputs.

from transient prediction error, and provide a direction for revising the affected region. This raises our central question:

Can a flow-matching video generator use its internal dynamics to steer an online, self-correcting trajectory against spatiotemporal structural hallucinations?

Our key observation (Fig. 1) is that low-noise reconstruction provides a geometric probe of structural integrity. Slightly re-noising and reconstructing a clean prediction draws it toward locally coherent, data-supported geometry. The resulting mean residual measures displacement from the local data manifold: its magnitude identifies structural inconsistency, while its negative direction guides recovery. Regions with persistent structural hallucinations retain larger residuals across sampling steps, allowing this geometric signal to locate failures and guide their correction as the video trajectory evolves.

We turn this observation into **Structura**, a training-free framework for online localization and closed-loop correction of spatiotemporal structural hallucinations in video flow-matching transformers. At each scheduled sampling step, Structura applies small re-noising probes to the current clean prediction and records the resulting reconstruction changes. It accumulates variance-corrected revision energy across solver steps to suppress transient variation and stabilize a spatiotemporal refinement mask. In parallel, it accumulates signed displacement to guide later re-noising within a predict-and-perturb loop (Jang et al., 2026). Paired reconstructions then determine whether a proposed local update should be committed.

Our main contributions are summarized as follows:

- **Low-noise geometric probing:** We connect the generator's mean reconstruction residual to local manifold displacement and use persistent residual energy to reveal structural hallucinations during sampling.
- **Trajectory feedback accumulation:** We derive a variance-corrected revision-energy estimator and accumulate it alongside signed displacement across sampling steps, providing a spatial refinement mask and a coherent correction direction.

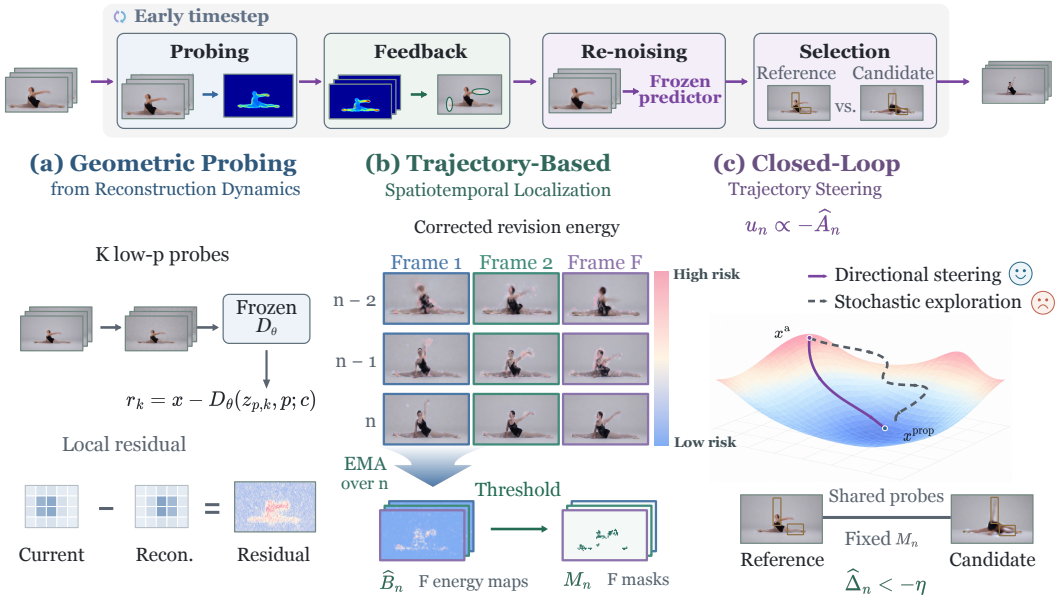


Figure 2: **Overview of Structura.** **(a)** Low-noise probes reveal reconstruction disagreement. **(b)** Accumulated revision energy localizes recurring errors, while signed displacement guides re-noising. **(c)** Directional exploration proposes updates; paired reconstructions select changes within a fixed mask. The anchor advances upon acceptance.

- **Feedback-guided refinement:** We combine localized directional re-noising with paired reconstruction tests to select and commit structural corrections along a frozen generator's sampling trajectory.
- **Structural hallucination evaluation:** We introduce Structura Bench and curate a high-quality dataset of 10,000 video-editing cases involving hands and complex object interactions. Experiments establish structural and anatomical gains across video backbones and extend reconstruction feedback to video editing and world action models.

2 Related Work

Structural hallucinations. Diffusion models can generate malformed structures through interpolation between data modes (Aithal et al., 2024). Prior work detects or corrects these failures using prediction variance, attention, score geometry, and local intrinsic dimension (Oorloff et al., 2025; Triaridis et al., 2025; Tian et al., 2025; Sobieski et al., 2026). Structura uses reconstruction feedback accumulated along the sampling trajectory to localize and revise persistent structural disagreement.

Motion and physical fidelity. VideoPhy and VBench-2.0 evaluate physical commonsense and video fidelity (Bansal et al., 2024; Zheng et al., 2025). VideoJAM and VideoREPA learn motion or physics priors (Chefer et al., 2025; Zhang et al., 2025); Structura operates on a frozen generator.

Inference-time refinement. Learned rewards, VLM guidance, and candidate search improve generation at inference time (Liu et al., 2025a; Kim et al., 2025; Cheng et al., 2026; Oshima et al., 2025). Generator-native methods refine temporal latent statistics (FlowMo), use fixed-noise predict-and-perturb loops (SRVS), or score perturbation residuals (Proprio) (Shaulov et al., 2025; Jang et al., 2026; Hassan et al., 2026). Structura combines trajectory feedback accumulation, directional re-noising, and paired reconstruction selection.

3 METHOD

Structura refines the trajectory of a frozen video flow-matching model through low-noise geometric probing, trajectory feedback accumulation, and feedback-guided re-noising with paired selection (Algorithm 1). Recurring reconstruction disagreement identifies the regions requiring structural revision.

3.1 PRELIMINARIES

For text c, clean latent $X_0 \in \mathbb{R}^{C\times T\times H\times W}$, and noise $\epsilon \sim \mathcal{N}(0, I)$, flow matching uses the linear path (Lipman et al., 2023)

$$Z_s = (1-s)X_0 + s\epsilon, \qquad s \in [0,1], \tag{1}$$

with noise at $s = 1$ and data at $s = 0$. The velocity v_θ regresses $\epsilon - X_0$; sampling integrates from $s = 1$ to 0 with clean endpoint estimate

$$\widehat{x}_{0,s} = D_\theta(z_s, s; c) \equiv z_s - s\, v_\theta(z_s, s; c). \tag{2}$$

This endpoint form lets us query the frozen generator as a low-noise reconstructor at $p \ll 1$ (Jang et al., 2026).

3.2 LOW-NOISE RESIDUALS AS GEOMETRIC PROBING

Our inspiration stems from the question: **What makes a hallucination-free latent along the video sampling process?**

Under the linear flow interpolant, the Bayes-optimal endpoint predictor computes the posterior mean of clean states: $D^*(z, p; c) = \mathbb{E}[X_0 \mid Z_p = z, c]$ (Lipman et al., 2023). For a smooth clean-data manifold $\mathcal{M}_c$ with positive smooth density and a unique, nondegenerate nearest point in the local neighborhood, this posterior contracts as $p \to 0$, reducing the conditional expectation to an orthogonal projection (Liu et al., 2025b; Permenter & Yuan, 2024):

$$D^*(z, p; c) = \Pi_{\mathcal{M}_c}\left(\frac{z}{1-p}\right) + \mathcal{O}(p^2), \qquad p \to 0. \tag{3}$$

Low-noise reconstruction therefore serves as a local manifold projection, actively pulling off-manifold states back into the spatiotemporally coherent, data-supported geometry.

