RETURNING CHARACTERS NEED THEIR ORIGINAL NOISE IN AUTOREGRESSIVE VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoregressive video generators such as NOVA produce a long video one latent frame at a time: causal attention reads earlier frames from a key–value cache, and a flow-matching sampler turns Gaussian noise into the new frame. When a character leaves the scene and later comes back, these generators often return a plausible but different person. We trace this failure to the noise. The cached features describe the earlier frames only partly, and the details they leave open were settled by the noise that generated those frames. Fresh, independent noise redraws these details, so even an exact conditional sampler keeps an expected appearance gap of twice the variance that the conditioning leaves open. We therefore store each generated frame's noise as an extra value in the attention cache and read it back with the attention weights the model already computes: the query that retrieves the character's earlier features also retrieves the noise that drew them. Mixing this noise with fresh noise lets the scene keep changing, and causal flow-matching fine-tuning adapts the generator to it. On 75 leave-and-return scripts with matched NOVA models, the method, SourceValue, raises judged same-person returns from 50.2% to 75.4% at an unchanged return rate; permuting which stored noise each query reads, at matched norm, removes 15.2 points. On Wan2.1-14B, same-person returns rise from 62.40% to 83.64%, 4.00 points above the strongest streaming comparison, with 0.5% (NOVA) and 1.67% (Wan) more generation time.

1 INTRODUCTION

Stories let characters leave and come back. In Figure 1, Luna leaves a market stall while her team prepares it, and 50 seconds into the video she is back sorting the apples. Autoregressive video generators are built for such long, open-ended videos (Deng et al., 2025; Yu et al., 2025a; Huang et al., 2025; Yang et al., 2026). They produce one latent frame or token group at a time, read earlier frames through causal attention over a key–value cache, and turn Gaussian noise into each new latent with a flow-matching sampler. When a character returns, however, it is often a different person of the same role and dress. With matched NOVA models on 75 leave-and-return scripts, human judges see the requested return in 79.9% of judgments but the same individual in only 50.2%.

The usual remedy is more memory: keep more frames, compress distant frames, or retrieve them when needed (Yang et al., 2026; Kim et al., 2026; Xue et al., 2026; Liu et al., 2026). Such memory brings the character's earlier features within reach of the returning frame and leaves the choice among the appearances they still allow to the sampler. A cached latent fixes coarse attributes, such as hair color and clothing, and leaves finer ones, such as face shape, hairline and fabric pattern, partly open. The noise drawn when the character first appeared settled these details; at the return, the sampler draws new, independent noise and settles them again, usually differently. Section 3 makes this precise: with independent noise, the expected squared difference between the two appearances is at least the variance that the conditioning leaves in the earlier one, and twice that for an exact conditional sampler. Better memory lowers this variance; what remains is redrawn at every return.

This analysis points to keeping the earlier draw together with the earlier features. A returning frame that starts from noise correlated with the noise that produced the character before settles the open details in the same way. Two requirements follow. The correlation must reach back across the whole absence, and it must connect the right positions: the noise that drew Luna must be placed where Luna reappears, which can be anywhere in the frame. Noise priors for video diffusion correlate each

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: A character leaves and returns while the scene moves on. Two 60-second SourceValue videos: each character's earlier appearance and return with enlarged crops (top), and timelines of the absence (bottom). The story needs the return to happen and the returning person to be the same individual.

frame with the previous frame or with a shared noise field (Ge et al., 2023; Qiu et al., 2024; Chang et al., 2024). A frame-to-frame chain fails the first requirement, because the link passes through every intermediate frame and decays geometrically once each frame adds fresh noise (Figure 2). A shared field fails the second: it is tied to screen positions, while the character moves.

Our key observation is that the generator already solves this correspondence problem. To draw the returning Luna, the query positions of the new frame attend to the cache entries that hold Luna's earlier features, and these are exactly the entries whose noise is needed. We therefore store, next to the key and value of every generated cache entry, the noise that produced it, and read this noise with the attention weights that the query already computes (Figure 3). The read reaches the earlier occurrence in one step, with a strength set by the current attention weight, for any length of absence. We mix the retrieved noise with fresh Gaussian noise so that the story can continue, and fine-tune the generator with flow matching on this noise. We call the method *SourceValue*, because every cache entry keeps its source noise as an extra value.

On matched NOVA models, SourceValue raises same-person confirmation from 50.2% to 75.4% at an unchanged return rate. Permuting which stored noise each query reads, at fixed norm, removes 15.2 points, and frame-to-frame or shared-field noise reaches only 57.1% and 57.6%. The drop in discrepancy equals twice the gain in cross-covariance, while futures sampled from one prefix differ as much as under independent noise. On a common Wan2.1-14B backbone, it reaches 83.64% Instance against 62.40% for matched IID and 79.64% for Ring Forcing. Our contributions are:

- A diagnosis: independent sampling noise redraws whatever the memory leaves undetermined, which limits how closely a returning appearance can match the earlier one (Section 3).
- SourceValue: the attention cache stores each generated frame's noise, the model's own attention weights retrieve it, and causal flow matching adapts the generator to it (Section 4).
- Experiments on four backbones: common-Wan comparisons, matched controls, and interventions isolating correspondence, the cached record, the covariance mechanism and the cost (Section 5).

2 Related Work

Memory in autoregressive video generation. Video generation has moved from denoising a fixed-length clip as a whole to autoregressive generation, which extends a video frame by frame and accepts new instructions while it runs. Autoregressive image and video models with continuous latents predict each frame or token group with a diffusion or flow head (Li et al., 2024b; Deng et al., 2025; Yu et al., 2025a; Wang et al., 2025a; Han et al., 2026), and causal video diffusion models denoise frames in order (Chen et al., 2024; Yin et al., 2025). Long rollouts raise two problems: errors accumulate, which rollout training (Huang et al., 2025) and error recycling (Li et al., 2026) address,

and a fixed attention window cannot hold a minute of video. Memory methods address the second problem in four ways: short- and long-term conditioning (Henschel et al., 2025; Yang et al., 2026), compression of distant frames into a fixed budget (Zhang et al., 2025; Kim et al., 2026), retrieval or learned storage of relevant distant frames (Yu et al., 2025b; Xiao et al., 2025; Xue et al., 2026), and explicit tracking of entities and their attributes (Liu et al., 2026). These methods decide which earlier content conditions the next frame, and a richer memory leaves less of an earlier appearance open. With independent source noise, details left open by the stored features are sampled anew at each return; this is the bottleneck we study. SourceValue adds the sampling noise to a backbone's existing memory, under the backbone's own retention and compression rule.

Structured noise in video diffusion and flow models. Standard samplers draw independent Gaussian noise for every frame, and a second line of work shapes this noise to control what the sampler produces. Correlated noise priors mix shared and per-frame components so that neighboring frames agree (Ge et al., 2023); noise rescheduling reuses shuffled noise across the windows of a long video (Qiu et al., 2024); and initialization refinement corrects the low-frequency content of the starting noise (Wu et al., 2024). Warping noise along optical flow keeps texture attached to moving surfaces and enables motion control (Chang et al., 2024; Burgert et al., 2025). Other methods change the source distribution of a pretrained flow (Wang et al., 2026b) or search over noise samples at test time (Liu et al., 2025). These works show that source noise affects generation. The temporal noise methods above use frame order, a shared field or motion fields to couple frames; source guidance and test-time search instead modify or select source samples. A character who has left the frame has no motion path to follow, and a frame-to-frame chain loses its link to an appearance many frames back. SourceValue takes correspondence from the generator's attention, which connects the returning positions directly to an earlier occurrence, and adapts the generator to the resulting noise with flow matching (Lipman et al., 2023; Liu et al., 2023).

3 Why a Returning Character Changes

Setting. A generator produces a video as a sequence of latent units $Z_1, \dots, Z_K$, where $Z_k \in \mathbb{R}^{N_k \times p}$ holds N_k spatial positions with p channels. A unit is a whole latent frame for NOVA and HMA, one masked token group for VideoMAR, or a block of up to three latent frames for the causal Wan interface (Appendix B.3). Unit k is generated from Gaussian *source noise* $S_k \in \mathbb{R}^{N_k \times p}$ by integrating a learned velocity field v_θ from time $t = 1$ to $t = 0$, given the conditioning C_k: everything the sampler sees apart from the noise, namely the earlier units in the attention cache, the text or actions released so far, and both classifier-free guidance branches. We write the result as $Z_k = \Phi_\theta(S_k; C_k)$, deterministic in the noise once C_k is fixed. Standard generators draw every S_k independently from $\mathcal{N}(0, I)$, which we call *IID noise*.

The independent-noise floor. Let the returning character last appear in unit $h < k$, and compare appearances by a fixed feature map ψ, e.g., a DINOv3 embedding of the character's segmented crop (Appendix E). With $Y_u = \psi(Z_u)$ for $u \in \{h, k\}$, write $\mu_u = \mathbb{E}[Y_u \mid C_k]$, $V_u = \operatorname{tr}\operatorname{Cov}(Y_u \mid C_k)$ and $K_{k,h} = \operatorname{tr}\operatorname{Cov}(Y_k, Y_h \mid C_k)$. The expected squared difference between them splits into four terms:

$$D = \mathbb{E}\big[\|Y_k - Y_h\|^2 \,\big|\, C_k\big] = \|\mu_k - \mu_h\|^2 + V_k + V_h - 2K_{k,h}. \tag{1}$$

The first three terms describe the two appearance distributions and are what memory acts on: it moves μ_k toward μ_h, and a more detailed memory lowers V_h. The last term, whether the two samples vary together, is decided by the noise. Under IID noise, S_k is independent of (Y_h, C_k), so Y_k and Y_h are conditionally independent and $K_{k,h} = 0$. Consequently $D_{\mathrm{IID}} \geq V_h$, and an exact conditional sampler ($\mu_k = \mu_h$, $V_k = V_h$) attains only $D_{\mathrm{IID}} = 2V_h$ (Appendix A.1). V_h is what the conditioning leaves open about the earlier appearance, and an independent draw redraws all of it.

What a remedy requires. Lowering D below this floor requires $K_{k,h} > 0$, that is, noise for unit k that is correlated with the noise that produced unit h. Correlation helps whenever the two units respond to their noise in the same way. If $X, X' \in \mathbb{R}^d$ are standard Gaussian with $\operatorname{Cov}(X, X') = rI$, and f is square-integrable with degree-q Hermite energy E_q, then

$$\mathbb{E}\,\|f(X) - f(X')\|^2 = 2\sum_{q \geq 1} \big(1 - r^q\big) E_q, \tag{2}$$

which decreases strictly in r for every nonconstant f (Theorem 2, Appendix A.3). Because r^q falls fastest for large q, fine details, which occupy high-degree terms, need strong correlation.

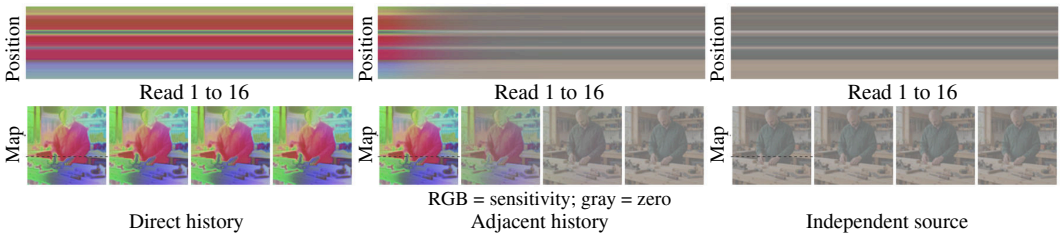


Figure 2: A direct read keeps the dependence on an earlier noise; a frame-to-frame chain loses it. Sensitivity of the current source noise to one fixed earlier source for a direct attention read, frame-to-frame inheritance and IID noise, along one scanline over reads 1–16 (strips) and at reads 1, 3, 6 and 16 (maps); gray marks zero.

The correlation must also survive the absence. If each unit inherits the previous unit's noise and adds fresh noise of variance at least $1 - \rho^2$ to remain Gaussian, the coefficient that links the current noise to the noise L units back is at most ρ^L in operator norm (Theorem 1, Appendix A.2). A ten-second absence spans 30 NOVA latent frames, so at $\rho = 0.5$ less than 10^{-9} of the link remains. A direct link to the earlier unit with weight a gives the coefficient ρa for any L (Figure 2). Finally, the link must reach the right positions, since a character can reappear anywhere; attention over the cache provides such direct, position-specific links.

4 SourceValue: Retrieve the Noise Together with the Features

SourceValue makes three changes to a pretrained generator (Figure 3): it stores each unit's source noise in the attention cache, builds the next unit's noise from a read of this store plus fresh noise, and fine-tunes the generator on the resulting noise. Algorithm 1 (Appendix B.1) gives the full procedure.

4.1 Store each unit's noise next to its features

An earlier draw can be reused only if it is kept, at the granularity at which attention retrieves content. When a generated unit is committed to the cache, each of its positions writes its key and value as usual and also its row $s_j \in \mathbb{R}^p$ of the source noise that produced the unit. Stacking the rows of all M generated entries gives $S_{\mathcal{H}} \in \mathbb{R}^{M \times p}$. The noise shares the key of the features it produced, so any query that selects these features also selects their noise. Input frames, such as an image prompt, have no sampling noise and add content only. Cache operations act on both fields: deleting an entry deletes its noise, and a linear compression W of the values also maps the noise, $S_{\mathcal{H}} \mapsto W S_{\mathcal{H}}$.

4.2 Read the noise with the model's own attention weights

Each position of unit k must decide which stored noise to inherit and how strongly. A position-based rule, such as frame-to-frame or shared-field noise, ignores where the character is; the backbone's attention already answers it. Before sampling unit k, the backbone reads its history: NOVA runs temporal attention before spatial sampling; VideoMAR and HMA mask current and unfinished tokens; and causal Wan queries fixed zero target placeholders with current and future content masked. Each read sees only history and released conditions. Let $A_k^{(\ell)}$ be the softmax attention of head $\ell \in \{1, \dots, H\}$ over all visible entries, and let $A_{k,\mathcal{H}}^{(\ell)} \in \mathbb{R}^{N_k \times M}$ be its columns on generated history, taken after the softmax. With fresh noise $\eta_k \sim \mathcal{N}(0, I)$ drawn after the read, SourceValue sets

$$R_k = \bar{A}_k \, S_{\mathcal{H}}, \quad \bar{A}_k = \frac{1}{H} \sum_{\ell=1}^{H} A_{k,\mathcal{H}}^{(\ell)}, \qquad S_k = \rho \, \mathrm{sg}(R_k) + \sqrt{1 - \rho^2} \, \eta_k, \tag{3}$$

with $0 < \rho < 1$, and $S_k = \eta_k$ when no generated history exists; sg denotes stop-gradient. Three properties of Equation (3) follow from the requirements of Section 3.

The noise follows the content. Each head applies $A_{k,\mathcal{H}}^{(\ell)}$ to its content values. Applying their average to the stored noise gives every position, without new parameters, a combination of the entries the model looks at. A returning position that attends to the character's earlier features thus inherits their noise, wherever they were in the frame, in one step for any gap length.

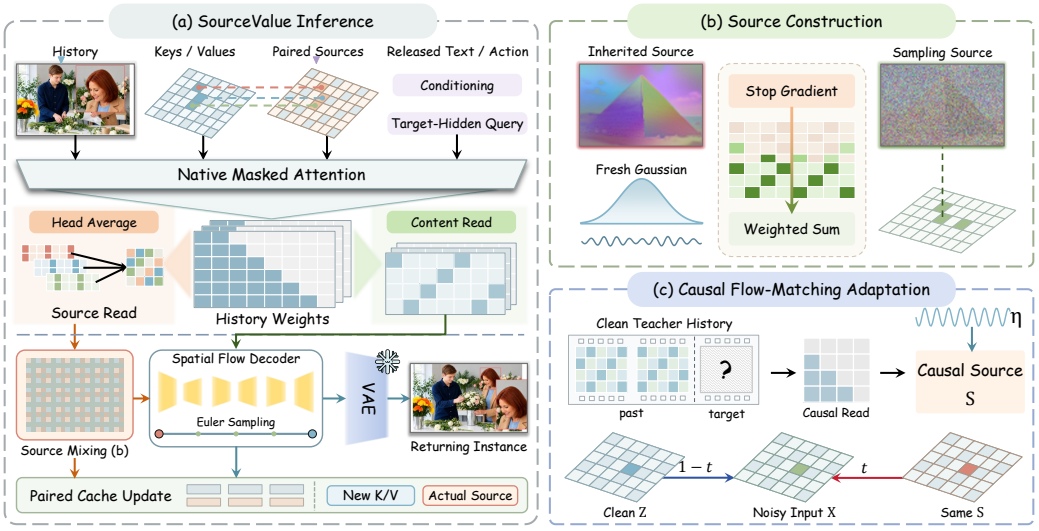


Figure 3: One attention read retrieves both the features and the noise of an earlier occurrence. (a) The head-averaged weights that select cached content also read the stored noise of the same entries; each completed unit commits its key, value and actual source noise. (b) The detached read is mixed with fresh Gaussian noise. (c) Fine-tuning builds this source causally from clean teacher history and uses it in both the noisy input and the velocity target.