This geometric behavior allows us to construct a *low-noise probe* to test the structural integrity of an intermediate prediction x. Specifically, we query the frozen reconstructor by injecting small, independent Gaussian perturbations at a designated level p:

$$z_{p,k} = (1-p)x + p\,\epsilon_k, \qquad \epsilon_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I), \qquad r_k \equiv x - D_\theta(z_{p,k}, p; c). \tag{4}$$

Taking the expectation over perturbation directions yields the mean response $\mu_p(x) \equiv \mathbb{E}_\epsilon[r(x, \epsilon)]$. By Eq. (3), the ideal counterpart satisfies:

$$\mu_p^*(x) = x - \Pi_{\mathcal{M}_c}(x) + \mathcal{O}(p^2). \tag{5}$$

(See Appendix A for details.) To leading order, $\mu_p^*(x)$ directly measures the displacement vector from the manifold to the current prediction: its magnitude reflects the severity of structural inconsistency, while its negative direction points toward the local manifold recovery.

In practice, each query operates over the full video latent space Ω. For each spatiotemporal token $q = (t_{\text{vid}}, h, w) \in \Omega$, we aggregate the response energy across all C latent channels to quantify localized off-manifold deviation:

$$\mathcal{E}_p(x; q) = \mathbb{E}_\epsilon \|r(x, \epsilon)(q)\|_2^2, \qquad \widehat{\mathcal{E}}_p(x; q) = \frac{1}{K}\sum_{k=1}^{K} \|r_k(q)\|_2^2. \tag{6}$$

For each prompt, six 40-step sampling trajectories generated with different random seeds illustrate this signal (Fig. 3). Regions with severe structural hallucinations maintain larger residual energy throughout sampling, showing that low-noise residuals serve as geometric probes that reveal persistent structural defects.

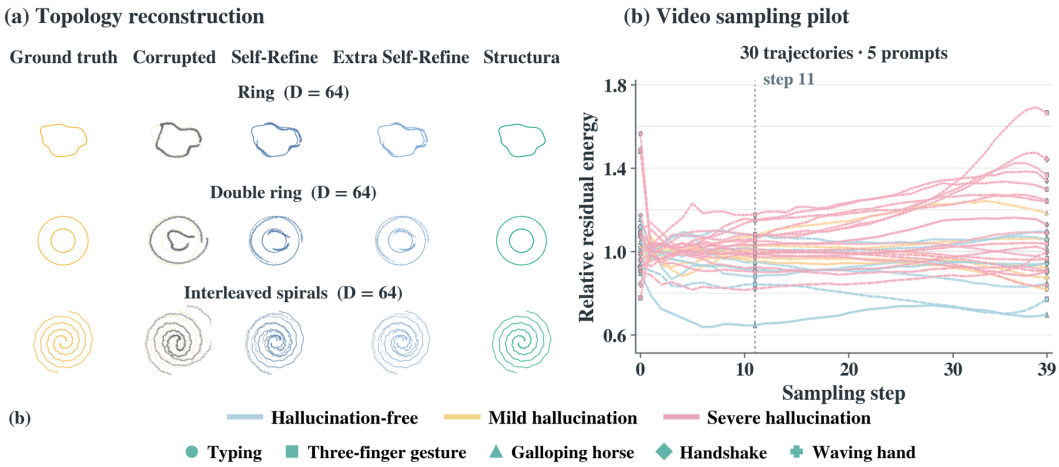


Figure 3: **Low-noise reconstruction as a structural probe.** Controlled topology reconstruction (left) and 40-step trajectories across five prompts and six random seeds per prompt (right). Regions with structural hallucinations retain higher residual energy throughout sampling.

3.3 Trajectory Feedback Accumulation

A region can repeatedly require revision while its correction direction changes. We accumulate revision energy separately from signed displacement to retain this recurring demand for correction.

At refinement round n, let $r_{n,k}(q)$ be the $K \geq 2$ independent probe responses to the current prediction x_n, with mean $\bar{r}_n(q) = K^{-1} \sum_k r_{n,k}(q)$. To measure systematic revision rather than perturbation-specific variation, we use

$$\widehat{b}_n(q) = \|\bar{r}_n(q)\|_2^2 - \frac{\widehat{v}_n(q)}{K}, \tag{7}$$

where $\widehat{v}_n(q)$ is the sample variance trace. This correction gives $\mathbb{E}[\widehat{b}_n(q) \mid x_n] = \|\mu_p(x_n; q)\|_2^2$, connecting $\widehat{b}_n$ to the geometric displacement of the preceding section (Appendix B.1).

For $n \geq 2$, we retain displacement $\widehat{A}_n$ and revision energy $\widehat{B}_n$ through exponential averaging with decay $\beta \in [0, 1)$:

$$\begin{aligned} \widehat{A}_n &= \beta \widehat{A}_{n-1} + (1 - \beta)\bar{r}_n, \\ \widehat{B}_n &= \beta \widehat{B}_{n-1} + (1 - \beta)\widehat{b}_n. \end{aligned} \tag{8}$$

Initialized from the first probe batch, these updates span sampling steps and their inner refinement rounds. The resulting $\widehat{B}_n$ records recurring revision demand, while $-\widehat{A}_n$ supplies its net correction direction. These complementary memories determine where to intervene and how to guide refinement. Figure 6 visualizes how spatial responses concentrate around hands and contact regions across feedback checkpoints.

3.4 Feedback-Guided Refinement (FGR)

We localize intervention with the binary mask $M_n(q) = \mathbf{1}\{[\widehat{B}_n(q)]_+ > \tau\}$ to select locations, where $\tau \geq 0$ and $[a]_+ = \max(a, 0)$. Within predict-and-perturb (P&P) sampling (Jang et al., 2026), which alternates re-noising and endpoint prediction at the current solver level, we bias re-noising toward the normalized correction $u_n(q) = -\widehat{A}_n(q)/\|\widehat{A}_n(q)\|_2$:

$$\zeta(q) = \sqrt{C}\, a_n(q) u_n(q) + \sqrt{1 - a_n(q)^2}\, \xi(q), \qquad \xi(q) \sim \mathcal{N}(0, I_C). \tag{9}$$

The weight $a_n(q) \in [0, 1)$ increases with displacement energy relative to its estimated probe noise, and is zero outside M_n. This converts part of the random exploration into directional steering while keeping $\mathbb{E}_\xi \|\zeta(q)\|_2^2 = C$. The exact strength rule is given in Appendix B.3.

Table 1. **Video generation on Dynamic-Bench and Structura Bench.** Smooth, Dyn., and Flicker are VBench metrics; SA and PC are VideoPhy2 scores. Joint requires both SA and PC ≥ 4. VLM as Judge (gemini-3.8-flash) scores physical, structural, and anatomical correctness from 1 to 5; Overall is their mean. Darker blue marks higher rank; ties share a rank.