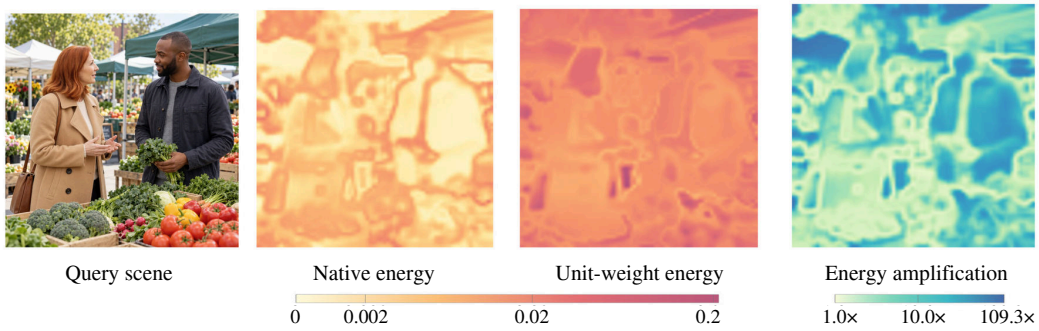


Figure 4: Native attention mass sets how much noise a position inherits. From left: query scene, energy of the native read, energy after renormalizing each row's history weights to one, and their ratio $1/\omega_n^2$. The two energy maps share one scale; the ratio has its own bar ($1.0\times$–$109.3\times$).

Inheritance is as strong as the dependence on history. A row of $\bar{A}_k$ sums to the share $\omega_n \leq 1$ of attention that position n spends on generated history; the rest goes to text, supplied frames and the current unit. We keep this native share. A position that draws a new object places little weight on history and should inherit little noise from it. Renormalizing each row to one would scale the read by $1/\omega_n$ and its energy by $1/\omega_n^2$. Figure 4 shows the consequence on a market scene: newly visible background attends mostly to text and to the current unit and receives a weak native read, while renormalization amplifies its read energy by up to $109\times$, placing strong inherited noise exactly where the model expressed little dependence on the past.

Every direction keeps fresh variation. Reusing the old noise unchanged would push the sampler to reproduce the old frame, whereas the character must move and the scene follow new instructions. The fresh term keeps variance at least $1 - \rho^2$ in every direction. Because stored rows have variance at most one and rows of $\bar{A}_k$ have mass at most one, the source variance at each position also stays at most one, by induction over units. We use $\rho = 0.5$ throughout; both guidance branches share S_k.

4.3 Fine-tune on the noise used at generation

A generator trained with IID noise has learned that its noise carries no information about the history. With correlated noise, the noisy input near $t = 1$ contains the pattern of an earlier frame, which the pretrained velocity field has never seen (Table 3: 53.9% without adaptation). We therefore fine-tune with flow matching on the new source distribution. Each training window of clean latents is

processed in causal order. For unit k we build S_k with Equation (3) from the cache of earlier clean units and their constructed noise, sample $t \sim \mathcal{U}[0,1]$, form $X_{k,t} = (1-t)Z_k + tS_k$, and minimize

$$\mathcal{L}(\theta) = \mathbb{E}\Big[\tfrac{1}{N_k p}\big\|v_\theta(X_{k,t}, t, C_k) - (S_k - Z_k)\big\|_F^2\Big] \tag{4}$$

before committing (Z_k, S_k) to the cache. The same S_k enters the input and the target, so the model learns the straight path from this source to the data, which the sampler integrates backward at generation. The stop-gradient keeps S_k a fixed regression endpoint while attention learns through the content path, and building S_k causally from clean latents reproduces the generation-time read without rolling out the model. The IID control is fine-tuned identically with $S_k = \eta_k$.

5 Experiments

Models and controls. Mechanism experiments use NOVA (0.646B parameters) (Deng et al., 2025), a text-to-video model that reads earlier latent frames by temporal attention, then samples each frame in 64 spatial token groups. From the public checkpoint, each noise rule is fine-tuned on MiraData (Ju et al., 2024) for 20k steps with three seeds; the IID control shares data, steps, learning-rate selection and loss. VideoMAR (Yu et al., 2025a) (masked image-to-video) and HMA (Wang et al., 2025a) (action-conditioned robot video) test two other interfaces. A common Wan2.1-T2V-14B comparison adds seven external methods and a matched IID/SourceValue pair (Appendices B.3 and D.1).

Benchmarks and primary endpoint. NarraStream-Bench releases six instructions at ten-second intervals in each of 324 scripts (Liu et al., 2026); NarrLV gives 360 narratives with one to six temporal narrative atoms (TNAs) up front (Feng et al., 2026); both yield 60-second videos. Conditional long-range consistency (CLC) compares DINO features between an entity's first and later occurrences. The primary endpoint uses the 75 NarraStream scripts whose text asks a character to leave and return. Five blinded judgments per script record whether the return happens (Return) and whether the returning character is the same individual (Instance); failed and unjudgeable outputs count as negatives. We report means of three complete repeats and paired within-repeat effects with 95% t-intervals (Appendix C).

5.1 Does the returning character stay the same?

Table 1: Long-video results (0–100; higher is better). SC/BC: subject/background consistency; Qual.: quality; Fid./Cov./Coh.: fidelity/coverage/coherence. NarraStream scores use 324 scripts; Return/Instance use the same 75 scripts with five judgments each. NarrLV uses 360 narratives, with coherence on the 300 multi-action cases. Means of three repeats. Native systems retain their checkpoints; Wan rows share the Wan2.1-T2V-14B root and use method-specific budgets (Appendix D.1). Each IID/SourceValue pair is matched. Dashes mark full-story methods without streaming support.

Method	NarraStream (streaming)						NarrLV (full narrative)		
	SC	BC	CLC	Qual.	Return	Instance	Fid.	Cov.	Coh.
Complete systems with streaming support (native backbones and budgets)									
Ring Forcing (Xue et al., 2026)	**96.2**	**94.2**	**94.2**	**84.1**	**82.8**	**77.7**	**83.5**	**74.7**	**74.2**
MemRoPE (Kim et al., 2026)	95.8	93.9	92.8	83.4	80.1	70.9	81.6	70.6	69.5
LongLive (Yang et al., 2026)	95.3	93.4	89.9	82.8	80.1	64.4	79.9	65.6	60.8
Stable Video Infinity (Li et al., 2026)	93.8	92.6	84.8	82.5	79.4	65.2	82.1	70.8	68.7
Self Forcing (Huang et al., 2025)	80.9	82.4	40.2	67.4	70.3	44.2	78.1	62.1	54.9
Matched NOVA d48w1024–osp480 (0.646B parameters) (Deng et al., 2025)									
IID	94.3	**93.4**	86.0	83.2	79.9	50.2	**80.2**	67.1	63.4
SourceValue	**94.5**	93.3	**93.4**	**83.3**	**80.7**	**75.4**	**80.2**	**69.3**	**68.0**
Common Wan2.1-T2V-14B (Team Wan et al., 2025) (480p, 16 fps; method-specific budgets)									
Ring Forcing (Xue et al., 2026)	**96.07**	**94.44**	94.16	**84.59**	85.16	79.64	84.28	74.85	75.08
MemRoPE (Kim et al., 2026)	95.81	94.23	93.05	84.31	83.82	75.02	83.26	72.53	72.29
LongLive (Yang et al., 2026)	95.59	93.89	90.89	84.08	83.73	70.84	82.58	69.56	66.03
World-R1-Large (Wang et al., 2026a)	—	—	—	—	—	—	**85.00**	75.87	73.17
Stable Video Infinity (Li et al., 2026)	95.14	93.61	89.21	83.97	83.11	72.36	83.99	73.45	71.05
Self Forcing (Huang et al., 2025)	89.32	88.77	68.15	80.09	77.69	55.73	81.21	67.35	62.30
Video-T1 (chunk tree) (Liu et al., 2025)	—	—	—	—	—	—	84.95	**77.60**	76.27
IID	95.89	94.42	89.00	84.30	84.80	62.40	84.19	73.07	69.91
SourceValue	95.94	94.15	**95.71**	84.31	**85.69**	**83.64**	84.15	75.42	**79.17**

On NOVA (Table 1), SourceValue raises Instance from 50.2% to 75.4% (+25.16 points; 95% CI 12.83–37.48), while Return moves from 79.9% to 80.7% (+0.80; −8.40–10.00). The two models



Figure 5: The person who leaves is the person who returns, in eight settings. Each column is one SourceValue video: an early appearance (5 s), the scene during the absence (15 s) and the return (55 s), with enlarged crops of the leaving and returning person.

carry out the return equally often; SourceValue changes who comes back. CLC, which measures the same property with features, rises by 7.35 points (6.00–8.70), while subject/background consistency and quality stay within 0.2 points. The gain stays above 16 points under any resolution of unjudgeable answers (Appendix C.3). On Wan2.1-14B, Instance rises from 62.40% to 83.64% (+21.24 points), exceeding Ring Forcing by 4.00 points; CLC rises from 89.00 to 95.71 and Return from 84.80% to 85.69%. The external Wan rows compare complete systems, each at its own method-specific budget (Appendix D.1).

Returns across settings. In the eight videos of Figure 5, the returning person keeps the face, hairstyle and clothing of the early appearance while the scene changes, such as the frames assembled in the two workshops.

Longer stories show a larger benefit. On NOVA, NarrLV coherence rises from 63.4 to 68.0 at equal fidelity (80.2). Figure 6 shows where this gain arises: the coherence gap widens from about 1 point for stories with two TNAs to about 8 points for six, coverage separates in the same direction, and fidelity stays within one point. From two to six TNAs, IID loses about 22 coherence points and SourceValue about 15. Each added narrative atom is another chance to redraw a character, and coherence registers each such break.

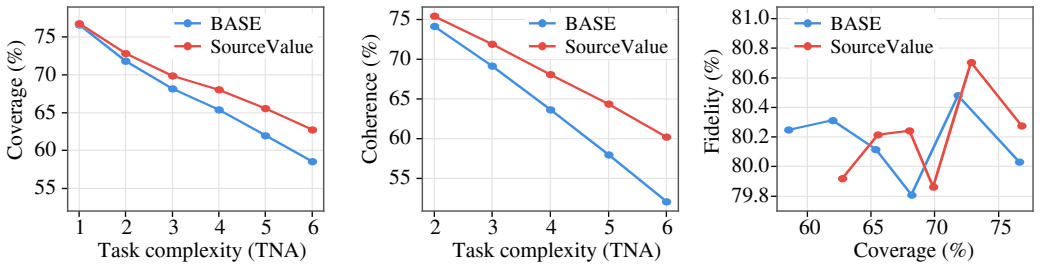
5.2 Does the gain come from retrieving the matching noise?

Which earlier noise is retrieved matters. At the first instruction transition, we permute the stored noise rows before the read and rescale each position's read to its original norm, keeping content, attention and fresh noise fixed; later updates follow the normal rule. This single change lowers Instance by 15.20 points (95% CI 8.58–21.82) and CLC by 2.7 points, while Return barely moves (Table 2). It also lowers VideoMAR transition completion (40.0 to 32.5) and HMA action sensitivity (4.71 to 4.46 dB). The permuted read keeps the original's strength and marginal statistics, so the loss comes from which earlier noise it contains: the character still returns on cue, but more often as someone else.

Table 2: Single-update interventions followed by normal rollout. NarraStream uses NOVA; TC-I2V uses VideoMAR; LT uses HMA on LanguageTable. All metrics are higher-is-better means; ΔPSNR is in dB, other scores are on 0–100 scales.

Variant	NarraStream			TC-I2V	LT
	CLC	Return	Inst.	TCR	ΔPSNR
SourceValue	93.4	**80.7**	**75.4**	**40.0**	**4.71**
Address shuffle	90.7	80.0	60.2	32.5	4.46
No innovation	**93.6**	73.4	65.1	33.7	4.31

The generator must learn to use the retrieved noise. The same read on IID-trained weights reaches only 53.9% Instance (Table 3), although its CLC of 89.2 already exceeds the IID level of 86.0: the retrieved noise carries identity information, and adaptation (Section 4.3) teaches the velocity field to use it.



(a) Required content (b) Cross-sentence coherence (c) Fidelity–coverage trajectory

Figure 6: The coherence advantage grows with the number of narrative atoms. Matched NOVA SourceValue and IID (BASE) on the 360 NarrLV narratives; each point averages 60 narratives and three seeds. (a) Coverage of requested content; (b) coherence of transitions between covered atoms; (c) fidelity against coverage, connected in order of TNA count (Appendix C.1).

Attention finds the matching noise; position rules miss it. Table 3 compares complete noise rules, each fine-tuned under its own rule at equal ρ. Frame-to-frame inheritance (the previous frame's noise at the same position) reaches 57.1% Instance and a shared field (the first generated frame's noise) 57.6%, against 75.4% for SourceValue. Both correlate noise across frames, yet neither follows the character: the chain fades during the absence, and the shared field is tied to screen positions. Their NarrLV coherence (64.6 and 63.4) stays at the IID level (63.4), below the 68.0 of SourceValue (Appendix D.2).

Table 3: Complete source laws on NOVA d48w1024 (0.646B). Adapted rows are separately trained under their own laws. Quality supplements identity and return; the final column is NarrLV coherence. Means on 0–100 scales; higher is better.

Source law	NarraStream				NLV
	CLC	Return	Inst.	Qual.	Coh.
Addressed, adapted	**93.4**	**80.7**	**75.4**	**83.3**	**68.0**
Addressed, IID-trained	89.2	78.5	53.9	82.9	65.1
Adjacent, adapted	86.9	77.7	57.1	83.2	64.6
Global, adapted	89.3	76.2	57.6	83.2	63.4

5.3 Does the gain depend on the earlier record in the cache?

Removing the earlier record removes the advantage. At the middle instruction transition (about 30 s), we delete all older cache entries, content and noise together, while keeping the latest segment, and cross this deletion with one SourceValue or IID update (Table 4). With history kept, the SourceValue update leads IID by 1.35 CLC points; with history deleted, by 0.32 (interaction 1.03; 95% CI 0.56–1.50). Instance behaves alike: 75.4% against 68.5% with history, and 50.8% against 49.0% without. The direct read reuses an earlier draw while its record is available, as Section 3 requires.

Table 4: History retention crossed with one source update (NOVA). The latest completed segment is retained. Relevant- and all-history deletion coincide on all 75 return scripts. Qual. adds the full-population quality endpoint. Means on 0–100 scales; higher is better.

Update	NarraStream			
	CLC	Return	Inst.	Qual.
Older relevant history retained				
Addressed	**93.4**	**80.7**	**75.4**	**83.3**
IID	92.0	79.0	68.5	**83.3**
Older relevant history deleted				
Addressed	**87.1**	**79.6**	**50.8**	**83.0**
IID	86.8	79.0	49.0	**83.0**

The content selects the person; the noise settles its details. At the return, we write original or donor content, crossed with original or donor noise, into the persistent cache and judge who appears later (Table 5). Donor content makes the donor recognizable in about 23% of judgments and cuts recognition of the original to 13–15%, under either noise. With original content, the original noise yields 43.0% original recognition against 38.6% with donor noise (interaction 2.49; −3.70–8.69). As Equation (1) describes, the noise acts on the appearance the content selects: it settles only what the conditioning leaves open.

Table 5: Future identity after content/source exchange. Mean percentages over 321 scripts with five judgments each. The six outcomes are exclusive; small deviations from 100 reflect rounding. Indist.: original/donor indistinguishable; Absent: required but absent; Unjudg.: unjudgeable. Bold marks the strongest confirmation of the written-content identity within its block, not a ranking of all response categories.

Source record	Recognized identity			Other outcomes		
	Original	Donor	Other	Indist.	Absent	Unjudg.
Original content written						
Original	**43.0**	1.8	3.1	39.3	3.6	9.3
Donor	38.6	2.7	4.4	38.5	4.8	11.1
Donor content written						
Original	14.7	23.1	4.9	37.8	7.0	12.5
Donor	12.7	**23.3**	5.5	37.5	7.3	13.6

A correct noise record keeps helping; a wrong one keeps hurting. After the first unit past ten seconds, we write or skip its content and noise in the persistent cache, with the noise retrieved normally or permuted (Appendix D.4). Writing widens the retrieved-minus-permuted gap by 2.02 CLC points (1.64–2.41) and 2.78 NarrLV coherence points (2.31–3.26), and the gap persists over later segments. Because later units reread every committed noise, one permuted update costs 15 Instance points (Table 2).

5.4 Does the gain act through the cross-covariance?

Equation (1) predicts that SourceValue lowers the discrepancy by raising $K_{k,h}$, without narrowing the returning appearance (V_k). Table 6 tests this on object features (SAM 3 masks, DINOv3 embeddings), comparing the SourceValue read with an IID draw on the same histories (Appendix E).

On NarraStream, the discrepancy falls from 42.2 to 24.7, a drop of 17.5; twice the gain in cross-covariance is 17.6; the current variance changes by 0.1; and the implied mean gap is zero to rounding. On NarrLV, the drop is 13.3 against 13.0. The improvement thus lies in the term that IID noise zeroes for an ideal sampler; the IID covariance is positive because pixel features also see history through the decoder. Simultaneous bounds on these moments are wide and include zero gain (Appendix E.3); the causal evidence is in Sections 5.2 and 5.3.

Table 6: Natural-condition object quantities in 10^{-2} units. D: discrepancy; V_k: current conditional variance; V_h: historical conditional variance; $K_{k,h}$: historical cross-covariance. Point means describe feature variation and historical dependence. Simultaneous uncertainty bounds appear in the text.

Source	D	V_k	V_h	$K_{k,h}$
NarraStream (NOVA d48w1024, 0.646B)				
Addressed	24.7	29.3	27.4	16.0
IID	42.2	29.2	27.4	7.2
NarrLV (NOVA d48w1024, 0.646B)				
Addressed	22.9	31.5	29.9	19.2
IID	36.2	31.7	29.9	12.7

5.5 What does inheriting noise cost?

Variation among futures is preserved. If inherited noise pulled the video toward its past, futures sampled from one prefix would look alike. Over ten 50-second futures for each of the 324 NarraStream prefixes (Appendix D.5), SourceValue raises full-video CLC from 85.7 to 93.3, while the LPIPS between futures stays at 0.144 against 0.143 (effect 0.0009; 95% CI −0.0003–0.0021). Removing the fresh noise at one update keeps CLC (93.2) but lowers LPIPS to 0.138 and Return to 73.4% (Table 2). SourceValue inherits only where, and as strongly as, a query attends to history, so the returning character becomes consistent while the rest still varies. Figure 7 shows this within one frame: among four futures that share a prefix, local differences concentrate on the moving child, whose new pose the history does not contain, and stay small on the corridor that earlier frames showed.



Figure 7: Futures differ where the content is new. Four SourceValue futures from one prefix (left) and one minus their mean local RGB cosine similarity (right).

Short tasks change little, at a small added cost. Effects on NOVA dynamic degree (−1.30; 95% CI −3.91–1.32), StoryEval events (−0.20; −0.65–0.25), VideoMAR transition completion (0.33; −0.08–0.75) and HMA action sensitivity (−0.014 dB; −0.435–0.407) are small: short tasks contain no long absence (Appendix D.6). A 60-second NOVA video takes 7.0 s (0.5%) longer and stores 15.91 MiB of noise; Wan adds 1.67% time and 8.76 MiB of source records (Appendix D.7).

6 Conclusion

A returning character is sampled twice. Memory lets the second sample see the first; SourceValue also lets it reuse the first sample's noise, found by the attention weights that find the character's features, raising same-person returns from 50.2% to 75.4% on NOVA and from 62.40% to 83.64% on Wan2.1-14B, while the fresh noise it adds keeps the rest of the video free to change. A sequential generator should keep the random choices behind its earlier outputs.

AI USE STATEMENT

Generative AI assisted related-work retrieval and synthesis, the refinement of research questions and experimental plans, mathematical and statistical checks, figure presentation, and the drafting and polishing of manuscript sections. The authors reviewed AI-assisted material and take responsibility for the final text, citations, methods, claims, and experimental evidence.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025b. URL https://arxiv.org/abs/2502.13923.

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 *CVPR*, pp. 13–23, 2025. URL https://arxiv.org/abs/2501.08331.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollar, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment Anything with Concepts. In *ICLR*, 2026. URL https://arxiv.org/abs/2511.16719.

Pascal Chang, Jingwei Tang, Markus Gross, and Vinicius C. Azevedo. How I Warped Your Noise: a Temporally-Correlated Noise Prior for Diffusion Models. In *ICLR*, 2024. URL https://openreview.net/forum?id=pzElnMrgSD.

Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion Forcing: Next-token Prediction Meets Full-Sequence Diffusion. In *NeurIPS*, 2024. URL https://arxiv.org/abs/2407.01392.

Ricky T. Q. Chen and Yaron Lipman. Flow matching on general geometries. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=g7ohDlTITL.

Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive Video Generation without Vector Quantization. In *ICLR*, 2025. URL https://arxiv.org/abs/2412.14169.

Weixi Feng, Jiachen Li, Michael Saxon, Tsu-Jui Fu, Wenhu Chen, and William Yang Wang. TC-Bench: Benchmarking Temporal Compositionality in Conditional Video Generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4638–4662, 2025. doi: 10.18653/v1/2025.findings-acl.241. URL https://aclanthology.org/2025.findings-acl.241/.

Xiaokun Feng, Haiming Yu, Meiqi Wu, Shiyu Hu, Jintao Chen, Chen Zhu, Jiahong Wu, Xiangxiang Chu, and Kaiqi Huang. NarrLV: Towards a Comprehensive Narrative-Centric Evaluation for Long Video Generation. In *ICLR*, 2026. URL https://arxiv.org/abs/2507.11245.

Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve Your Own Correlation: A Noise Prior for Video Diffusion Models. In *ICCV*, 2023. URL https://arxiv.org/abs/2305.10474.

Chunrui Han, Guopeng Li, Jingwei Wu, Quan Sun, Yan Cai, Yuang Peng, Zheng Ge, Deyu Zhou, Haomiao Tang, Hongyu Zhou, Kenkun Liu, Shu-Tao Xia, Binxing Jiao, Daxin Jiang, Xiangyu Zhang, and Yibo Zhu. NextStep-1: Toward Autoregressive Image Generation with Continuous Tokens at Scale. In *ICLR*, 2026. URL https://openreview.net/forum?id=Ndnwg9oOQO.

Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. StreamingT2V: Consistent, Dynamic, and Extendable Long Video Generation from Text. In *CVPR*, 2025. URL https://arxiv.org/abs/2403.14773.

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*, volume 30, 2017. URL https://arxiv.org/abs/1706.08500.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the Train-Test Gap in Autoregressive Video Diffusion. In *NeurIPS*, 2025. URL https://arxiv.org/abs/2506.08009.

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 *CVPR*, pp. 21807–21818, 2024. doi: 10.1109/CVPR52733.2024.02060. URL https://arxiv.org/abs/2311.17982.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and Versatile Benchmark Suite for Video Generative Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3268–3285, 2026. doi: 10.1109/TPAMI.2025.3633890. URL https://doi.org/10.1109/TPAMI.2025.3633890.

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. MiraData: A Large-Scale Video Dataset with Long Durations and Structured Captions. In *NeurIPS*, 2024. URL https://arxiv.org/abs/2407.06358.

Youngrae Kim, Qixin Hu, C.-C. Jay Kuo, and Peter A. Beerel. MemRoPE: Training-Free Infinite Video Generation via Evolving Memory Tokens. In *ECCV*, 2026. URL https://arxiv.org/abs/2603.12513.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy Visual Task Transfer. *arXiv preprint arXiv:2408.03326*, 2024a. URL https://arxiv.org/abs/2408.03326.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive Image Generation without Vector Quantization. In *NeurIPS*, 2024b. URL https://arxiv.org/abs/2406.11838.

Wuyang Li, Wentao Pan, Po-Chien Luan, Yang Gao, and Alexandre Alahi. Stable Video Infinity: Infinite-Length Video Generation with Error Recycling. In *ICLR*, 2026. URL https://arxiv.org/abs/2510.09212.

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

Fangfu Liu, Hanyang Wang, Yimo Cai, Kaiyan Zhang, Xiaohang Zhan, and Yueqi Duan. Video-T1: Test-time Scaling for Video Generation. In *ICCV*, pp. 18671–18681, 2025. doi: 10.1109/ICCV51701.2025.01735. URL https://arxiv.org/abs/2503.18942.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved Reasoning, OCR, and World Knowledge, 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/. Research blog, January 30; accessed 2026-09-24.

Jinzhuo Liu, Jiangning Zhang, Wencan Jiang, Yabiao Wang, Dingkang Liang, Zhucun Xue, Ran Yi, and Yong Liu. Advancing Narrative Long Video Generation via Training-Free Identity-Aware Memory. *arXiv preprint arXiv:2605.18733*, 2026. URL https://arxiv.org/abs/2605.18733.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *ICLR*, 2023. URL https://arxiv.org/abs/2209.03003.

Corey Lynch, Ayzaan Wahid, Jonathan Tompson, Tianli Ding, James Betker, Robert Baruch, Travis Armstrong, and Pete Florence. Interactive Language: Talking to Robots in Real Time. *arXiv preprint arXiv:2210.06407*, 2022. URL https://arxiv.org/abs/2210.06407.

Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. FreeNoise: Tuning-Free Longer Video Diffusion via Noise Rescheduling. In *ICLR*, 2024. URL https://arxiv.org/abs/2310.15169.

Qwen Team. Qwen3.5-27B, 2026. URL https://huggingface.co/Qwen/Qwen3.5-27B. Official model card; accessed 2026-09-24.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021. URL https://arxiv.org/abs/2103.00020.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. URL https://arxiv.org/abs/2508.10104.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2V-CompBench: A Comprehensive Benchmark for Compositional Text-to-video Generation. In *CVPR*, pp. 8406–8416, 2025. URL https://arxiv.org/abs/2407.14505.

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.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards Accurate Generative Models of Video: A New Metric & Challenges. *arXiv preprint arXiv:1812.01717*, 2018. URL https://arxiv.org/abs/1812.01717.

Lirui Wang, Kevin Zhao, Chaoqi Liu, and Xinlei Chen. Learning Real-World Action-Video Dynamics with Heterogeneous Masked Autoregression. *arXiv preprint arXiv:2502.04296*, 2025a. URL https://arxiv.org/abs/2502.04296.

Weijie Wang, Xiaoxuan He, Youping Gu, Yifan Yang, Zeyu Zhang, Yefei He, Yanbo Ding, Xirui Hu, Donny Y. Chen, Zhiyuan He, Yuqing Yang, and Bohan Zhuang. World-R1: Reinforcing 3D Constraints for Text-to-Video Generation. In *ICML*, 2026a. URL https://arxiv.org/abs/2604.24764.

Yiping Wang, Xuehai He, Kuan Wang, Luyao Ma, Jianwei Yang, Shuohang Wang, Simon Shaolei Du, and Yelong Shen. Is Your World Simulator a Good Story Presenter? A Consecutive Events-Based Benchmark for Future Long Video Generation. In *CVPR*, pp. 13629–13638, 2025b. URL https://arxiv.org/abs/2412.16211.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861. URL https://doi.org/10.1109/TIP.2003.819861.

Zifan Wang, Alice Harting, Matthieu Barreau, Michael M. Zavlanos, and Karl H. Johansson. Source-Guided Flow Matching. In *ICLR*, 2026b. URL https://arxiv.org/abs/2508.14807.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-Align: Teaching LMMs for Visual Scoring via Discrete Text-Defined Levels. *arXiv preprint arXiv:2312.17090*, 2023. URL https://arxiv.org/abs/2312.17090.

Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. FreeInit: Bridging Initialization Gap in Video Diffusion Models. In *ECCV*, 2024. URL https://arxiv.org/abs/2312.07537.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. WorldMem: Long-term Consistent World Simulation with Memory. In *NeurIPS*, 2025. URL https://arxiv.org/abs/2504.12369.

Bowen Xue, Brandon Y. Feng, Chenguo Lin, Yuchen Lin, Yujia Zeng, Lvmin Zhang, Maneesh Agrawala, Honglei Yan, and Panwang Pan. Ring Forcing: Towards Precise Long-Term Memory for Autoregressive Video Diffusion. In *ECCV*, 2026. URL https://arxiv.org/abs/2608.26794.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. LongLive: Real-time Interactive Long Video Generation. In *ICLR*, 2026. URL https://arxiv.org/abs/2509.22622.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From Slow Bidirectional to Fast Autoregressive Video Diffusion Models. In *CVPR*, 2025. URL https://arxiv.org/abs/2412.07772.

Hu Yu, Biao Gong, Hangjie Yuan, DanDan Zheng, Weilong Chai, Jingdong Chen, Kecheng Zheng, and Feng Zhao. VideoMAR: Autoregressive Video Generation with Continuous Tokens. In *NeurIPS*, 2025a. doi: 10.52202/085713-1907. URL https://doi.org/10.52202/085713-1907.

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as Memory: Scene-Consistent Interactive Long Video Generation with Memory Retrieval. In *SIGGRAPH Asia*, 2025b. URL https://arxiv.org/abs/2506.03141.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame Context Packing and Drift Prevention in Next-Frame-Prediction Video Diffusion Models. In *NeurIPS*, 2025. URL https://arxiv.org/abs/2504.12626.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*, 2018. URL https://arxiv.org/abs/1801.03924.

A THEORY

This appendix proves the statements used in Section 3. Section A.1 derives the discrepancy decomposition and the floor under IID noise. Section A.2 bounds how much dependence on an earlier noise a frame-to-frame chain can retain (Theorem 1). Section A.3 relates noise correlation to appearance agreement for a shared response (Theorem 2). Section A.4 combines the two results into a strict comparison and treats normalization, and Section A.5 lists the conditions under which the conclusions fail.

A.1 THE DISCREPANCY DECOMPOSITION AND THE IID FLOOR

Fix the current unit k and write $C = C_k$. All conditional identities hold almost surely in C. The readout ψ is fixed before sampling, and Y_h, Y_k are square-integrable. For $u \in \{h, k\}$ let $\widetilde{Y}_u = Y_u - \mu_u$. Since $\mathbb{E}[\widetilde{Y}_u \mid C] = 0$, the cross terms between means and residuals vanish, and

$$\mathbb{E}\big[\|Y_k - Y_h\|^2 \;\big|\; C\big] = \|\mu_k - \mu_h\|^2 + \mathbb{E}\big[\|\widetilde{Y}_k - \widetilde{Y}_h\|^2 \;\big|\; C\big] = \|\mu_k - \mu_h\|^2 + V_k + V_h - 2K_{k,h}, \quad (5)$$

which is Equation (1). Under IID noise, the conditional law of S_k given (Y_h, C) equals its unconditional law. The map $S_k \mapsto \psi(\Phi_\theta(S_k; C))$ is fixed given C, so Y_k and Y_h are conditionally independent and $K_{k,h} = 0$. The remaining terms are nonnegative, hence $D_{\mathrm{IID}} \geq V_h$, with equality when the current readout equals μ_h deterministically. If $\mu_k = \mu_h$ and $V_k = V_h$, then $D_{\mathrm{IID}} = 2V_h$: a sampler with the correct conditional mean and variance still disagrees with the earlier realization.

To compare two noise rules $\pi \in \{\mathrm{IID}, \mathrm{SV}\}$ under the same historical law and the same C, write $G^\pi = \|\mu_k^\pi - \mu_h\|^2$ for the mean gap, V_k^π for the current variance and $K_{k,h}^\pi$ for the cross-covariance. Subtracting the two instances of Equation (5) gives

$$D^{\mathrm{IID}} - D^{\mathrm{SV}} = 2K_{k,h}^{\mathrm{SV}} + \big(G^{\mathrm{IID}} - G^{\mathrm{SV}}\big) + \big(V_k^{\mathrm{IID}} - V_k^{\mathrm{SV}}\big). \quad (6)$$

The historical variance cancels. A positive cross-covariance therefore lowers the discrepancy unless the new noise rule also enlarges the mean gap or the current variance by more; Section 5.4 measures all three terms. Averaging over a common distribution of C preserves both identities. A readout that depends on decoder history outside C needs an extra covariance term, treated in Appendix E.1.