	Video Quality Metrics ↑						VLM as Judge gemini-3.8-flash ↑			
Method	**Smooth**	**Dyn.**	**Flicker**	**SA**	**PC**	**Joint**	**Phys.**	**Struct.**	**Anat.**	**Overall**
Dynamic-Bench (120 prompts × 3 seeds, $N = 360$)										
SRVS	**0.9895**	0.5500	0.9842	3.1167	3.6667	23.61	2.54	3.06	3.41	3.00
FlowMo	0.9809	**0.7500**	0.9699	3.2000	3.4667	23.06	2.47	2.81	3.04	2.77
NFE×2	0.9816	0.7333	0.9699	3.1833	3.4833	23.33	2.41	2.66	3.11	2.73
CFG-Zero	0.9852	**0.7500**	0.9740	3.1333	3.4333	18.61	2.39	2.61	3.19	2.73
Best-of-4	0.9818	0.7306	0.9712	**3.2306**	3.6083	25.28	2.46	2.71	3.24	2.80
CachedSearch	0.9827	0.7361	0.9735	3.1361	3.5639	23.61	2.52	2.68	3.26	2.82
Structura	**0.9895**	0.6000	**0.9849**	3.1000	**3.7500**	**27.22**	**2.58**	**3.24**	**3.46**	**3.09**
Structura Bench (93 prompts × 3 seeds, $N = 279$)										
SRVS	0.9862	0.7828	0.9776	3.6105	3.9625	47.57	2.77	2.85	2.98	2.87
FlowMo	0.9739	0.9401	0.9577	3.6142	3.7603	42.29	2.49	2.48	2.65	2.54
NFE×2	0.9742	0.9326	0.9581	3.6142	3.7678	43.73	2.57	2.54	2.71	2.61
CFG-Zero	0.9769	**0.9513**	0.9606	3.5843	3.6779	41.94	2.38	2.32	2.48	2.39
Best-of-4	0.9752	0.9319	0.9613	**3.6487**	3.8638	45.52	2.51	2.46	2.68	2.55
CachedSearch	0.9760	0.9355	0.9634	3.5950	3.8244	44.44	2.55	2.44	2.67	2.55
Structura	**0.9873**	0.8390	**0.9790**	3.5843	**3.9813**	**50.19**	**2.81**	**2.94**	**3.01**	**2.92**

Re-noising a stored clean proposal anchor x^{a} at solver level s yields $x^{\mathrm{prop}} = D_{\mathrm{CFG}}((1-s)x^{\mathrm{a}} + s\zeta, s; c)$, where D_{CFG} is the frozen endpoint predictor with classifier-free guidance (Ho & Salimans, 2022). We then test the proposal in the region that motivated intervention. Holding M_n fixed, we evaluate x^{prop} and the current reference x^{ref} with the same K low-noise probes:

$$\widehat{\Delta}_n = \sum_{q \in \Omega} M_n(q) \left[\widehat{\mathcal{E}}_p(x^{\mathrm{prop}}; q) - \widehat{\mathcal{E}}_p(x^{\mathrm{ref}}; q) \right]. \tag{10}$$

Candidate changes within M_n are committed when $\widehat{\Delta}_n < -\eta$, where $\eta \geq 0$ is the acceptance margin; elsewhere we retain P&P's uncertainty-based updates. The anchor advances only after acceptance. Paired evaluation and masked state updates are specified in Appendix B.4–B.5.

4 Experiments

Benchmarks and setup. Existing video benchmarks emphasize motion and physical plausibility; their aggregate scores do not directly capture localized structural hallucinations. We introduce **Structura Bench**, 93 independently collected prompts targeting hallucination-prone anatomy, articulation, contact, occlusion, and object structure. We also curate a high-quality dataset of 10,000 video-editing cases involving hands and complex object interactions to evaluate artifact removal. Dynamic-Bench provides a complementary set of 120 dynamic prompts. We use three seeds per prompt and Wan2.1-T2V-1.3B with a native 40-step schedule, comparing SRVS (Jang et al., 2026), FlowMo (Shaulov et al., 2025), NFE×2, CFG-Zero, Best-of-4, and CachedSearch. **VLM as Judge** uses gemini-3.8-flash to rate physical plausibility, structural correctness, and anatomical correctness; VideoPhy2 reports semantic adherence (SA) and physical commonsense (PC).

Main results. Structura achieves the highest structural, anatomical, and overall VLM-as-judge scores on both benchmarks. Relative to SRVS, structural correctness rises from 3.06 to 3.24 on Dynamic-Bench and from 2.85 to 2.94 on Structura Bench; joint success increases by 3.61 and 2.62 points. Structural fidelity and physical commonsense improve, while SA decreases slightly (3.1000



Figure 4: **Structural refinement across five qualitative cases.** Frames progress left to right within each case; rows compare sampling methods. Red boxes mark baseline artifacts. The bottom panel shows a handshake; EACC denotes Structura.

versus 3.1167; 3.5843 versus 3.6105). Figure 4 shows the resulting body structure and hand–object contacts across five qualitative cases, including a handshake; Appendix E adds the Wan2.2 example.

Backbone and action transfer. VLM-as-judge Overall improves across all three additional backbones (Table 4); joint success increases on Wan2.2 and CogVideoX-2B and is maintained on CogVideoX-5B. On OpenWAM Full, RoboTwin success increases by 0.36 and 0.62 points (Table 7), reaching gains of 5.50 and 3.60 points for tasks below 80% baseline success (Table 2). Additional settings appear in Appendix G.

Trajectory feedback and directional steering. Table 3 isolates trajectory accumulation and directional steering in FGR. Removing directional tilt replaces the guided re-noising direction with isotropic exploration. Removing accumulation uses current-step feedback in place of accumulated revision energy $\widehat{B}_n$ and signed displacement $\widehat{A}_n$. Without tilt, structural scores drop by 0.21 and 0.18; without accumulation, they drop by 0.36 and 0.29, with joint success falling to 17.22% and 33.33%. Both components improve the structural and anatomical dimensions that measure hallucinations. Figure 6 visualizes the spatial feedback; the noise sweep in Fig. 5 favors local probing.

Table 2. **Success by difficulty.** Strata use each setting's baseline rates. Rates are percentages; Δ is in percentage points.

Stratum	N	Base	Ours	Δ
Clean				
$< 80\%$	2	59.00	**64.50**	+5.50
80–89%	6	84.50	**84.67**	+0.17
90–100%	42	97.29	**97.43**	+0.14
Randomized				
$< 80\%$	5	68.60	**72.20**	+3.60
80–89%	6	85.17	**87.00**	+1.83
90–100%	39	97.54	**97.59**	+0.05

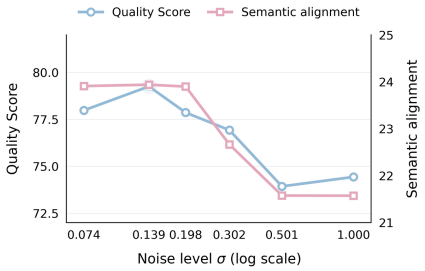


Figure 5: **Probe noise.** Quality peaks near $\sigma = 0.139$.

Table 3. **Ablation of trajectory feedback accumulation and directional steering.** Removing accumulation uses only the current step's reconstruction feedback. VLM as Judge uses gemini-3.8-flash. Metrics and rankings follow Table 1.