A.2 DEPENDENCE RETAINED BY A FRAME-TO-FRAME CHAIN

We model frame-to-frame inheritance as a linear chain $U_{i+1} = B_i U_i + \epsilon_{i+1}$ in $\mathbb{R}^d$, $i = 0, \ldots, L-1$, with deterministic matrices B_i, $\mathrm{Cov}(U_0) \preceq I$, and centered innovations ϵ_{i+1} with covariance Q_i, independent of U_0 and of one another. The influence bound needs only these covariances; the response analysis in Section A.4 additionally assumes a jointly Gaussian chain.

Theorem 1 (Frame-to-frame dependence). *If $B_i B_i^\top + Q_i \preceq I$ and $Q_i \succeq (1-\rho^2)I$ for all i, the L-step coefficient satisfies $\|B_{L-1} \cdots B_0\|_{\mathrm{op}} \leq \rho^L$. A direct read that gives the earlier source weight a has coefficient $\rho a I$ on that source when the other records are held fixed.*

Lemma 1 (Stable transport). *If $B_i B_i^\top + Q_i \preceq I$ and $\delta_i = \lambda_{\min}(Q_i)$, then*

$$\mathrm{Cov}(U_i) \preceq I, \qquad \|B_{L-1} \cdots B_0\|_{\mathrm{op}} \leq \prod_{i=0}^{L-1} \sqrt{1-\delta_i}. \quad (7)$$

Proof. Independence gives $\mathrm{Cov}(U_{i+1}) = B_i \mathrm{Cov}(U_i) B_i^\top + Q_i$, and the covariance bound follows by induction. Moreover $B_i B_i^\top \preceq I - Q_i \preceq (1-\delta_i)I$, so $\|B_i\|_{\mathrm{op}} \leq \sqrt{1-\delta_i}$, and submultiplicativity gives the product bound. Unrolling the recursion writes U_L as $(B_{L-1} \cdots B_0)U_0$ plus a sum of transformed innovations; the matrix product is the coefficient of the earlier state. □

Taking $\delta_i \geq 1-\rho^2$ proves the first claim of Theorem 1. For a direct read $S = \rho\big(aU_0 + \sum_{j \neq 0} a_j U_j\big) + \sqrt{1-\rho^2}\,\eta$, varying U_0 with the other records fixed gives the coefficient $\rho a I$, which exceeds the chain bound when $a > \rho^{L-1}$. This is a partial, direct-edge coefficient; correlated records and feedback through later units can change the total dependence.

Keeping a link costs innovation. If a chain keeps a coefficient of at least $\kappa \in (0, 1]$ over L steps, then $\kappa^2 \le \prod_i (1 - \delta_i)$ by Equation (7), and $\log(1 - x) \le -x$ gives

$$\frac{1}{L} \sum_{i=0}^{L-1} \delta_i \le \frac{-2 \log \kappa}{L}. \tag{8}$$

With a common innovation floor δ, the sharper constraint is $\delta \le 1 - \kappa^{2/L}$, and $L(1 - \kappa^{2/L}) \to -2 \log \kappa$. A chain that keeps a long-range link must therefore add almost no fresh noise per step, which is the opposite of what a changing scene needs. The bound is attained: with $\delta = 1 - \kappa^{2/L}$, $B_i = \sqrt{1 - \delta}\, I$, Gaussian innovations of covariance δI and standard Gaussian U_0, every state is standard Gaussian and the endpoint coefficient is exactly κI; the logarithmic bound is sharp as $L \to \infty$ with κ fixed. A retained direct edge keeps the same coefficient for any number of intervening units. Figure 2 visualizes the two cases.

A.3 Noise correlation and appearance agreement

The response analysis holds the conditioning and the generator fixed. Let $X, X' \in \mathbb{R}^d$ be jointly standard Gaussian. An orthogonal change of coordinates brings their cross-covariance to $\mathrm{diag}(r_1, \ldots, r_d)$ with $0 \le r_i \le 1$; the response functions are rotated together with their inputs. For a multi-index $\alpha = (\alpha_1, \ldots, \alpha_d)$ write $|\alpha| = \sum_i \alpha_i$, $\alpha! = \prod_i \alpha_i!$ and $r^\alpha = \prod_i r_i^{\alpha_i}$ with $0^0 = 1$, and let $H_\alpha(x) = \prod_i \mathrm{He}_{\alpha_i}(x_i)/\sqrt{\alpha!}$ be the normalized probabilists' Hermite basis of $L^2(\gamma_d)$.

Theorem 2 (Agreement under a shared response)**.** *Let $\mathrm{Cov}(X, X') = rI$ with $0 \le r < 1$. For $f \in L^2(\gamma_d; \mathbb{R}^{d_Y})$ with Hermite coefficients f_α, define the degree-q energy $E_q = \sum_{|\alpha|=q} \|f_\alpha\|^2$. Then $\mathbb{E}\|f(X) - f(X')\|^2 = 2 \sum_{q \ge 1} (1 - r^q) E_q$, as in Equation (2), and this quantity decreases strictly in r when f is nonconstant. A one-to-one direct read with $r = \rho$ yields a smaller discrepancy than any chain of Theorem 1 with $L > 1$, when the chain endpoints are standard Gaussian and share the response f.*

Lemma 2 (Cross-response expansion)**.** *For vector-valued $f, g \in L^2(\gamma_d)$ with coefficients f_α, g_α in this basis,*

$$\mathrm{tr}\,\mathrm{Cov}\big(g(X'), f(X)\big) = \sum_{\alpha \ne 0} r^\alpha \langle g_\alpha, f_\alpha \rangle, \tag{9}$$

where the summands can have either sign.

Proof. For one Gaussian pair with correlation r, the moment generating function gives $\mathbb{E}\big[e^{sX - s^2/2} e^{s'X' - s'^2/2}\big] = e^{rss'}$. Expanding $e^{sx - s^2/2} = \sum_{n \ge 0} \mathrm{He}_n(x) s^n / n!$ and matching coefficients yields $\mathbb{E}[\mathrm{He}_n(X) \mathrm{He}_{n'}(X')] = \mathbf{1}_{n=n'}\, n!\, r^n$. The diagonal pairs are mutually independent, so $\mathbb{E}[H_\alpha(X) H_\beta(X')] = \mathbf{1}_{\alpha=\beta}\, r^\alpha$, which proves the formula for polynomials. For general responses, truncate at a total degree; by Cauchy–Schwarz the error is bounded by the L^2 truncation errors times the response norms, and it vanishes by completeness. The series converges absolutely because $\sum_\alpha |r^\alpha \langle g_\alpha, f_\alpha \rangle| \le \big(\sum_\alpha \|g_\alpha\|^2\big)^{1/2} \big(\sum_\alpha \|f_\alpha\|^2\big)^{1/2}$. Removing the degree-zero term removes the means. □

For a shared response and $r_i = r$, Parseval gives $\mathrm{tr}\,\mathrm{Cov} f(X) = \sum_{q \ge 1} E_q$, and substituting Equation (9) into the decomposition proves Equation (2). If $r_2 > r_1 \ge 0$ and f is nonconstant, some $E_q > 0$, so $D(r_1) - D(r_2) = 2 \sum_{q \ge 1} (r_2^q - r_1^q) E_q > 0$. The generator need not be linear for this monotonicity.

A.4 Strict comparison and normalization

Let $T = B_{L-1} \cdots B_0$ and take jointly Gaussian endpoints U_0, U_L with standard marginals. Unrolling the recursion and using independence of the innovations gives $\mathrm{Cov}(U_L, U_0) = T$, whose singular values are the canonical correlations and are at most ρ^L by Theorem 1. With the singular value decomposition $T = P\, \mathrm{diag}(\sigma_i) P'^\top$, rotate the two endpoint inputs by $P^\top$ and $P'^\top$. The rotated responses can differ even when both endpoints originally share f, but orthogonal rotations preserve the Gaussian measure and each total-degree Hermite subspace, so both rotated degree-q

coefficient arrays still have squared norm E_q. For $|\alpha| = q$, Lemma 2 weights the cross term by $\sigma^\alpha \le \rho^{qL}$, and Cauchy–Schwarz gives

$$\left|K_q^{\mathrm{adj}}\right| \le \rho^{qL}\Big(\sum_{|\alpha|=q}\|f_{P,\alpha}\|^2\Big)^{1/2}\Big(\sum_{|\alpha|=q}\|f_{P',\alpha}\|^2\Big)^{1/2} = \rho^{qL}E_q, \tag{10}$$

where $f_{P,\alpha}, f_{P',\alpha}$ are the rotated coefficients. One-to-one direct retrieval has cross-covariance ρI and $K_q^{\mathrm{SV}} = \rho^q E_q$. Both constructions have the same means and marginal variances, so these cancel in the difference, leaving

$$D^{\mathrm{adj}} - D^{\mathrm{SV}} \ge 2\sum_{q\ge 1}\big(\rho^q - \rho^{qL}\big)E_q > 0 \qquad (L > 1,\ 0 < \rho < 1) \tag{11}$$

for a nonconstant response; absolute convergence (Lemma 2) justifies summing the degree-wise bounds. This proves the last claim of Theorem 2, including matrix-valued chains. At $L = 1$ the direct and scalar chain constructions coincide.

Normalization. A transport coefficient becomes a correlation only after dividing by the source scale. For a single fixed read of an independent standard entry, $S = \rho a X + \sqrt{1-\rho^2}\,\eta$ has covariance $\tau^2 I$ with $\tau^2 = \rho^2 a^2 + 1 - \rho^2$, and $\mathrm{Cov}(X, S/\tau) = (\rho a/\tau)I$. A chain endpoint with the same variance has canonical correlation at most $\min(1, \rho^L/\tau)$, and the equal-variance comparison uses these correlations with the response $x \mapsto \psi(\Phi_\theta(\tau x; C))$. With several correlated entries, the source covariance need not be a multiple of the identity, and whitening gives the corresponding coordinates when the covariance permits it.

First-degree approximation. The first Hermite degree gives a controlled approximation,

$$0 \le D(0) - D(r) - 2rE_1 = 2\sum_{q\ge 2} r^q E_q \le 2r^2 \sum_{q\ge 2} E_q. \tag{12}$$

For a Gaussian Sobolev response, integration by parts gives $f_{e_i} = \mathbb{E}[\partial_i f(X)]$ and hence $E_1 = \sum_i \|\mathbb{E}[\partial_i f(X)]\|^2$. The derivative formula needs Sobolev regularity, whereas the full identity needs only square integrability; a first-order approximation is informative only when the higher-degree energy is controlled.

A.5 Conditions under which the conclusions fail

Different responses. If the two units respond to the noise differently, correlation can hurt. For scalar standard Gaussians with historical response $f(x) = x$ and current response $g(x') = -x'$, the discrepancy is $2 + 2r$, which grows with r. A constant response gains nothing at any correlation, and if the conditioning determines the historical readout, $V_h = 0$ and the IID floor is zero.

Chain assumptions. A chain can keep a direction indefinitely if its innovation covariance vanishes in that direction, which violates the uniform innovation floor. The update $U_{i+1} = U_i + \epsilon_{i+1}$ keeps its coefficient but lets the variance grow, which violates the stability condition. A state that keeps adding untouched coordinates of history lies outside the fixed-dimensional comparison.

Data-dependent attention. Real attention weights depend on the generated history. Conditioning on their realized values can select earlier innovations and destroy the independent Gaussian law assumed above: if a read weight is $a = \mathbf{1}\{X_1 > 0\}$ for standard Gaussian X, fixing the observed weight at one truncates X_1 to a half-normal law. Freezing an observed weight therefore differs from choosing a weight independently of the source before sampling. A deterministic sampler must also include every random draw of an otherwise stochastic solver, and the implemented finite-precision noise differs from ideal continuous Gaussians. The decomposition of Section A.1 holds for the implemented finite law, while the Hermite analysis assumes continuous Gaussian sources. These distinctions motivate the natural-condition bounds and the controlled replay in Appendix E.

B Implementation

This appendix specifies how SourceValue is trained and run (Section B.1), shows what the two noise components look like (Section B.2), describes the interface of each backbone (Section B.3), and lists the data and optimization shared by all matched models (Section B.4).

B.1 TRAINING AND GENERATION PROCEDURE

Algorithm 1 gives the shared source recursion, with Euler generation shown for NOVA, VideoMAR and HMA; its organization of sampling, path construction and optimizer steps follows Algorithm 1 of Chen & Lipman (2024). Here k indexes a unit for which one source noise is built: a whole latent frame for NOVA and HMA, one masked group for VideoMAR, or one causal block for Wan. Within a unit, $\mathcal{G}_k$ lists the backbone's sampling groups (64 spatial groups for NOVA, one group for VideoMAR, HMA and Wan), and $T_{k,g}$ is the backbone's reversible rearrangement of the unit's noise restricted to group g. In the training loop, v_θ collects the backbone's group predictions over the unit, each computed from a view in which the target is hidden; summing squared errors and dividing by $\sum_k N_k p$ gives the per-scalar training loss.

At generation, the velocity is $\tilde{v}_\theta = v_\theta^{\text{uncond}} + w_g\big(v_\theta^{\text{cond}} - v_\theta^{\text{uncond}}\big)$ with the native guidance scale w_g, and both branches start from the same source. For a head with $2p$ output channels, v_θ denotes its first p channels. TARGETHIDDENNATIVEREAD applies the backbone's own query/key normalization, scaling and mask, with only previously released conditions and completed history visible. NATIVECOMMIT runs at the backbone's usual content-commit point, and supplied observations initialize a content-only memory. Training keeps θ fixed within each window and recomputes the differentiable content path for every required view, while the stored noise recursion is detached.

For NOVA, VideoMAR and HMA, Euler integration uses $t_j = 1 - j/J$ and $X^{(j+1)} = X^{(j)} - J^{-1}\tilde{v}_\theta(X^{(j)}, t_j, C)$, starting at $X^{(0)} = S$; the time code passed to the network is $999t$. NOVA's p-channel head predicts velocity directly. VideoMAR and HMA keep their $2p$-channel heads and use the first p channels for velocity, leaving the others outside the loss and the state update. The backbone runs in BF16; attention softmax, the noise read, the noise store and Euler updates run in FP32. Matched interventions share all random draws except the noise component that the comparison varies.

B.2 WHAT THE TWO NOISE COMPONENTS LOOK LIKE

Figure 4 in Section 4.2 shows the effect of keeping the native history mass: for a position with mass $\omega_n > 0$, renormalization turns the read R_n into R_n/ω_n and multiplies its energy by $1/\omega_n^2$.

Figure 8 separates the two terms of Equation (3) on four query scenes. The inherited term ρR_k carries spatial structure aligned with the scene, because the read combines the noise of entries that the query attends to. The fresh term $\sqrt{1-\rho^2}\,\eta_k$ is unstructured and supplies the fine-scale variation visible in the full source. The full source therefore looks like noise at every position while carrying a component that is tied to earlier content.

B.3 BACKBONE INTERFACES

NOVA. We use the public 48-layer, width-1024, 480p model (Deng et al., 2025) at 480×768, 12 fps, guidance scale 5, 64 spatial sampling groups and 25 Euler steps per group. A temporal-attention frame has 360 blocks of 64 source channels, reversibly rearranged to the spatial decoder's 1,440 blocks of 16 channels. One whole-frame source is built after the last temporal read and before spatial decoding. A 60-second run uses 181 latent frames and decodes 721 RGB frames, of which the last endpoint is removed. The model's nonrotary temporal positions are interpolated over all 181 positions throughout the run. Newly released text enters the first unstarted latent update, and earlier text remains available. In training, one shifted causal temporal pass supplies 16 target-hidden reads per window, followed by source construction and spatial losses normalized over target scalars.

VideoMAR. The public 480p model (Yu et al., 2025a) generates 49 RGB frames at 480×768 and 12 fps, i.e., seven latent frames with 64 groups per generated frame and 100 Euler steps per group. The guidance scale increases from 1 to 3 along the native group schedule. The first latent frame is supplied content. For each group of the remaining six frames, the final visual attention reads earlier generated frames and completed groups of the current frame, while the target and unfinished groups stay masked. Training uses $6 \times 64 = 384$ target-hidden views per window, and each group loss is weighted by its target-token count divided by $6 \times 1{,}440$. Inference uses 384 masked reads, one supplied-frame encoding and five generated-frame memory updates, for 390 trunk calls and 38,400 head evaluations.