	Video Quality Metrics ↑						VLM as Judge gemini-3.8-flash ↑			
Variant	**Smooth**	**Dyn.**	**Flicker**	**SA**	**PC**	**Joint**	**Phys.**	**Struct.**	**Anat.**	**Overall**
Dynamic-Bench (120 prompts × 3 seeds, $N = 360$)										
w/o tilt	0.9868	0.6222	0.9811	3.0806	3.6333	21.39	2.56	3.03	3.28	2.96
w/o accumulation	0.9854	**0.6417**	0.9789	3.0611	3.5250	17.22	2.44	2.88	3.14	2.82
Structura (Full)	**0.9895**	0.6000	**0.9849**	**3.1000**	**3.7500**	**27.22**	**2.58**	**3.24**	**3.46**	**3.09**
Structura Bench (93 prompts × 3 seeds, $N = 279$)										
w/o tilt	0.9841	0.8495	0.9742	3.5627	3.8638	38.71	**2.82**	2.76	2.85	2.81
w/o accumulation	0.9824	**0.8638**	0.9713	3.5412	3.7921	33.33	2.62	2.65	2.72	2.66
Structura (Full)	**0.9873**	0.8390	**0.9790**	**3.5843**	**3.9813**	**50.19**	2.81	**2.94**	**3.01**	**2.92**

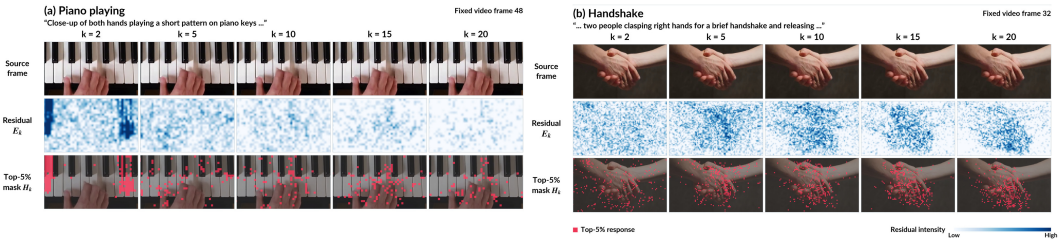


Figure 6: **Spatial feedback across sampling.** Residual maps and top-5% masks at $k = 2, 5, 10, 15, 20$ for piano playing (left) and a handshake (right). The video frame remains fixed; later responses concentrate on the hands and contact regions.

Video editing. Table 5 evaluates paired native and +Structura outputs on DAVIS-2017 validation and Own-100, an evaluation subset of 100 instances drawn from our curated 10k video-editing dataset. Edited-region temporal differences decrease for MiniMax-Remover and YOSE on both datasets, with closely preserved background appearance. ProPainter and DiffuEraser show lower BG LPIPS alongside increased edited-region temporal differences. These measurements characterize how reconstruction feedback changes artifact removal and temporal refinement across editing systems.

Table 4. **Transfer across video backbones.** Phys., Struct., Anat., and Overall use the same 1–5 VLM-as-judge evaluation with gemini-3.8-flash as Table 1. Smooth and Flicker are multiplied by 100; PC is on a 1–5 scale and Joint is a percentage. Higher is better. Blue shades indicate ranks within each backbone; ties share the same rank.

		VLM as Judge gemini-3.8-flash ↑				**AutoEval Metrics ↑**			
Backbone	**Method**	Phys.	Struct.	Anat.	Overall	Smooth	Flicker	PC	Joint (%)
Wan2.2-TI2V-5B	SRVS	3.09	3.41	3.33	3.28	99.32	98.91	4.33	71.11
	Structura	**3.28**	**3.59**	**3.44**	**3.44**	**99.45**	**99.21**	**4.36**	**74.44**
CogVideoX-2B	SRVS	2.82	2.76	2.63	2.74	97.99	96.85	4.36	60.00
	Structura	**2.83**	**2.77**	**2.69**	**2.76**	**98.59**	**97.69**	**4.43**	**72.22**
CogVideoX-5B	SRVS	2.91	2.98	2.84	2.91	98.48	97.41	4.38	**70.00**
	Structura	**3.08**	**3.20**	**3.08**	**3.12**	**98.67**	**97.79**	**4.39**	**70.00**

Table 5. **Video object removal.** D: DAVIS-2017 val (61 instances); O: Own-100, 100 evaluation instances drawn from our curated 10k dataset. Paired rows show native and +Structura means. BG metrics measure background preservation; Edit L1 measures edit activity. Blue shading marks the better value within each pair; Edit L1 is unranked.

Editor	**Set**	**Method**	**BG LPIPS ↓**	**BG L1 ↓**	**BG PSNR ↑**	**Edit L1**	**Edit temp. ↓**	**BG temp. ↓**
MiniMax-Remover	D	Native	**0.129942**	0.023251	29.781806	0.200624	0.028058	0.023150
		+Structura	0.130068	**0.022957**	**29.873736**	0.199401	**0.026324**	**0.023110**
	O	Native	**0.053701**	0.011581	36.685199	0.224363	0.010499	0.005021
		+Structura	0.053805	**0.011321**	**36.841852**	0.224075	**0.010022**	**0.005014**
YOSE	D	Native	0.127607	0.022149	29.729291	0.204853	0.030897	0.023851
		+Structura	**0.127312**	**0.022054**	**29.793262**	0.202882	**0.028937**	**0.023787**
	O	Native	0.057119	0.010213	36.870607	0.235504	0.012232	0.005105
		+Structura	**0.056928**	**0.010184**	**36.922566**	0.235331	**0.011357**	**0.005094**
ProPainter	D	Native	0.033730	**0.008251**	**39.111136**	0.181665	**0.025796**	**0.022562**
		+Structura	**0.033690**	0.008253	39.108641	0.182494	0.028337	0.022563
	O	Native	0.038031	0.005701	**42.487232**	0.211145	**0.014941**	**0.005040**
		+Structura	**0.037979**	**0.005700**	42.483004	0.211543	0.017068	0.005042
DiffuEraser	D	Native	0.064345	**0.018298**	**33.023348**	0.194121	**0.031566**	**0.022326**
		+Structura	**0.064307**	0.018305	33.015121	0.197216	0.034211	0.022333
	O	Native	0.064406	**0.016547**	**34.427252**	0.225673	**0.015984**	**0.005232**
		+Structura	**0.064081**	0.016556	34.416824	0.227329	0.017230	0.005237

5 Conclusion

Structura turns a frozen video generator's reconstruction dynamics into a structural refinement signal. Low-noise probes identify disagreement; trajectory feedback accumulates revision energy and signed displacement, and FGR uses them for localized re-noising and paired selection. Structura Bench measures hallucinations in challenging anatomy and object interactions. Across video backbones, Structura improves structural and anatomical correctness; editing and action-conditioned experiments extend this feedback mechanism to temporal refinement and task completion.

AUTHOR STATEMENTS

AI USE STATEMENT

Generative AI tools assisted with language editing and manuscript preparation. The authors are responsible for the technical content and final manuscript.

ETHICS STATEMENT

Structura aims to reduce structural artifacts in generated videos. Improved visual fidelity can also increase the realism of synthetic media; generated content should be identified as synthetic when presented to users.

REPRODUCIBILITY STATEMENT

The appendix provides the reconstruction geometry, feedback estimators, directional-noise rule, paired comparison, and complete sampling algorithm, together with metric definitions and the action-evaluation configuration.

REFERENCES

Sumukh K Aithal, Pratyush Maini, Zachary C. Lipton, and J. Zico Kolter. Understanding Hallucinations in Diffusion Models through Mode Interpolation. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.09358.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. VideoPhy: Evaluating Physical Commonsense for Video Generation. arXiv preprint arXiv:2406.03520, 2024. URL https://arxiv.org/abs/2406.03520.

Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. VideoJAM: Joint Appearance-Motion Representations for Enhanced Motion Generation in Video Models. In *International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2502.02492.

Junhao Cheng, Liang Hou, Tianxiong Zhong, Xin Tao, Pengfei Wan, Kun Gai, and Jing Liao. VLMs are Good Teachers for Video Reasoning via Adaptive Test-Time Optimization. arXiv preprint arXiv:2606.02564, 2026. URL https://arxiv.org/abs/2606.02564.

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

Mariam Hassan, Kaouther Messaoud, Wuyang Li, and Alexandre Alahi. Proprio: Latent Self-Scoring and Inference-Time Refinement for Physically Plausible Video Generation. arXiv preprint arXiv:2605.28230, 2026. URL https://arxiv.org/abs/2605.28230.

Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. arXiv preprint arXiv:2207.12598, 2022. URL https://arxiv.org/abs/2207.12598.

Sangwon Jang, Taekyung Ki, Jaehyeong Jo, Saining Xie, Jaehong Yoon, and Sung Ju Hwang. Self-Refining Video Sampling. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2601.18577.

Jaemin Kim, Bryan Sangwoo Kim, and Jong Chul Ye. Free^{2}Guide: Training-Free Text-to-Video Alignment using Image LVLM. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2411.17041.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02747.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di Zhang, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving Video Generation with Human Feedback. arXiv preprint arXiv:2501.13918, 2025a. URL https://arxiv.org/abs/2501.13918.

Zichen Liu, Wei Zhang, and Tiejun Li. Improving the Euclidean diffusion generation of manifold data by mitigating score function singularity. In *Advances in Neural Information Processing Systems*, 2025b. URL https://arxiv.org/abs/2505.09922.

NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos World Foundation Model Platform for Physical AI. arXiv preprint arXiv:2501.03575, 2025. URL https://arxiv.org/abs/2501.03575.

Trevine Oorloff, Yaser Yacoob, and Abhinav Shrivastava. Mitigating Hallucinations in Diffusion Models through Adaptive Attention Modulation. arXiv preprint arXiv:2502.16872, 2025. URL https://arxiv.org/abs/2502.16872.

Yuta Oshima, Masahiro Suzuki, Yutaka Matsuo, and Hiroki Furuta. Inference-Time Text-to-Video Alignment with Diffusion Latent Beam Search. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2501.19252.

Frank Permenter and Chenyang Yuan. Interpreting and improving diffusion models from an optimization perspective. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 40461–40483. PMLR, 2024. URL https://proceedings.mlr.press/v235/permenter24a.html.

Ariel Shaulov, Itay Hazan, Lior Wolf, and Hila Chefer. FlowMo: Variance-Based Flow Guidance for Coherent Motion in Video Generation. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2506.01144.

Bartlomiej Sobieski, Matthew Tivnan, Dawid Płudowski, Michał Jan Włodarczyk, Pengfei Jin, Przemyslaw Biecek, and Quanzheng Li. Local Intrinsic Dimension Unveils Hallucinations in Diffusion Models. arXiv preprint arXiv:2605.05026, 2026. URL https://arxiv.org/abs/2605.05026.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and Advanced Large-Scale Video Generative Models. arXiv preprint arXiv:2503.20314, 2025. URL https://arxiv.org/abs/2503.20314.

Yiqi Tian, Pengfei Jin, Mingze Yuan, Na Li, Bo Zeng, and Quanzheng Li. RODS: Robust Optimization Inspired Diffusion Sampling for Detecting and Reducing Hallucination in Generative Models. In *Advances in Neural Information Processing Systems*, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/24c79c21eb699e9efdebaafa40ed0e4d-Abstract-Conference.html.

Kostas Triaridis, Alexandros Graikos, Aggelina Chatziagapi, Grigorios G. Chrysos, and Dimitris Samaras. Mitigating Diffusion Model Hallucinations with Dynamic Guidance. arXiv preprint arXiv:2510.05356, 2025. URL https://arxiv.org/abs/2510.05356.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2408.06072.

Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. VideoREPA: Learning Physics for Video Generation through Relational Alignment with Foundation Models. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2505.23656.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing Video Generation Benchmark Suite for Intrinsic Faithfulness. arXiv preprint arXiv:2503.21755, 2025. URL https://arxiv.org/abs/2503.21755.

A Low-Noise Probe Geometry

Let the clean conditional distribution have a positive smooth density ρ_c on a smooth embedded manifold $\mathcal{M}_c \subset \mathbb{R}^d$. We work in a neighborhood with a unique nearest-point projection $\Pi_{\mathcal{M}_c}$ and a nondegenerate distance minimum. Define $y = z/(1-p)$ and $\sigma = p/(1-p)$. The linear flow observation becomes $y = X_0 + \sigma\epsilon$, and its Bayes-optimal endpoint predictor is the posterior mean

$$D^*(z,p;c) = \frac{\displaystyle\int_{\mathcal{M}_c} X\,\rho_c(X)\exp\!\left(-\frac{\|y-X\|_2^2}{2\sigma^2}\right)\,dV(X)}{\displaystyle\int_{\mathcal{M}_c} \rho_c(X)\exp\!\left(-\frac{\|y-X\|_2^2}{2\sigma^2}\right)\,dV(X)}. \tag{11}$$

Here dV denotes manifold volume. The posterior-mean interpretation follows from squared-error endpoint prediction (Lipman et al., 2023); the Gaussian observation also gives Tweedie's denoising identity (Efron, 2011).

Local projection. Write $X_* = \Pi_{\mathcal{M}_c}(y)$ and introduce tangent coordinates h around X_*. Orthogonality at the nearest point removes the linear term:

$$\|y - X(h)\|_2^2 = \|y - X_*\|_2^2 + h^\top H_y h + \mathcal{O}(\|h\|_2^3), \tag{12}$$

where the nondegenerate minimum gives $H_y \succ 0$. The leading posterior in h is centered with covariance of order σ^2. Smooth density variation and curvature shift its mean by order σ^2; the normal displacement of $X(h)$ is also quadratic. Consequently,

$$D^*(z,p;c) = \Pi_{\mathcal{M}_c}(y) + \mathcal{O}(\sigma^2) = \Pi_{\mathcal{M}_c}\!\left(\frac{z}{1-p}\right) + \mathcal{O}(p^2). \tag{13}$$

This recovers Eq. (3) and the local projection view of denoising (Liu et al., 2025b; Permenter & Yuan, 2024).

Mean probe response. For $z = (1-p)x + p\epsilon$, the rescaled observation is $y = x + \sigma\epsilon$. Smoothness of the projection gives

$$\Pi_{\mathcal{M}_c}(x+\sigma\epsilon) = \Pi_{\mathcal{M}_c}(x) + \sigma J_\Pi(x)\epsilon + \frac{\sigma^2}{2}D^2\Pi(x)[\epsilon,\epsilon] + \mathcal{O}(\sigma^3\|\epsilon\|_2^3). \tag{14}$$

Under the local regularity and integrability conditions of this expansion, the linear term averages to zero for centered Gaussian noise. Thus,

$$\mu_p^*(x) = x - \mathbb{E}_\epsilon D^*((1-p)x + p\epsilon, p; c) = x - \Pi_{\mathcal{M}_c}(x) + \mathcal{O}(p^2). \tag{15}$$

The leading response supplies a deviation magnitude and correction direction. At token q, total response energy decomposes as

$$\mathcal{E}_p(x;q) = \|\mu_p(x;q)\|_2^2 + \operatorname{tr}\operatorname{Cov}_\epsilon[r(x,\epsilon;q)]. \tag{16}$$

This motivates correcting probe variance before accumulating systematic revision energy.