Algorithm 1 SourceValue: causal adaptation and generation with stored source noise

Require: Backbone θ; teacher windows $\mathcal{D}$; release schedule c; $\rho \in (0,1)$;
native units k, groups $\mathcal{G}_k$, grid maps $T_{k,g}$, Euler counts J_g;
native masks, scale, retention/compression rule; optimizer settings in Section B.4

Ensure: Adapted θ; generated latents $\widehat{Z}_{1:K}$ and memory $\mathcal{M}$

function READSOURCE($\theta, \mathcal{M}, c_k$)
  $(C_k, \{A_k^{(\ell)}\}_{\ell=1}^{H}) \leftarrow$ TARGETHIDDENNATIVEREAD($\theta, \mathcal{M}, c_k$)
  $\mathcal{H} \leftarrow \{j \in \mathcal{M} : j$ has a stored source row$\}$
  if $\mathcal{H} = \varnothing$ **then**
    $\eta_k \sim \mathcal{N}(0, I)$; $S_k \leftarrow \eta_k$
  else
    $\bar{A}_k \leftarrow H^{-1} \sum_{\ell=1}^{H} A_{k,\mathcal{H}}^{(\ell)}$
    $R_k \leftarrow \bar{A}_k S_{\mathcal{H}}$ ▷ native softmax mass kept
    $\eta_k \sim \mathcal{N}(0, I)$ independently of the completed read
    $S_k \leftarrow \rho\, \mathrm{sg}(R_k) + \sqrt{1-\rho^2}\, \eta_k$
  end if
  return (C_k, S_k)
end function
function COMMITPAIR($\mathcal{M}, Z_k, S_k$)
  $(\mathbf{K}_k, \mathbf{V}_k) \leftarrow$ NATIVECOMMIT(Z_k); $\mathcal{M} \leftarrow \mathcal{M} \cup \{(\mathbf{K}_k, \mathbf{V}_k, \mathrm{sg}(S_k))\}$
  Apply the native retention indices to both $\mathbf{V}$ and $S_{\mathcal{H}}$
  Apply each native linear aggregation W as $(\mathbf{V}, S_{\mathcal{H}}) \leftarrow (W\mathbf{V}, WS_{\mathcal{H}})$
  return $\mathcal{M}$
end function
Adaptation
for each optimizer update **do**
  Sample a minibatch $\mathcal{B} \subset \mathcal{D}$; $\mathcal{L}_{\mathcal{B}} \leftarrow 0$
  for each teacher window $(Z_{1:K}, c) \in \mathcal{B}$ **do**
    $\mathcal{M} \leftarrow$ NATIVEPREFIXMEMORY(Z_{observed}, c); $\mathcal{L}_{\mathrm{win}} \leftarrow 0$
    for $k = 1, \ldots, K$ in native causal order **do**
      $(C_k, S_k) \leftarrow$ READSOURCE($\theta, \mathcal{M}, c_k$); $t_k \sim \mathcal{U}[0,1]$
      $X_{k,t_k} \leftarrow (1-t_k) Z_k + t_k S_k$
      $\mathcal{L}_{\mathrm{win}} \leftarrow \mathcal{L}_{\mathrm{win}} + \|v_\theta(X_{k,t_k}, t_k, C_k) - (S_k - Z_k)\|_F^2$
      $\mathcal{M} \leftarrow$ COMMITPAIR($\mathcal{M}, Z_k, S_k$)
    end for
    $\mathcal{L}_{\mathcal{B}} \leftarrow \mathcal{L}_{\mathcal{B}} + \mathcal{L}_{\mathrm{win}} / \sum_{k=1}^{K} N_k p$
  end for
  $\theta \leftarrow$ ADAMW($\theta, \nabla_\theta(\mathcal{L}_{\mathcal{B}} / |\mathcal{B}|)$)
end for
Generation
$\mathcal{M} \leftarrow$ NATIVEPREFIXMEMORY(Z_{observed}, c)
for $k = 1, \ldots, K$ in native causal order **do**
  $(C_k, S_k) \leftarrow$ READSOURCE($\theta, \mathcal{M}, c_k$)
  for $g \in \mathcal{G}_k$ in native sampling order **do**
    $X_g^{(0)} \leftarrow T_{k,g}(S_k)$; obtain the native content condition $C_{k,g}$
    for $j = 0, \ldots, J_g - 1$ **do**
      $t_j \leftarrow 1 - j/J_g$; $X_g^{(j+1)} \leftarrow X_g^{(j)} - J_g^{-1} \tilde{v}_\theta(X_g^{(j)}, t_j, C_{k,g})$
    end for
    $\widehat{Z}_{k,g} \leftarrow X_g^{(J_g)}$; expose this completed group to later native groups
  end for
  $\widehat{Z}_k \leftarrow$ ASSEMBLE($\{\widehat{Z}_{k,g}\}_{g \in \mathcal{G}_k}$)
  $\mathcal{M} \leftarrow$ COMMITPAIR($\mathcal{M}, \widehat{Z}_k, S_k$)
end for
return $(\theta, \widehat{Z}_{1:K}, \mathcal{M})$

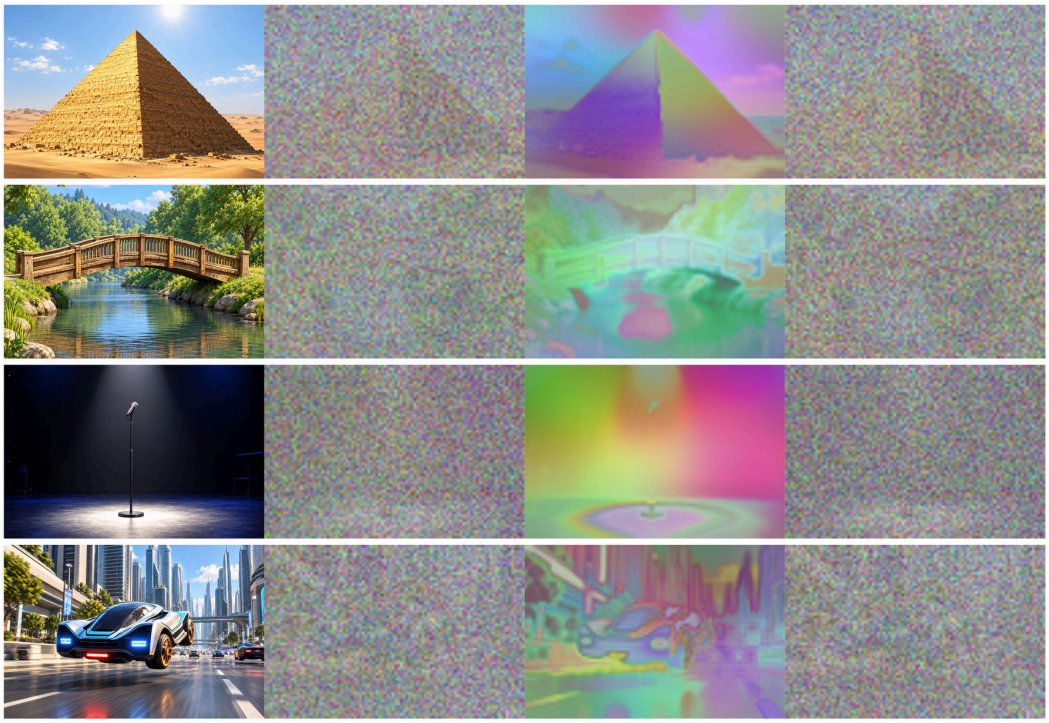


Figure 8: The inherited and fresh parts of the source play different roles. Rows: pyramid, bridge, stage and vehicle. Columns: query scene, full source $S_k = \rho R_k + \sqrt{1-\rho^2}\,\eta_k$, inherited part only ($\rho R_k$, labeled No innovation) and fresh part only ($\sqrt{1-\rho^2}\,\eta_k$, labeled No inherited read). Source vectors are normalized per position, and all twelve maps share one three-component PCA basis; colors encode direction, not energy.

HMA. The public continuous base model (Wang et al., 2025a) uses 256×256 observations at 2 Hz, a 12-frame window, four observed frames, guidance scale 1 and 100 Euler steps for each of eight generated frames. Its single masked group contains 256 blocks of 16 channels. The final temporal attention reads noise along the native same-position temporal edges, $R_{i,n} = \frac{1}{H}\sum_{\ell}\sum_{j<i} A^{(\ell)}_{n,i,j} S_{j,n}$ for frame i and position n, with the inner sum restricted to generated frames. Training uses eight equally weighted views, each exposing earlier teacher frames and masking current and future slots. The native action modulation, query/key normalization and attention scaling are unchanged. An eight-frame prediction requires eight trunk calls and 800 head evaluations.

Wan2.1. The common root is Wan2.1-T2V-14B, run in BF16 at 832×480 and 16 fps. The matched causal interface generates blocks of at most three latent frames with 50 denoising steps. Long runs decode 961 RGB frames and remove the final endpoint, retaining 960 frames (60 seconds); short T2V uses 81 frames. New instructions enter the first unstarted block. Before drawing its noise, fixed zero target placeholders query the last layer's head-averaged attention, with current and future content masked. IID and SourceValue share a 20-frame-equivalent history: the first two frames, the latest nine and nine contiguous equal-time buckets over the remaining history. Buckets average K/V linearly; SourceValue applies the same weights to FP32 source records. Early runs retain only available history, and the pending block adds at most three frames. SourceValue uses $\rho = 0.5$ and the source-consistent input and target of Section 4.3; IID uses independent noise.

B.4 Training data and matched optimization

NOVA, VideoMAR and the Wan comparison use the MiraData v1 330K release and its captions (Ju et al., 2024). Clips are grouped by source video for a deterministic 98/1/1 split into training, development and calibration sets, and matching source identifiers, normalized captions and decoded frames remove any overlap with evaluation material. HMA uses the 442,226 episodes of LanguageTable 0.1.0 (Lynch et al., 2022); the first 200 episodes of the public evaluation split are held out, and the re-

maining 442,026 episodes receive the same grouped split. Windows stay inside episode boundaries. Development data select the learning rate, and calibration data set evaluator procedures.

The four noise rules on NOVA, VideoMAR and HMA use AdamW with $(\beta_1, \beta_2) = (0.9, 0.95)$ and $\epsilon = 10^{-8}$. NOVA and VideoMAR use weight decay 0.02, global batch 16, 20,000 updates and gradient clipping at 1; HMA uses weight decay 0.01, global batch 32, 10,000 updates and clipping at 10. Each model draws 320,000 training windows. Warmup lasts 500 updates and is followed by a constant learning rate, and each target receives one flow-matching time and source sample. Training uses four A100 80GB PCIe GPUs. For each model and noise rule, three 2,000-update runs compare learning rates $\{5 \times 10^{-6}, 10^{-5}, 2 \times 10^{-5}\}$; the lowest development flow-matching loss selects the rate, with ties broken toward the lower rate. Three independent runs with seeds 17, 29 and 43 then start from the public model and are evaluated at the final update. The noise rule defines the regression target, and the same selection procedure is applied within each rule. The Wan IID/SourceValue pair shares the NOVA optimizer, batch size, update count, warmup, learning-rate selection and final-checkpoint protocol. External Wan adaptations use 20,000 total updates; two-stage methods allocate 10,000 updates to each stage, retaining their own training objectives and teacher calls. MemRoPE reuses the adapted LongLive checkpoint, and Video-T1 reuses the Wan IID generator; the inherited adaptation cost is included in Table 14.

C Evaluation protocol

This appendix defines the benchmarks and metrics (Section C.1), the identity judgments and the uncertainty calculation (Section C.2), and the full distribution of answers at the primary endpoint together with examples (Section C.3).

C.1 Benchmarks and metrics

NarraStream-Bench. We use all 324 six-instruction scripts and their entity-occurrence groups (Liu et al., 2026). Conditional long-range consistency (CLC) measures the DINO appearance consistency between an entity's first occurrence and its later occurrences, aggregated with the benchmark's target-segment weights and shrinkage rule. Entity grounding (EG) measures visual agreement with the entities and attributes extracted from each segment's text: Qwen3.5-27B (Qwen Team, 2026) extracts the requirements, which are shared across methods, and Qwen3-VL-30B-A3B-Instruct (Bai et al., 2025a) answers visual questions on five frames per segment with deterministic decoding. VLM Alignment scores narrative order and state consistency on a 1–100 scale; the whole-video score weights segment scores by 0.8 and a full-video judgment by 0.2, and Figure 15 uses the segment scores. Quality averages subject and background consistency, flickering, motion smoothness and video–text similarity. Each metric uses its applicable script population, and generation failures stay in it.

NarrLV. The benchmark has 360 narratives: three factors (scene attributes, target attributes and target actions), six temporal narrative atom counts (TNA1–6) and 20 inputs per factor and count (Feng et al., 2026). Each point pooled across factors therefore uses 60 narratives, and coherence, defined only with at least two atoms, uses the 300 narratives of TNA2–6. Qwen2.5-VL-72B-Instruct (Bai et al., 2025b) gives five answers to each of the 4,978 benchmark questions, sampling video at 2 fps; initial-content questions read the first frame, and anomaly, coverage and transition questions read the full output. Let q_F, q_C, q_H be the fidelity, coverage and coherence question scores, with anomaly answers reversed to the benchmark's quality direction; let QA be 0.2 times the first-frame Q-Align score (Wu et al., 2023) and CR the fraction of coverage questions scoring above 0.2. The benchmark combines them as

$$\text{Fidelity} = \tfrac{1}{2}\big(q_F + \text{QA}\big), \qquad \text{Coverage} = \tfrac{1}{2}\big(q_C + \text{QA}\big), \qquad \text{Coherence} = \tfrac{1}{3}\big(q_H + \text{QA} + \text{CR}\big).$$

Scene or target category scores below 0.3 set the associated layout term to zero, and a transition receives zero coherence if either of its units has coverage below 0.3. Within each factor, fidelity and coverage average the six TNA counts and coherence the five counts with transitions; the mean weights the three metrics and three factors equally. Figure 6 pools factors at each count, and Figure 10 shows each factor separately.

Short-video benchmarks. These tests use each model's native duration, including NOVA's 61 RGB frames; the common-Wan comparison uses 81 frames. VBench uses the original metric populations

and 25 samples per text input, and its image-to-video extension contains 355 images and 763 camera instructions (Huang et al., 2024; 2026). TC-Bench evaluates 150 text-to-video descriptions and 120 image-to-video inputs with five videos per input and repeat (Feng et al., 2025). LLaVA-v1.6-Mistral-7B (Liu et al., 2024) judges assertions from 16 uniformly sampled frames with greedy decoding, and ambiguous answers count as unsatisfied. Assertion accuracy averages binary answers within each video and then across videos; transition completion (TCR) averages whether all critical transition-completion and transition-object assertions hold. The text-to-video assertion-count groups 4, 6, 7, 9, 10 and 11 contain 10, 21, 52, 31, 16 and 20 descriptions. The image-to-video TC-Score combines assertion accuracy with the benchmark's adjacent-frame CLIP term (Radford et al., 2021). StoryEval has 423 descriptions with 1,164 events, of which 133 descriptions contain two events, 262 three and 28 four (Wang et al., 2025b). An event succeeds when two independent LLaVA-OneVision-Qwen2-72B (Li et al., 2024a) judgments both succeed, each sampling $\max(4, \min(32, \lfloor F/4 \rfloor))$ frames from an F-frame output. For a required fraction u, panel (c) of Figure 20 reports the share of descriptions whose fraction of completed events is at least u, weighting descriptions equally. T2V-CompBench V2 evaluates seven categories of 200 descriptions (Sun et al., 2025). For Wan, StoryEval uses one output per description and repeat; TC-TCR uses 750 outputs per repeat, and TC-Score averages verified assertions within each video before averaging videos. VBench total uses the official normalization of all 16 dimensions and the 4:1 quality/semantic weighting; dynamic degree has weight 0.5 and uses 72 inputs with 25 outputs each.

Action-conditioned prediction. For HMA, nonoverlapping 12-frame windows start at multiples of 12 within each held-out episode, and the 200 episodes supply 198 full windows. PSNR and action sensitivity use the last predicted frame; SSIM (Wang et al., 2004) and LPIPS (Zhang et al., 2018) average the eight future frames. FID (Heusel et al., 2017) uses all predicted frames, and FVD (Unterthiner et al., 2018) uses the four observed and eight predicted frames together to meet its minimum clip length; distribution metrics use the complete population. Action sensitivity compares the PSNR under the supplied actions with that under five Gaussian action sequences in the model's action normalization.