B Feedback Estimation and Refinement Details

B.1 Estimating systematic revision energy

Fix a queried prediction x_n and a latent location q, and suppress these arguments. Let the independent probe responses $r_k \in \mathbb{R}^C$ have mean μ and covariance Σ, with finite second moments. The sample mean and sample variance trace are

$$\bar{r} = \frac{1}{K}\sum_{k=1}^K r_k, \qquad \widehat{v} = \frac{1}{K-1}\sum_{k=1}^K \|r_k - \bar{r}\|_2^2, \qquad K \geq 2. \tag{17}$$

Independence gives

$$\mathbb{E}\|\bar{r}\|_2^2 = \|\mu\|_2^2 + \frac{\operatorname{tr}\Sigma}{K}, \qquad \mathbb{E}\widehat{v} = \operatorname{tr}\Sigma. \tag{18}$$

Subtracting the second term proves the target identity for Eq. (7):

$$\mathbb{E}\widehat{b} = \mathbb{E}\left[\|\overline{r}\|_2^2 - \frac{\widehat{v}}{K}\right] = \|\mu\|_2^2. \tag{19}$$

Equivalently,

$$\widehat{b} = \frac{1}{K(K-1)} \sum_{k \neq l} r_k^{\mathsf{T}} r_l. \tag{20}$$

For $K = 2$, this reduces to $r_1^{\mathsf{T}} r_2$: independent queries contribute through their shared displacement. We accumulate the signed estimate; the positive-part operation is applied when constructing M_n.

B.2 Memory initialization and directional-noise calibration

The vector and scalar memories in Eq. (8) start from $\widehat{A}_1 = \overline{r}_1$ and $\widehat{B}_1 = \widehat{b}_1$. The index n counts refinement rounds throughout the trajectory, including multiple refinement rounds within a solver step. At round n, their weights are

$$w_{n,1} = \beta^{n-1}, \qquad w_{n,j} = (1-\beta)\beta^{n-j} \ (2 \leq j \leq n), \qquad \sum_{j=1}^{n} w_{n,j} = 1. \tag{21}$$

Thus $\widehat{A}_n = \sum_j w_{n,j} \overline{r}_j$ and $\widehat{B}_n = \sum_j w_{n,j} \widehat{b}_j$. For the corresponding mean responses μ_j, define $A_n^{\mu} = \sum_j w_{n,j} \mu_j$ and $B_n^{\mu} = \sum_j w_{n,j} \|\mu_j\|_2^2$. Then

$$B_n^{\mu} - \|A_n^{\mu}\|_2^2 = \sum_j w_{n,j} \|\mu_j - A_n^{\mu}\|_2^2. \tag{22}$$

This identity explains why revision strength is accumulated before signed responses are averaged: disagreement among correction directions remains visible in the energy memory.

For directional calibration, we maintain the smoothed probe variance trace

$$\widehat{V}_1 = \widehat{v}_1, \qquad \widehat{V}_n = \beta \widehat{V}_{n-1} + (1-\beta)\widehat{v}_n. \tag{23}$$

Let $e_j = \overline{r}_j - \mu_j$ be the error of the mean response. With fresh probes at each queried state, e_j has zero mean conditional on the sampling history preceding that batch. Its conditional squared norm has expectation v_j/K, where v_j is the probe covariance trace at that state. Since cross-round error products have zero expectation,

$$\mathbb{E}\left\| \sum_j w_{n,j} e_j \right\|_2^2 = \frac{1}{K} \sum_j w_{n,j}^2 \mathbb{E}[v_j]. \tag{24}$$

Under a locally constant variance scale $v_j \simeq v$, this becomes $v W_n / K$, where

$$W_1 = 1, \qquad W_n = \beta^2 W_{n-1} + (1-\beta)^2. \tag{25}$$

Using $\widehat{V}_n$ for the local variance scale gives the directional probe-noise estimate

$$\widehat{N}_n(q) = \frac{W_n \widehat{V}_n(q)}{K}. \tag{26}$$

B.3 Directional strength and perturbation energy

The directional weight a_n is computed as

$$a_n(q) = \gamma M_n(q) \left[1 - \frac{\widehat{N}_n(q)}{\|\widehat{A}_n(q)\|_2^2}\right]_+, \qquad \gamma \in [0, 1). \tag{27}$$

Here γ caps directional strength, and $M_n(q) = \mathbf{1}\{[\widehat{B}_n(q)]_+ > \tau\}$ selects regions using the energy threshold $\tau \geq 0$. The bracketed term is the fraction of observed displacement energy remaining

after subtracting the estimated probe-noise contribution. If this contribution dominates, the proposal stays isotropic; as the ratio decreases, more of the perturbation is directed along $u_n = -\widehat{A}_n/\|\widehat{A}_n\|_2$. We set $u_n = 0$ and $a_n = 0$ wherever $\widehat{A}_n = 0$.

Conditioned on the accumulated feedback, the fresh Gaussian field ξ is independent of the probe batches. At a location with $\|u_n\|_2 = 1$, Eq. (9) has moments

$$\mathbb{E}_\xi[\zeta] = \sqrt{C}\, a_n u_n, \qquad \mathrm{Cov}_\xi(\zeta) = (1 - a_n^2) I_C. \tag{28}$$

Consequently,

$$\mathbb{E}_\xi\|\zeta\|_2^2 = C a_n^2 + C(1 - a_n^2) = C. \tag{29}$$

For zero displacement, $a_n = 0$ gives the same result. Thus the construction allocates a fixed expected perturbation energy between directed and random components; $\gamma < 1$ retains a stochastic component.

B.4 Paired reconstruction evaluation

With the candidate, reference, and mask fixed, draw K independent evaluation perturbations and use each draw for both predictions. For either prediction x, the local empirical response energy is

$$\widehat{\mathcal{E}}_p(x; q) = \frac{1}{K} \sum_{k=1}^{K} \|r(x, \epsilon_k; q)\|_2^2. \tag{30}$$

Writing $\ell_n(x, \epsilon) = \sum_q M_n(q)\|r(x, \epsilon; q)\|_2^2$, the main-text comparison is exactly

$$\widehat{\Delta}_n = \frac{1}{K} \sum_{k=1}^{K} \left[\ell_n(x^{\mathrm{prop}}, \epsilon_k) - \ell_n(x^{\mathrm{ref}}, \epsilon_k)\right]. \tag{31}$$

This evaluates total reconstruction response within the fixed mask; localization uses the corrected mean-response energy $\widehat{B}_n$. The binary acceptance decision is

$$\chi = \mathbf{1}\{\widehat{\Delta}_n < -\eta\}, \qquad \eta \geq 0. \tag{32}$$

The margin η is measured in summed masked reconstruction energy.

For fixed predictions and mask, the difference has expectation

$$\mathbb{E}[\widehat{\Delta}_n] = \sum_q M_n(q) \left[\mathcal{E}_p(x^{\mathrm{prop}}; q) - \mathcal{E}_p(x^{\mathrm{ref}}; q)\right]. \tag{33}$$

Let $L_{\mathrm{p}} = \ell_n(x^{\mathrm{prop}}, \epsilon)$ and $L_{\mathrm{r}} = \ell_n(x^{\mathrm{ref}}, \epsilon)$. With finite second moments of these losses,

$$\mathrm{Var}(\widehat{\Delta}_n) = \frac{1}{K}\left[\mathrm{Var}(L_{\mathrm{p}}) + \mathrm{Var}(L_{\mathrm{r}}) - 2\,\mathrm{Cov}(L_{\mathrm{p}}, L_{\mathrm{r}})\right]. \tag{34}$$

Shared draws create a paired comparison under matched perturbations; positive covariance reduces the difference variance relative to independent evaluations. This is the common-random-number construction.

B.5 Proposal and reference updates

At the beginning of each solver step, the clean proposal anchor x^{a} and the comparison reference x^{ref} are both initialized to the current clean endpoint prediction. Using guidance scale w_{cfg}, the proposal predictor is

$$\begin{aligned} v_{\mathrm{CFG}}(z, s; c) &= v_\theta(z, s; \varnothing) + w_{\mathrm{cfg}}\big[v_\theta(z, s; c) - v_\theta(z, s; \varnothing)\big], \\ D_{\mathrm{CFG}}(z, s; c) &= z - s v_{\mathrm{CFG}}(z, s; c), \end{aligned} \tag{35}$$

where $\varnothing$ denotes the unconditional input (Ho & Salimans, 2022). The low-noise queries use the conditional predictor D_θ.

Outside M_n, we inherit the uncertainty-based P&P update of Jang et al. (2026). Its binary mask U selects locations using the channel-averaged absolute difference between consecutive P&P clean predictions. Denoting these by $\widetilde{x}^{(m-1)}$ and $\widetilde{x}^{(m)}$, the thresholded mask is

$$U(q) = \mathbf{1}\left\{\frac{1}{C}\|\widetilde{x}^{(m)}(q) - \widetilde{x}^{(m-1)}(q)\|_1 > \tau_U\right\}, \tag{36}$$

where m indexes the P&P iteration and τ_U is the inherited uncertainty threshold. With $\odot$ denoting elementwise multiplication and masks broadcast over channels, local commitment is

$$x^{\mathrm{ref}}_{\mathrm{new}} = x^{\mathrm{ref}} + [\chi M_n + (1 - M_n)U] \odot \left(x^{\mathrm{prop}} - x^{\mathrm{ref}}\right). \tag{37}$$

The anchor advances according to

$$x^{\mathrm{a}}_{\mathrm{new}} = \begin{cases} x^{\mathrm{ref}}_{\mathrm{new}}, & \chi = 1, \\ x^{\mathrm{a}}, & \chi = 0. \end{cases} \tag{38}$$

The paired comparison evaluates the full proposal and reference within M_n; the update above commits the selected spatial components.

C Complete Sampling Algorithm

Algorithm 1 uses completed probe batches to construct each proposal. At every solver step, the anchor and reference reset to the endpoint prediction while feedback memory persists. The first proposal is isotropic because $M_0 = 0$. For a linear-flow Euler step from s to s', the next noisy state is $(s'/s)z_s + (1 - s'/s)D_{\mathrm{CFG}}(z_s, s; c)$.

Algorithm 1 Structura: feedback-guided video sampling

Input: Frozen $D_{\mathrm{CFG}}, D_\theta$, condition c; schedule $1 = s_0 > \cdots > s_L = 0$, rounds $\{m_i\}$; $p, K, \beta, \tau, \gamma, \tau_U, \eta$ with $K \geq 2$.

Sample $z_{s_0} \sim \mathcal{N}(0, I)$; set $n = 0$, $M_0 = 0$, $\widehat{A}_0 = \widehat{B}_0 = \widehat{V}_0 = 0$.
for $i = 0, \ldots, L-1$, with $(s, s') = (s_i, s_{i+1})$:
  $x^{\mathrm{ref}} \leftarrow D_{\mathrm{CFG}}(z_s, s; c)$; $x^{\mathrm{a}} \leftarrow x^{\mathrm{ref}}$.
  $z^{\mathrm{ref}} \leftarrow (s'/s)z_s + (1 - s'/s)x^{\mathrm{ref}}$.
  for $j = 1, \ldots, m_i$:
    Form M_n, u_n, a_n from prior feedback using Eq. (27).
    Draw fresh ξ and construct ζ by Eq. (9).
    $\widetilde{z}_s \leftarrow (1 - s)x^{\mathrm{a}} + s\zeta$; $x^{\mathrm{prop}} \leftarrow D_{\mathrm{CFG}}(\widetilde{z}_s, s; c)$.
    $z^{\mathrm{prop}} \leftarrow (s'/s)\widetilde{z}_s + (1 - s'/s)x^{\mathrm{prop}}$.
    Draw $\epsilon_1, \ldots, \epsilon_K$ independently of each other and of ξ.
    Probe x^{prop} and x^{ref} with these same draws using Eq. (4).
    Evaluate $\widehat{\Delta}_n$ on fixed M_n by Eq. (10); set $\chi \leftarrow \mathbf{1}\{\widehat{\Delta}_n < -\eta\}$.
    $U(q) \leftarrow \mathbf{1}\{\|x^{\mathrm{prop}}(q) - x^{\mathrm{ref}}(q)\|_1 / C > \tau_U\}$.
    $F \leftarrow \chi M_n + (1 - M_n)U$.
    $x^{\mathrm{ref}} \leftarrow F \odot x^{\mathrm{prop}} + (1 - F) \odot x^{\mathrm{ref}}$.
    $z^{\mathrm{ref}} \leftarrow F \odot z^{\mathrm{prop}} + (1 - F) \odot z^{\mathrm{ref}}$.
    if $\chi = 1$: $x^{\mathrm{a}} \leftarrow x^{\mathrm{ref}}$.
    Form $\overline{r}, \widehat{b}, \widehat{v}$ from the proposal probe batch; set $n \leftarrow n + 1$.
    if $n = 1$: initialize $(\widehat{A}_1, \widehat{B}_1, \widehat{V}_1) = (\overline{r}, \widehat{b}, \widehat{v})$, $W_1 = 1$.
    else: update $(\widehat{A}_n, \widehat{B}_n)$ by Eq. (8); update $\widehat{V}_n, W_n$ as in Appendix B.2.
    Set $\widehat{N}_n = W_n \widehat{V}_n / K$.
  end for; set $z_{s'} \leftarrow z^{\mathrm{ref}}$.
end for; **return** z_{s_L}.

D Evaluation Metrics and Additional Analysis

Temporal and task metrics. Motion smoothness and temporal flickering use 0–1 units in Tables 1 and 3, and are multiplied by 100 in Table 4. A higher flickering score means less flicker. Dynamic degree measures the prevalence of motion. VideoPhy2 SA and PC use a 1–5 scale. Joint success is

$$\mathrm{Joint} = \frac{100}{N} \sum_{i=1}^{N} \mathbf{1}\{\mathrm{SA}_i \geq 4 \text{ and } \mathrm{PC}_i \geq 4\}. \tag{39}$$

SA measures adherence to the requested content and action; PC measures physical commonsense.