C.2 Identity judgments and paired uncertainty

Return applicability is decided from the input text alone: a departure from a room, scene or activity area followed by a return qualifies. This selects 75 NarraStream scripts and leaves 249 others. Reviewers see anonymized complete outputs, the original text and the relevant earlier appearances, with randomized left/right order and replay available. Five judgments per script separately assess whether the return happens and whether the returning character is the same individual; staying continuously visible fails the leave-and-return requirement. Missing entities, failed returns and unjudgeable outputs remain in the denominator. The donor comparison of Table 5 uses 321 text-defined later recurrences and six exclusive outcome categories.

For three complete repeat scores x_1, x_2, x_3 we compute

$$\bar{x} = \frac{1}{3}\sum_{i=1}^{3} x_i, \qquad \hat{\sigma}_x^2 = \frac{1}{2}\sum_{i=1}^{3}(x_i - \bar{x})^2, \qquad \mathrm{CI}_{95} = \bar{x} \pm 4.30265273\,\hat{\sigma}_x/\sqrt{3}. \tag{13}$$

Effects apply this formula to the within-repeat differences $x_i - y_i$, and interactions to within-repeat differences of differences. Inputs, frames and judgments stay grouped within each repeat. These pointwise intervals describe training variability for matched models and trained Wan adaptations, or generation variability for fixed native external systems, under an approximate normal-mean model. Contrasts are computed before rounding, and tables omit interval half-widths. Wan tables average seeds 17, 29 and 43; MemRoPE/LongLive and Video-T1/IID share their respective parent checkpoints within each seed. The bands in Figure 10 and the error bars in Figure 15 use this paired calculation; other figures show means or empirical distributions.

C.3 Judgments at the primary endpoint

Table 7 lists the NOVA answer categories at the 75-script endpoint of Table 1; each repeat has $75 \times 5 = 375$ judgments per model, and "No" includes unsuccessful returns. Over the three equally sized repeats, confirmation rates are 75.38% for SourceValue and 50.22% for IID, with 6.76% and 9.07% unjudgeable answers.

To measure how far unjudgeable answers can move the comparison, we resolve them in the two extreme ways. Resolving every unjudgeable SourceValue answer against confirmation and every unjudgeable IID answer in favor of it gives an observed difference of 16.09 points; the reverse assignment gives 31.91 points. The difference therefore stays between 16.09 and 31.91 points under any resolution of ambiguous answers, and sampling uncertainty is reported separately by the paired interval in Section 5.1.

Table 7: Same-instance judgment outcomes for the 75-script return endpoint. Mean percentages across three repeats; each repeat has 375 assigned judgments per model. "No" includes unsuccessful returns. Rounding may prevent totals from summing to 100.

Method	Confirmed	No	Unjudg.
NOVA d48w1024 (0.646B)			
IID	50.2	40.7	9.1
SourceValue	**75.4**	**17.9**	**6.8**

Figure 9 shows how judges compare a return: each model is judged against its own earlier character. The two models start the same script from different people, so a return can only be judged within a model, and a single script shows one outcome of a random process; the endpoint therefore reports rates over all 75 scripts and three repeats. Figure 5 in Section 5.1 shows SourceValue returns in eight settings.



Figure 9: Each model is judged against its own earlier character. IID and SourceValue follow one script: Clara leaves, equipment is assembled, and Clara returns. Lower views pair each model's first appearance at 0 s with its returns at 50 s and 59 s. The two models start from different people, so comparisons are made within a row. Table 7 gives the answer distribution over all 75 scripts.

D Additional results and intervention details

This appendix follows the order of Section 5. It first places the matched models among complete systems (Section D.1), then details the interventions on correspondence (Section D.2), on the cached record (Sections D.3 and D.4), and the measurements of variation, short tasks and cost (Sections D.5–D.7).

D.1 Complete-system comparisons

Table 8 places matched NOVA and Wan models among complete systems on NarrLV and four short text-to-video benchmarks, together with their sampling budgets, and Table 9 gives the component scores. The native block retains external checkpoints, training and sampling budgets; the common-Wan block fixes the backbone root and evaluation protocol. The IID/SourceValue pairs isolate the noise rule. In the native block, Ring Forcing uses its Wan2.2 A14B memory model (Xue et al., 2026). MemRoPE uses evolving memory on LongLive with Wan 1.3B (Kim et al., 2026), and LongLive uses its four-step generator (Yang et al., 2026). Stable Video Infinity uses text-to-video error recycling (Li et al., 2026), and Self Forcing uses four-step chunkwise generation (Huang et al., 2025). World-R1-Large combines geometric noise transport with a reinforcement-trained Wan2.1-14B model (Wang et al., 2026a). Video-T1 runs NOVA with Tree-of-Frames search and candidate

verification (Liu et al., 2025). World-R1 and Video-T1 receive the complete text at once and therefore appear only on NarrLV and the short full-text benchmarks; the NarraStream comparison uses systems that accept instructions as they arrive and keep previously emitted video.

Common Wan comparison. The added blocks start from Wan2.1-T2V-14B and share evaluation inputs, evaluators, resolution and frame rate. Ring Forcing retains ring training and history compression on the dense 14B root; MemRoPE uses the adapted LongLive model; LongLive and Self Forcing retain causal distillation and long-video training; SVI retains error recycling; and World-R1 retains geometric transport and its reinforcement objective. Video-T1 searches over Wan blocks with four candidates per decision, using a calibration-fixed verifier and candidate-index tie breaking, and keeps the selected prefix. MemRoPE, LongLive and Self Forcing use four denoising steps; the other Wan rows use 50. Candidate count is one except for Video-T1. All Wan inference uses one A100 80GB. External methods retain their own history policies and training objectives; only Wan IID and SourceValue share the causal cache, optimization and sampling budgets that isolate the source rule. World-R1 and Video-T1 retain the full-story-only boundary in Table 1.

Table 8: Five-benchmark results and budgets (0–100; higher is better). NarrLV: 60-second videos; the four short-T2V benchmarks retain their full input populations, with 81-frame outputs for Wan. VB: VBench total; TC: completion and TC-Score; Events: event completion; Stable/Dyn.: attribute binding. NarrLV time is in minutes on the listed A100 80GB GPU count: sampling for native rows, end-to-end inference for Wan. Native Video-T1 uses NOVA Tree-of-Frames; its Wan version uses chunk search. Wan steps/candidates are specified in Appendix D.1; only IID/SourceValue pairs isolate the noise rule.

Method	NarrLV				VB	TC-Bench		Story	CompBench		Budget	
	Fid.	Cov.	Coh.	Mean	Total	TCR	Score	Events	Stable	Dyn.	GPUs	Min.
Complete-system comparison (native checkpoints, training and sampling budgets)												
Ring Forcing (Xue et al., 2026)	**83.5**	**74.7**	**74.2**	**77.4**	**84.2**	49.6	88.8	32.5	**71.8**	**6.0**	4	105.64
MemRoPE (Kim et al., 2026)	81.6	70.6	69.5	73.9	83.4	44.7	87.8	27.5	67.5	3.7	1	2.18
LongLive (Yang et al., 2026)	79.9	65.6	60.8	68.8	82.9	39.4	86.1	23.5	65.3	3.5	1	2.13
World-R1-Large (Wang et al., 2026a)	83.2	73.0	70.7	75.6	84.0	**51.7**	**89.9**	**35.0**	70.1	5.8	4	132.40
Stable Video Infinity (Li et al., 2026)	82.1	70.8	68.7	73.9	83.7	45.4	87.9	29.5	68.9	5.6	2	66.56
Self Forcing (Huang et al., 2025)	78.1	62.1	54.9	65.0	83.3	44.0	87.3	26.5	67.1	2.8	1	**2.04**
Video-T1 (NOVA ToF) (Liu et al., 2025)	82.7	73.4	71.4	75.9	83.3	48.0	89.1	32.8	70.9	4.8	4	145.20
Matched NOVA d48w1024–osp480 (0.646B) (Deng et al., 2025)												
IID	**80.2**	67.1	63.4	70.2	**82.8**	39.5	86.2	**25.1**	**64.7**	**3.0**	1	**22.90**
SourceValue	**80.2**	**69.3**	**68.0**	**72.5**	82.7	**39.6**	**86.3**	24.9	64.4	2.8	1	23.01
Common Wan2.1-T2V-14B (Team Wan et al., 2025) (480p, 16 fps; method-specific budgets)												
Ring Forcing (Xue et al., 2026)	84.28	74.85	75.08	78.07	85.33	54.36	90.98	38.26	73.69	7.17	1	188.48
MemRoPE (Kim et al., 2026)	83.26	72.53	72.29	76.03	84.73	48.13	88.67	34.82	70.61	5.46	1	24.49
LongLive (Yang et al., 2026)	82.58	69.56	66.03	72.73	84.47	45.47	87.25	32.82	69.10	5.08	1	23.76
World-R1-Large (Wang et al., 2026a)	**85.00**	75.87	73.17	78.01	**85.69**	56.49	92.01	40.12	72.39	**8.11**	1	240.42
Stable Video Infinity (Li et al., 2026)	83.99	73.45	71.05	76.16	85.30	56.00	91.36	37.08	71.96	7.44	1	166.25
Self Forcing (Huang et al., 2025)	81.21	67.35	62.30	70.29	84.32	47.64	88.20	34.11	69.58	4.51	1	**21.52**
Video-T1 (chunk tree) (Liu et al., 2025)	84.95	**77.60**	76.27	**79.61**	85.65	**58.27**	**92.77**	**41.61**	**74.20**	7.05	1	718.04
IID	84.19	73.07	69.91	75.72	85.51	51.29	90.27	35.77	72.49	6.56	1	180.22
SourceValue	84.15	75.42	**79.17**	79.58	85.57	51.47	90.28	35.68	72.66	6.42	1	183.24

On NOVA, SourceValue's NarrLV coherence of 68.0 lies above LongLive and Self Forcing and below the other five complete systems, which use larger backbones or, in the case of Video-T1, test-time search on NOVA; it takes 23.01 minutes per video on one GPU. On the short benchmarks the matched NOVA models differ by at most 0.3 points on every summary score, for example TC-Score 86.3 against 86.2 and StoryEval event completion 24.9 against 25.1. Short videos rarely contain an absence long enough for a character to be redrawn, so the matched comparison changes little there, while the long-video comparison in Table 1 changes substantially.

On Wan, SourceValue's NarrLV coherence is 79.17, against 69.91 for IID and 76.27 for Video-T1. Its three-metric mean is 79.58, close to Video-T1's 79.61. VBench total is 85.57 against IID's 85.51; TC-TCR is 51.47 against 51.29, TC-Score 90.28 against 90.27, and StoryEval 35.68 against 35.77. Table 9 also retains the mixed short-task ordering: dynamic degree is 55.52 against 55.93 and dynamic binding 6.42 against 6.56. The common-backbone result therefore supports higher long-range identity and narrative coherence with similar short-task scores, while four-step methods retain their speed advantage.

Table 9: Short-T2V breakdowns for Table 8 (0–100; higher is better). VBench: subject/background consistency (SC/BC), motion smoothness (MS), dynamic degree (DD), aesthetic/imaging quality (AQ/IQ). CompBench: consistent/dynamic attributes, spatial relations, motion/action binding, interactions and numeracy. Native and common-Wan blocks use the same methods and protocols as Table 8. VBench total uses all 16 official dimensions; the six shown here are not averaged to obtain it.

Method	VBench T2V						T2V-CompBench						
	SC	BC	MS	DD	AQ	IQ	Stable	Dyn.	Spatial	Mot.	Action	Inter.	Num.
Complete systems (native short-video tasks and sampling budgets)													
Ring Forcing (Xue et al., 2026)	97.8	**97.9**	**99.2**	44.5	**68.2**	**72.6**	**71.8**	**6.0**	58.4	30.8	60.2	**66.4**	44.1
MemRoPE (Kim et al., 2026)	97.5	97.6	99.1	49.5	66.8	70.9	67.5	3.7	55.3	29.8	57.0	63.1	42.0
LongLive (Yang et al., 2026)	97.3	97.6	99.1	48.4	66.4	70.4	65.3	3.5	51.6	27.7	54.9	61.6	38.9
World-R1-Large (Wang et al., 2026a)	**97.9**	96.9	98.6	**60.2**	67.4	68.5	70.1	5.8	**63.6**	**34.8**	61.0	66.0	**45.4**
Stable Video Infinity (Li et al., 2026)	96.9	97.3	98.8	57.0	67.8	70.6	68.9	5.6	57.9	33.2	60.5	65.4	41.5
Self Forcing (Huang et al., 2025)	96.9	97.3	98.8	56.4	64.7	68.4	67.1	2.8	54.4	28.9	57.9	62.7	42.5
Video-T1 (NOVA ToF) (Liu et al., 2025)	96.8	97.5	98.8	48.2	65.2	70.2	70.9	4.8	56.4	28.9	**61.2**	66.0	43.9
Matched NOVA d48w1024–osp480 (0.646B) (Deng et al., 2025)													
IID	**96.9**	97.4	**98.9**	**53.1**	**64.7**	69.8	**64.7**	**3.0**	51.7	**28.7**	**55.4**	61.5	38.8
SourceValue	96.6	**97.5**	**98.9**	51.8	**64.7**	**70.0**	64.4	2.8	**52.1**	28.3	55.3	**61.9**	**39.0**
Common Wan2.1-T2V-14B (Team Wan et al., 2025) (480p, 16 fps; method-specific budgets)													
Ring Forcing (Xue et al., 2026)	98.10	**98.09**	**99.28**	50.48	**71.21**	**74.12**	73.69	7.17	61.13	34.05	62.74	68.41	46.15
MemRoPE (Kim et al., 2026)	97.92	97.85	99.22	52.87	69.07	73.21	70.61	5.46	58.96	32.78	60.72	65.44	43.95
LongLive (Yang et al., 2026)	97.61	97.83	99.09	52.83	68.27	72.75	69.10	5.08	55.31	30.68	59.05	64.29	41.47
World-R1-Large (Wang et al., 2026a)	97.90	97.63	98.92	**63.35**	70.63	72.34	72.39	**8.11**	**66.66**	**37.82**	65.15	69.13	**48.64**
Stable Video Infinity (Li et al., 2026)	97.59	97.75	98.96	58.76	70.18	73.06	71.96	7.44	61.60	36.63	64.17	67.91	45.11
Self Forcing (Huang et al., 2025)	97.28	97.56	98.90	56.87	68.24	71.47	69.58	4.51	58.46	32.14	61.20	65.15	44.43
Video-T1 (chunk tree) (Liu et al., 2025)	97.77	97.96	99.01	55.24	70.51	73.61	**74.20**	7.05	63.28	35.28	**66.20**	**69.53**	47.18
IID	97.98	97.86	99.12	55.93	70.60	73.70	72.49	6.56	61.13	35.14	63.07	67.10	45.90
SourceValue	**98.13**	98.00	99.11	55.52	70.86	73.74	72.66	6.42	61.17	34.88	62.93	67.50	45.57

D.2 Noise rules and the address permutation

Noise rules. Table 3 compares complete noise rules on NOVA. Frame-to-frame (adjacent) inheritance reads the previous completed latent frame at matching spatial coordinates. Shared-field (global) inheritance reads the actual source of the first generated frame in the current generation window, and this field persists across online text changes. Both use $S_k = 0.5R_k + \sqrt{0.75}\,\eta_k$ from the second generated frame on; the first generated frame follows SourceValue, including groups already completed within it. Supplied image or video prefixes contribute content only. Each rule is fine-tuned under its own source distribution with the procedure of Section B.4.

Figure 10 resolves the NarrLV coherence of these rules by narrative factor and TNA count. The SourceValue read gains 7 to 9 coherence points over the fine-tuned IID model at six TNAs in all three factors, and its gain grows steadily with the number of TNAs. Frame-to-frame and shared-field noise stay within a few points of zero at every count, and the SourceValue read on IID-trained weights lies between the two groups. Longer stories contain more departures and returns, so the rule that retrieves the matching earlier noise gains most where stories are longest.

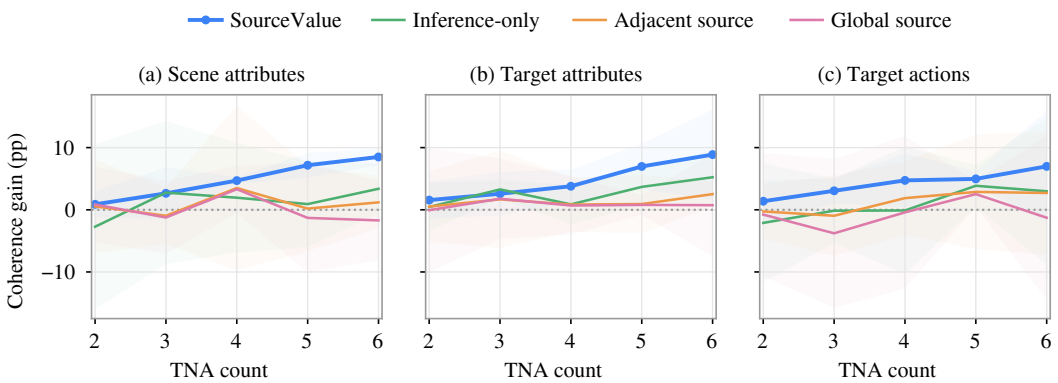


Figure 10: Only attention-addressed noise gains more coherence as stories get longer. NarrLV coherence gain over the fine-tuned IID model per factor (20 narratives per factor and count). Adjacent (frame-to-frame) and global (shared-field) noise are fine-tuned under their own rules; inference-only uses IID-trained weights. Bands: pointwise 95% t-intervals over three seeds.