VLM as Judge (gemini-3.8-flash). gemini-3.8-flash evaluates the generated videos. Physical plausibility concerns support, contact, and motion consistent with the scene. Structural correctness concerns geometry, topology, part counts, and connections over time. Anatomical correctness concerns visible bodies, limbs, hands, and articulation. The 1–5 scales run from severe visible defects to coherent structure. Overall averages the three aspect scores and is distinct from an aggregate VBench score. Tables 1 and 3 use 360 videos per method on Dynamic-Bench and 279 on Structura Bench.

E Additional Qualitative Results

Figure 7 complements the five main-text cases with a Wan2.2 dance sequence. The spatial diagnostics appear in Figure 6.

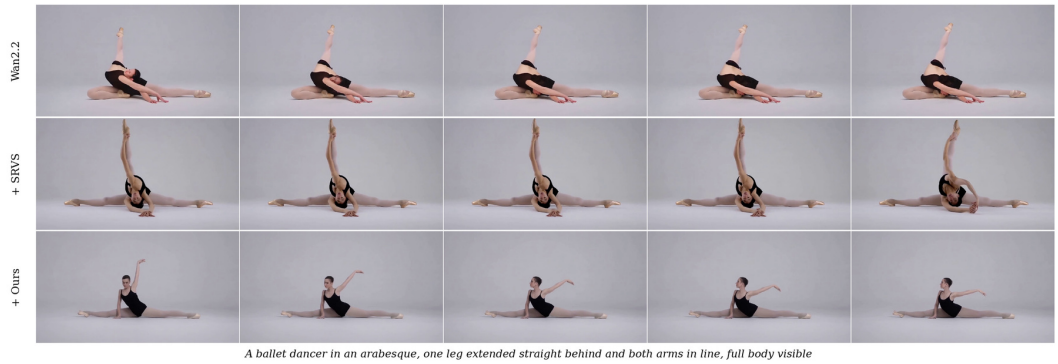


Figure 7: **Additional qualitative comparisons.** Wan2.2 dance. Frames progress left to right; methods are labeled by row.

F VIDEO EDITING EVALUATION

Dataset and evaluation subset. We curate a high-quality dataset of 10,000 video-editing cases involving hands and complex object interactions. Own-100 comprises 100 evaluation instances drawn from this collection and complements the 61 DAVIS-2017 validation object instances. Each native-base/+Structura pair uses the same source instance. Table 5 reports all paired results for the four editors.

Automatic metrics. BG LPIPS, BG L1, and BG PSNR measure preservation of visible background. Edit temporal and BG temporal measure temporal differences in the edited and background regions. Edit L1 measures the amount of change inside the edit region and is an activity diagnostic.

G World Action Model Evaluation

Protocol. We use OpenWAM Full at checkpoint step 118655 with ten denoising steps, disabled DiT cache, and unseen instructions. Each setting contains 50 tasks and 100 episodes per task, paired by task, instruction type, and episode index. Clean uses probe level 0.09, uncertainty threshold 0.4, and at most one refinement per step. Randomized uses probe level 0.12, uncertainty threshold 0.2, and the reference 3/2/1 refinement plan.

Table 6. **Paired outcomes.** Recovered and Lost count episodes won exclusively by Structura and the baseline, respectively.

Outcome	Clean	Random
Baseline	4,711	4,658
Structura	**4,729**	**4,689**
Recovered	123	154
Lost	105	123
Net	+18	+31

Table 7. **RoboTwin success (%).** 5,000 paired episodes per setting.

OpenWAM	Clean	Randomized
Native	94.22	93.16
+Structura	**94.58**	**93.78**

Paired outcomes. Table 6 reports recovered failures and lost successes. The exact two-sided McNemar test on discordant pairs gives $p = 0.2602$ for clean and $p = 0.0713$ for randomized. Table 2 groups tasks by baseline success within each setting. Gains are largest below 80%; perfect-baseline tasks show small decreases. Table 8 gives all per-task results.

Table 8. **Complete RoboTwin results.** Each entry uses 100 episodes. Rates are percentages; Δ is the change in percentage points. Blue shading ranks the paired rates within each task and setting; ties share first rank.

	Clean			Randomized		
Task	**Base**	**Ours**	**Δ**	**Base**	**Ours**	**Δ**
hanging_mug	45	**52**	+7	54	**57**	+3
turn_switch	73	**77**	+4	69	**74**	+5
blocks_ranking_size	**81**	78	−3	78	**81**	+3
put_object_cabinet	81	**83**	+2	**85**	84	−1
stack_bowls_three	84	**88**	+4	80	**88**	+8
pick_diverse_bottles	86	**88**	+2	84	**90**	+6
place_can_basket	**87**	**87**	0	77	**80**	+3
open_microwave	**88**	84	−4	65	**69**	+4
handover_mic	**92**	89	−3	96	**97**	+1
press_stapler	92	**93**	+1	91	**93**	+2
rotate_qrcode	92	**94**	+2	**94**	**94**	0
place_bread_skillet	**93**	91	−2	**88**	87	−1
put_bottles_dustbin	93	**95**	+2	**85**	84	−1
stack_bowls_two	93	**96**	+3	95	**97**	+2
place_burger_fries	94	**98**	+4	**97**	**97**	0
place_dual_shoes	94	**99**	+5	**98**	97	−1
place_phone_stand	**95**	94	−1	**98**	97	−1
dump_bin_bigbin	**96**	95	−1	**97**	**97**	0
pick_dual_bottles	**96**	**96**	0	**96**	95	−1
place_a2b_right	96	**99**	+3	**98**	95	−3
place_bread_basket	96	**97**	+1	**92**	**92**	0
place_object_basket	**96**	**96**	0	**89**	**89**	0
place_object_stand	**96**	**96**	0	**97**	**97**	0
stamp_seal	96	**97**	+1	**96**	95	−1
move_stapler_pad	**97**	**97**	0	96	**99**	+3
scan_object	**97**	94	−3	93	**95**	+2
click_alarmclock	**98**	96	−2	**99**	**99**	0
handover_block	**98**	**98**	0	93	**94**	+1
place_a2b_left	**98**	**98**	0	**99**	98	−1
place_fan	**98**	95	−3	93	**96**	+3
place_object_scale	**98**	96	−2	**98**	96	−2
stack_blocks_three	**98**	**98**	0	**97**	95	−2
beat_block_hammer	99	**100**	+1	**99**	**99**	0
blocks_ranking_rgb	**99**	**99**	0	98	**99**	+1
move_can_pot	99	**100**	+1	**99**	98	−1
place_cans_plasticbox	99	**100**	+1	**100**	98	−2
place_container_plate	99	**100**	+1	**100**	99	−1
place_mouse_pad	99	**100**	+1	99	**100**	+1
adjust_bottle	**100**	**100**	0	99	**100**	+1
click_bell	**100**	**100**	0	**100**	**100**	0
grab_roller	**100**	**100**	0	99	**100**	+1
lift_pot	**100**	**100**	0	**100**	**100**	0
move_pillbottle_pad	**100**	98	−2	**99**	**99**	0
move_playingcard_away	**100**	**100**	0	**100**	**100**	0
open_laptop	**100**	**100**	0	**100**	**100**	0
place_empty_cup	**100**	**100**	0	**100**	**100**	0
place_shoe	**100**	98	−2	**100**	**100**	0
shake_bottle	**100**	**100**	0	**100**	**100**	0
shake_bottle_horizontally	**100**	**100**	0	**100**	99	−1
stack_blocks_two	**100**	**100**	0	99	**100**	+1
All 50 tasks	94.22	**94.58**	+0.36	93.16	**93.78**	+0.62