Address permutation. For the one-update control, let P_π uniformly permute whole rows of the stored noise, compute $R^\pi = \bar{A} P_\pi S_{\mathcal{H}}$, and for each current position n use

$$R_n^\star = \begin{cases} \|R_n\| \, R_n^\pi / \|R_n^\pi\|, & \|R_n^\pi\| > 0, \\ R_n, & \|R_n^\pi\| = 0. \end{cases} \tag{14}$$

Norm matching keeps the magnitude of the inherited read and changes only its direction. The branches share fresh noise, content attention and permanent memory indexing, and every sampled permutation is kept, including identity permutations. The intervention happens at NOVA's first update that consumes the second instruction and at the second generated frame for VideoMAR and HMA. Innovation deletion (No innovation in Table 2) happens at the same update and removes the fresh term while keeping the scale of the inherited term; an empty-history source becomes zero under this intervention. All later updates follow the normal rule. Figure 11 shows the effect of Equation (14) on the conditional mean of the source: the retrieved noise follows the layout of the two readers and the table, while the permuted noise keeps the same magnitude and loses this layout.

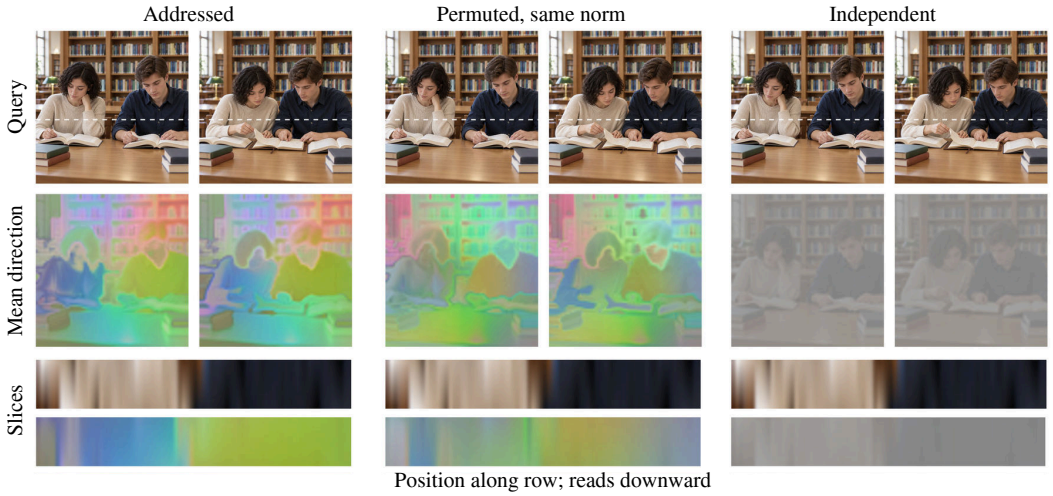


Figure 11: Permuting the stored noise changes its direction at a fixed read norm. Retrieved (addressed), norm-matched permuted and IID conditions on the same query frames and scanline. Middle: conditional mean direction of the source in a shared PCA-to-RGB basis. Bottom: RGB appearance and source direction along the scanline, with reads running downward. Equation (14) holds content attention and fresh noise fixed; the IID conditional mean is zero and appears neutral.

D.3 History retention and record exchange

Retention. Table 4 intervenes at the middle instruction transition, about 30 seconds into the video. In one-based NOVA indexing, update 1 emits RGB frame 0 and update $k \geq 2$ emits frames $4k - 7$ to $4k - 4$. At update 92, memory first receives the completed content and noise of update 91; deletion then removes all older entries and keeps that pair, and the branch occurs at $361/12$ seconds. Relevant-history sets contain the earlier segments showing entities required by the current text. They coincide with the all-older deletion set for 322 of 324 scripts, including all 75 return scripts. The CLC retention-by-noise interaction is 1.03 points (95% CI 0.56–1.50), and the same-instance interaction is 4.98 points (-3.46–13.42).

For Figure 12, let $\mathcal{G}$ contain the historical entries of the text-identified entities and $A_{n,j}^{(\ell)}$ the native pre-sampling attention. The plotted attention mass is $\mathrm{AM} = \frac{1}{HN} \sum_{\ell=1}^{H} \sum_{n=1}^{N} \sum_{j \in \mathcal{G}} A_{n,j}^{(\ell)}$, which keeps the original softmax mass; deleted entries contribute zero, while the latest retained entry can still contribute. The 324 scripts are ordered by their three-seed mean AM with history kept, and every panel uses this order. The source gain is the within-script difference in CLC between the SourceValue update and an IID update, averaged over seeds at the same checkpoint and history condition. Deletion lowers the attention mass on the relevant entities from about 0.2–0.4 to about 0.04 and the source gain from about 1.4 to about 0.3 points on almost every script.

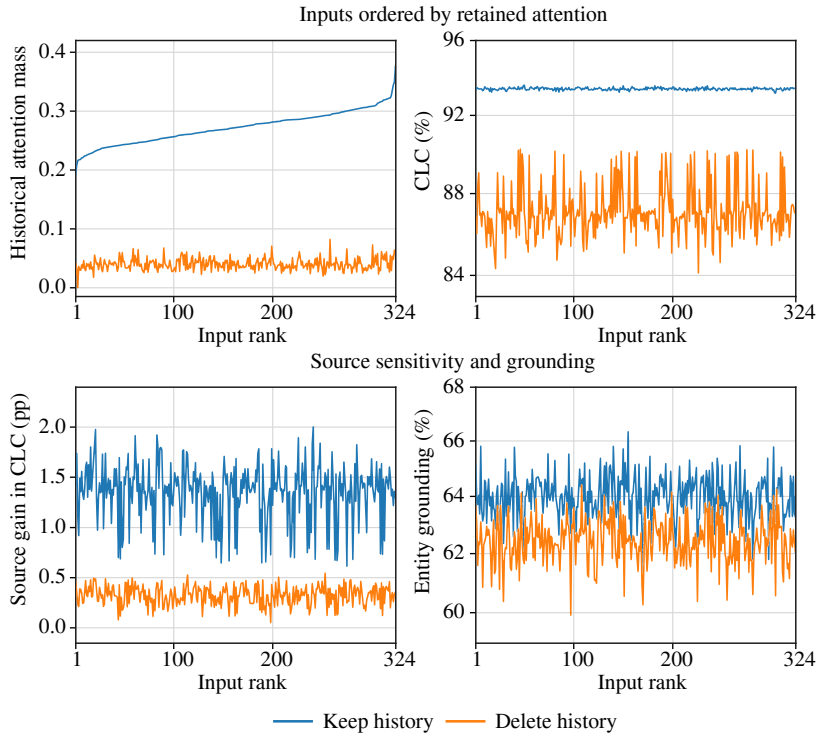


Figure 12: The benefit follows the availability of the earlier record. All 324 NarraStream scripts, ordered by attention mass with history kept; each curve averages three seeds. Keeping history and deleting all older entries share the latest completed entry. Panels: attention mass on text-relevant history, CLC, the SourceValue-minus-IID-update CLC gain, and entity grounding. CLC and grounding use the SourceValue branch.

Record exchange. Table 5 exchanges records at the text-defined return. A normal branch and a branch with permuted noise start from a shared pre-return state and generate the return segment; their actual records supply the original content and noise (M_1, T_1) and the donor content and noise (M_0, T_0). The four writes are (M_1, T_1), (M_1, T_0), (M_0, T_1) and (M_0, T_0). All branches keep the released video prefix and the immediate clean input, so only the persistent memory differs, and each noise record includes both its inherited and fresh parts. Later identity judgments use six categories: original, donor, other, indistinguishable, absent and unjudgeable. Figure 13 shows the spatial response of the source to each replacement at a shared query.

D.4 Persistent write

This intervention separates immediate access to a record from its persistence. We cross the norm-matched permutation with writing the sampled content and noise into persistent memory. The selected unit is update 32, the first to begin after ten seconds. Each noise condition generates this unit once and then forks into a write branch and a skip branch. Both branches use its content and actual noise at the next local read; the skip branch omits the subsequent persistent key, value and noise write. Older records, text, absolute positions and future fresh noise are shared at the fork, and later writes resume normally. A later divergence therefore traces the retained record. Table 10 reports the four conditions.

Table 10: Source correspondence × one persistent write (NOVA). The four evaluated cells report mean raw scores (0–100; higher is better). Return/Inst. use the fixed 75-script endpoint. Paired contrasts and their uncertainty are reported in the text rather than repeated as derived table rows.

Source	Commit	NarraStream			NLV
		CLC	Return	Inst.	Coh.
Addressed	Write	**93.40**	**80.7**	**75.4**	**68.02**
Permuted	Write	90.69	80.0	60.2	64.59
Addressed	Skip	92.09	77.9	68.4	67.05
Permuted	Skip	91.41	80.2	64.2	66.40

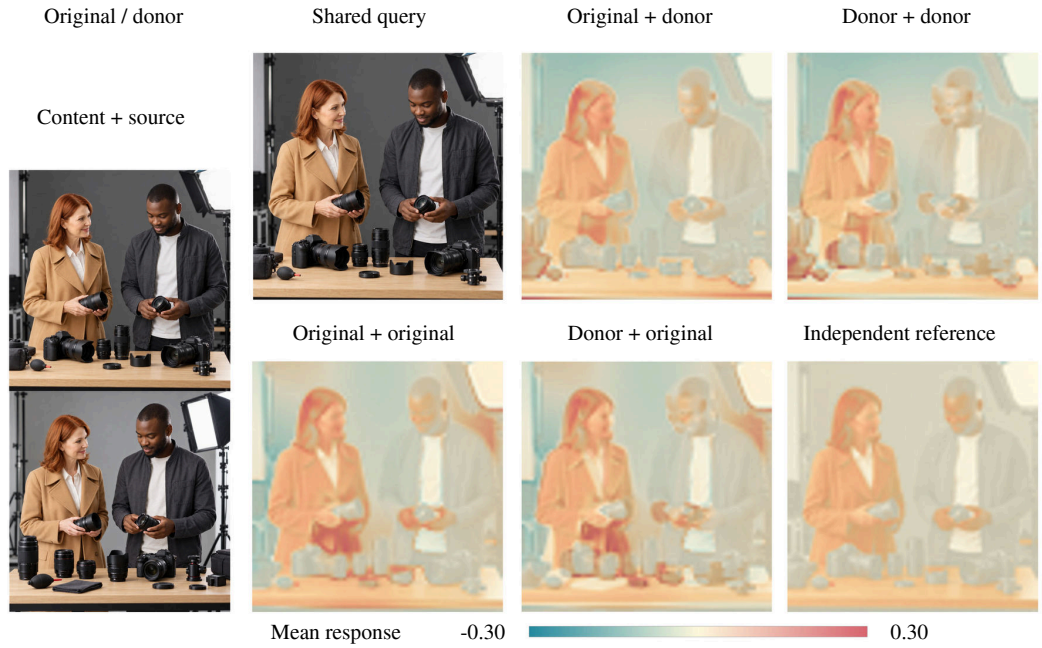


Figure 13: Exchanging content and noise records at a shared query. Left: original and donor records. Each label names the content first and the noise second; original + original is the reference, and the other cells exchange one record or both. The IID response is at lower right. All maps share one signed mean-response scale (-0.30 to 0.30); the query and fresh noise are fixed. Table 5 reports the identity outcomes.

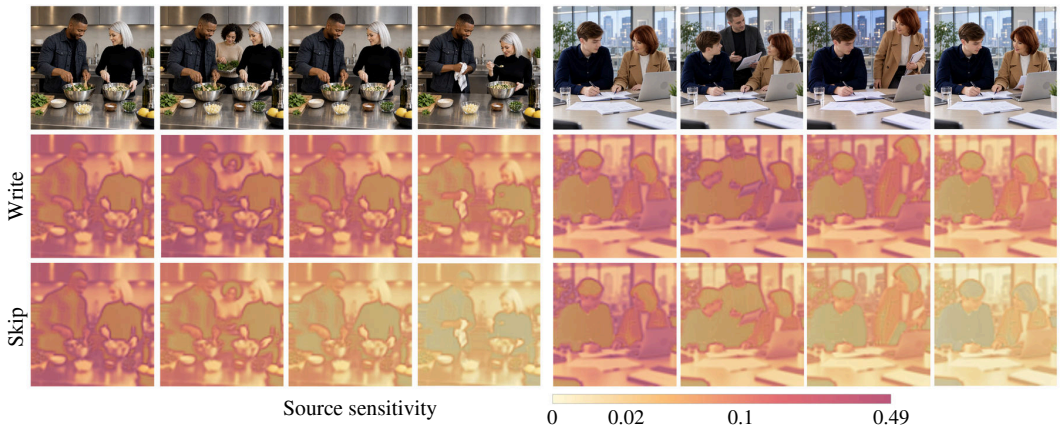


Figure 14: A persistent write keeps later reads sensitive to the record. Two examples with four reads each, left to right. Query frames sit above paired Write and Skip sensitivity maps on a shared 0–0.49 scale. The first read precedes the effect of the write; later reads keep a stronger response after writing. Compare Write and Skip within a column.

For a metric Q, the interaction is $(Q_{\mathrm{ret,write}} - Q_{\mathrm{perm,write}}) - (Q_{\mathrm{ret,skip}} - Q_{\mathrm{perm,skip}})$, where ret denotes the retrieved (addressed) noise. It is 2.02 points for NarraStream CLC (95% CI 1.64–2.41) and 2.78 for NarrLV coherence (2.31–3.26). Writing helps the retrieved branch, and skipping the permuted record partly undoes its loss, so the committed record carries the correspondence effect into later generation. Figure 14 follows the written record through later reads, and Figure 15 shows the population-level effect. Its left panel compares each condition with the separately trained IID model within seed, averaging the three NarrLV factors at each TNA count; its right panel plots the retrieved-minus-permuted segment alignment on all 324 NarraStream scripts, which is zero before the intervention. The complete-output interactions in Table 10 use the full 60-second videos.

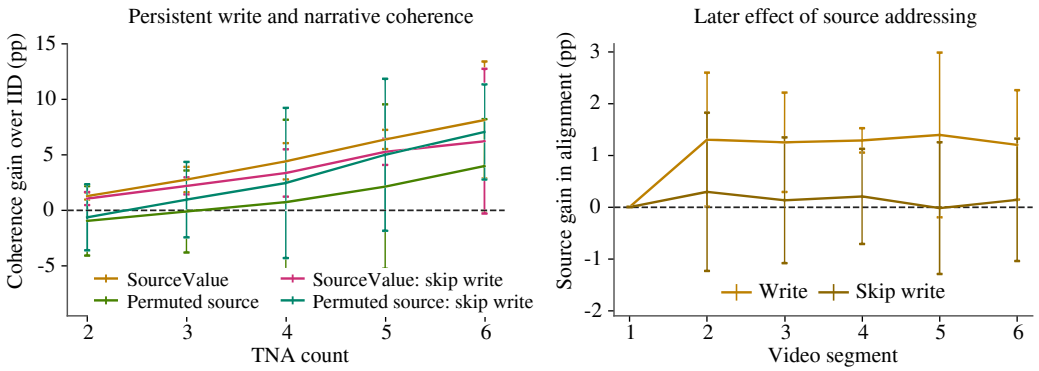


Figure 15: A committed noise record keeps affecting later generation. NOVA with retrieved or norm-matched permuted noise, crossed with writing or skipping one content/noise record after ten seconds. Left: NarrLV coherence gain over IID by TNA count (60 narratives per point). Right: retrieved-minus-permuted segment alignment on all 324 NarraStream scripts for each ten-second segment. Points and bars are paired means and pointwise 95% t-intervals over three seeds.

D.5 Variation among sampled futures

Each NOVA method generates a ten-second prefix for each of the 324 NarraStream scripts and then ten 50-second continuations per prefix, with independent fresh noise and the original instruction-release times. LPIPS-Alex v0.1 (Zhang et al., 2018) compares corresponding native-resolution RGB frames of futures in the cyclic pairs $(1,2),\ldots,(9,10),(10,1)$, averaging over frames and pairs and then equally over scripts. Joining each future to its prefix gives the full-video CLC. This yields 3,240 continuations per method and seed, and 9,720 over three seeds. Figure 16 and Table 11 use this population.

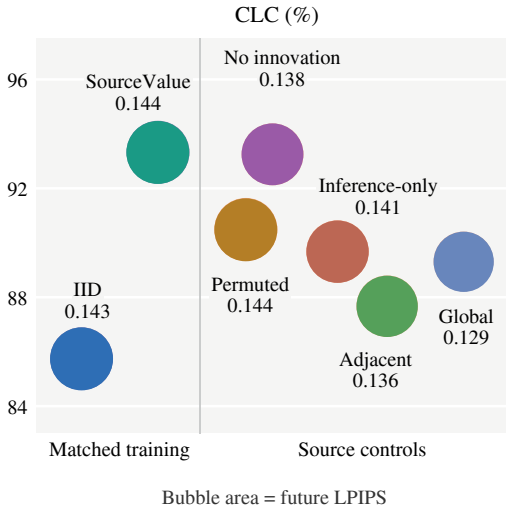


Figure 16: Consistency rises while futures stay diverse. NOVA on 324 NarraStream prefixes with ten 50-second futures each and three seeds. Height: full-video CLC; bubble area and labels: LPIPS between futures. Points are seed means.

Table 11 adds NOVA's native dynamic-degree and event tests. Dynamic degree (51.8 against 53.1) and StoryEval event completion (24.9 against 25.1) stay within their paired intervals, -3.91–1.32 and -0.65–0.25, so the adapted model still moves and still completes the requested events. On the future population, CLC rises from 85.7 to 93.3, while the LPIPS between futures changes from 14.27 to 14.36 (in units of 10^{-2}). High consistency together with unchanged LPIPS means that the ten futures agree on who is present but not on what happens. This is the behavior that Equation (3) is designed to produce: the inherited term is weighted by the attention on history, which is high where the returning character is drawn, and the fresh term adds variation at every position.

The prefix is generated once per method and seed, so the ten futures of a script differ only in the fresh noise drawn after ten seconds and in the noise that later units inherit from those draws. The LPIPS between futures therefore measures how much the futures differ given the same history; a drop would mean that inherited noise pulls the futures together.

In Figure 16, SourceValue reaches 93.3% CLC and an LPIPS of 0.144, against 85.7% and 0.143 for IID. Removing the fresh noise at one update gives 93.2% and 0.138, and frame-to-frame and shared-field noise give 87.7% and 89.3% with LPIPS 0.136 and 0.129: similar consistency can come with different variation, and only the attention read raises consistency at unchanged variation.

Table 11: NOVA responsiveness and conditional futures. DD: VBench dynamic degree; Events: StoryEval completion. Future CLC and LPIPS use the same ten 50-second continuations per prefix. Scores use 0–100; LPIPS uses 10^{-2} and measures variability, not quality.

Method	Native tasks		Future population	
	DD	Events	CLC	LPIPS
IID	**53.1**	**25.1**	85.7	14.27
SourceValue	51.8	24.9	**93.3**	14.36

Figure 17 shows ten futures from one prefix: the returning participant stays recognizable while poses and interactions change from future to future. Figure 7 in Section 5.5 locates this variation within the frame: local differences concentrate around the moving person and stay small in the corridor. This is the pattern that Equation (3) is built to produce: the corridor was visible in earlier frames, so its positions can attend to history and inherit its noise, while a person in a new pose shows content that the history does not contain and is drawn largely from fresh noise. The map is the spatial counterpart of the population result in Figure 16.

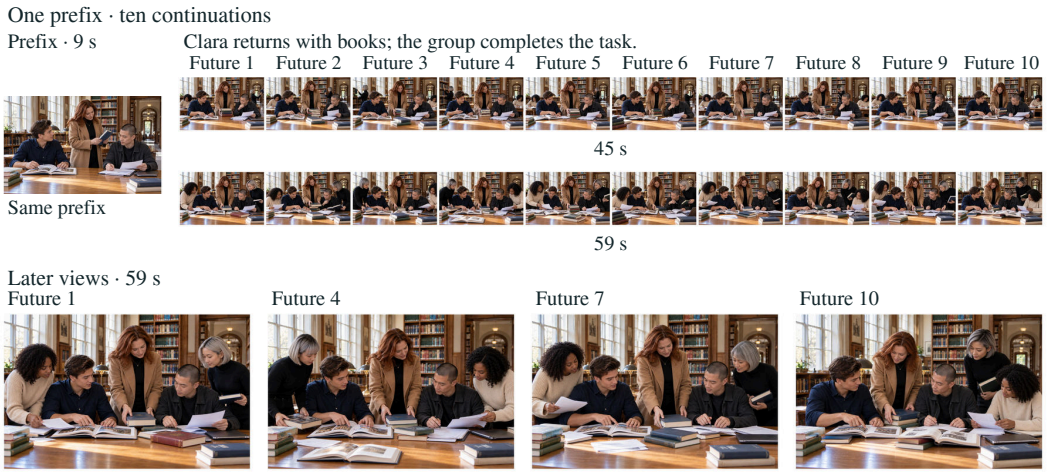


Figure 17: Ten futures from one shared prefix. Ten independently sampled SourceValue futures share the first nine seconds. Columns align the 45-second and 59-second views; four enlarged later views show different poses and interactions while the room and the participant remain recognizable.

D.6 Short-video and action-conditioned results

These tests check that SourceValue leaves tasks without a long absence as they were, on two further interfaces. Table 12 reports the complete image-to-video results for NOVA and VideoMAR. Every score changes by less than one point. Agreement with the input image (I-S, I-B) is unchanged, so the supplied first frame still anchors the video, and dynamic degree rises slightly for both models (38.5 to 38.7 and 43.6 to 44.4), so the videos keep their motion. Transition completion rises by 0.2 and 0.3 points. In a four-second clip the entities stay in view, so the conditioning already fixes most of their appearance and leaves little variance for the noise to settle, which is the regime where Equation (1) predicts a small effect.

Figure 18 shows what a completed transition entails: the requested attribute or action must change within a continuing scene, which requires the model to depart from the previous frame at the positions that change. The unchanged completion rate therefore also shows that content still changes where the instruction asks for change.

Table 13 reports HMA's action-conditioned prediction on LanguageTable. HMA reads its noise along same-position temporal edges (Appendix B.3). Prediction quality improves slightly (PSNR 25.70 to 25.86, FVD 153.5 to 150.0), and action sensitivity, the gap between final-frame PSNR under the supplied actions and under random actions, is unchanged (-0.014 dB; 95% CI -0.435–0.407). A model that copied its past would lose action sensitivity, because a random action sequence would

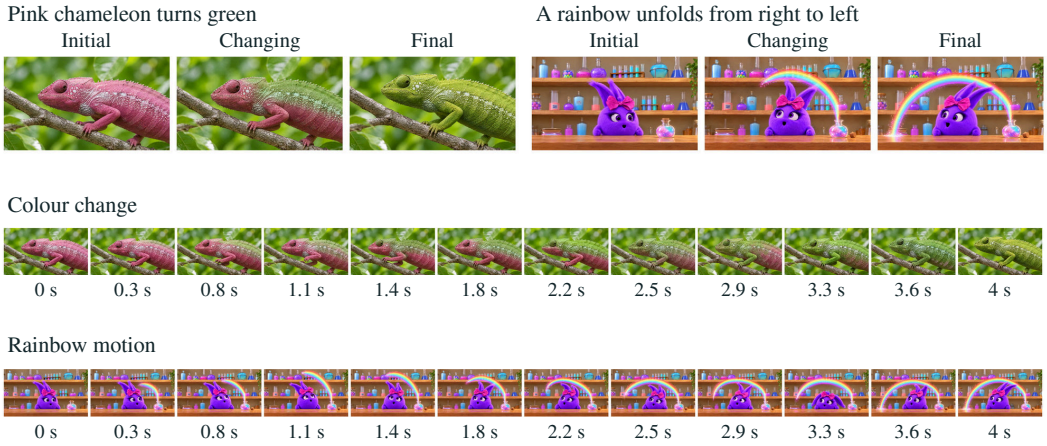


Figure 18: Temporal changes within a continuing scene. A pink chameleon turns green, and a rainbow unfolds from right to left. Upper panels show the initial, changing and final states; each timeline shows twelve frames spanning the four-second change. Table 12 reports the matched evaluation.

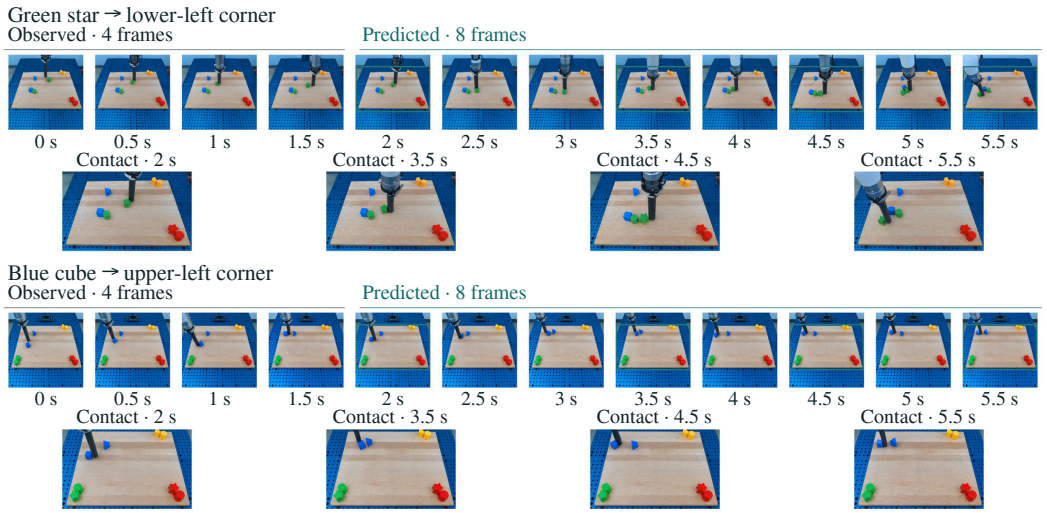


Figure 19: Action-conditioned prediction from four observed frames. The two tasks move a green star toward the lower-left corner and a blue cube toward the upper-left. Each timeline separates four observed frames from eight predicted frames; enlarged views show successive contact configurations. Table 13 reports the evaluation over 198 windows.

Table 12: Complete matched image-to-video scope. All values are means on 0–100 scales, higher is better. I-S/I-B: input-image subject/background agreement; Cam.: camera motion; SC/BC: within-video subject/background consistency; MS/DD: motion smoothness/dynamic degree; AQ/IQ: aesthetic/imaging quality; Q/C/T: official quality/condition/total. TC columns give transition completion, TC-Score, assertion accuracy and mapped adjacent-CLIP similarity. Native durations and all official input populations are retained.

Method	VBench I2V												TC-Bench I2V			
	I-S	I-B	Cam.	SC	BC	MS	DD	AQ	IQ	Q	C	T	TCR	Score	Acc.	CLIP
NOVA d48w1024–osp480 (0.646B), native 61-frame I2V (Deng et al., 2025)																
IID	97.0	**95.7**	19.0	**96.9**	**97.7**	**98.8**	38.5	65.1	68.9	**80.6**	91.7	86.14	32.8	78.1	83.9	**66.4**
SourceValue	**97.1**	**95.7**	**19.4**	**96.9**	**97.7**	**98.8**	**38.7**	**65.3**	**69.1**	**80.6**	**91.8**	**86.21**	**33.0**	**78.2**	**84.1**	**66.4**
VideoMAR checkpoint-736–480768 (1.394B), native 49-frame I2V (Yu et al., 2025a)																
IID	**97.8**	**96.6**	**23.6**	**97.6**	**98.1**	**99.1**	43.6	**67.2**	**71.4**	**82.3**	92.9	87.60	39.7	81.0	86.5	**70.2**
SourceValue	**97.8**	**96.6**	**23.6**	97.5	**98.1**	99.0	**44.4**	67.1	71.3	**82.3**	**93.0**	**87.63**	**40.0**	**81.1**	**86.6**	**70.2**

then produce nearly the same prediction as the true one. Figure 19 shows two predicted episodes in which the arm reaches and pushes the target object.

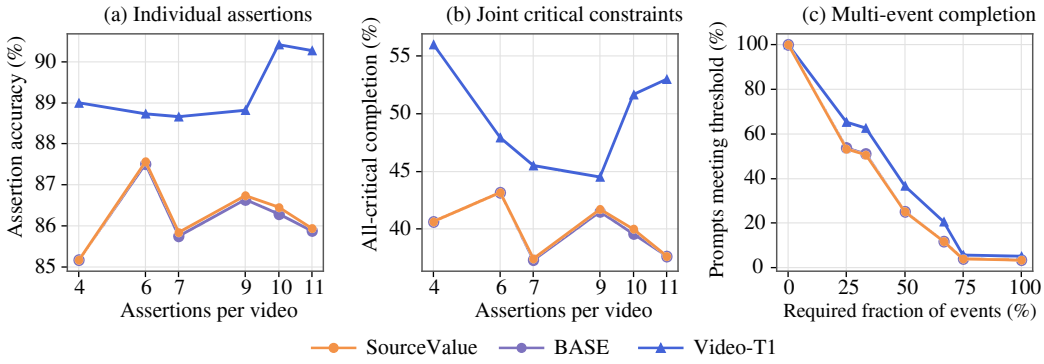


Figure 20: Individual requirements and complete task success. NOVA SourceValue, trained IID (BASE) and Video-T1 on TC-Bench text-to-video and StoryEval. (a,b) Assertion accuracy and the fraction of videos satisfying all critical assertions, over 150 descriptions with five videos each per repeat. (c) Fraction of the 423 StoryEval descriptions meeting each event-completion requirement. Curves average three seeds for SourceValue and BASE and three generation repeats for Video-T1.

Table 13: Action-conditioned HMA base-cont prediction (1.062B parameters) (Wang et al., 2025a). All 198 complete windows from 200 held-out LanguageTable episodes are retained. PSNR and ΔPSNR use dB; SSIM/LPIPS use 10^{-2} units. Means over three training seeds. The first three columns are higher-is-better; the last three are lower-is-better.

Method	Fidelity and action sensitivity ↑			Perceptual / distribution error ↓		
	PSNR	ΔPSNR	SSIM	LPIPS	FID	FVD
IID	25.70	**4.72**	78.5	11.8	43.5	153.5
SourceValue	**25.86**	4.71	**78.7**	**11.7**	**42.9**	**150.0**

Figure 20 separates individual requirements from complete task success for the NOVA systems on TC-Bench and StoryEval. Joint success is rarer than individual success for every method, SourceValue and IID have nearly identical curves, and Video-T1 attains higher completion through candidate search at a sampling budget of 145.20 minutes on four GPUs per NarrLV video (Table 8).

D.7 Inference and adaptation cost

Table 14 reports the time for one output and the optimizer-stage adaptation cost. Inference for all four backbones uses batch one on an A100 80GB PCIe GPU and includes input encoding, native generation, VAE decoding and device synchronization; benchmark scoring runs separately. Training uses four such GPUs, and GPU-hours are wall time multiplied by the device count; the totals include the target-hidden views of Appendix B.3, among them VideoMAR's 384 views per window, which dominate its cost.

Storing M source rows of width p takes $4Mp$ bytes in FP32, and the head-averaged read costs $O(NMp)$ multiply–adds for N queries. A NOVA source frame occupies $360\times64\times4 = 92{,}160$ bytes. At 60 seconds, 180 committed frames occupy 16,588,800 bytes, and the pending 181st frame raises the peak to 16,680,960 bytes (15.91 MiB). The native run performs 181 temporal passes, 11,584 spatial passes and 289,600 head evaluations, and SourceValue adds none of them.

The measured device peak also includes the model, the content cache, activations and tem-

Table 14: Inference and adaptation cost; three-repeat means. Time: end-to-end seconds per output on one A100 80GB PCIe (batch one); Peak: allocated GiB; Opt.: adaptation GPU-hours on four GPUs. Inherited checkpoints are included; base pretraining and hyperparameter search are excluded. Wan uses 60-second NarrLV outputs; Time/60 matches Table 8. Lower is better.

Method	Time (s)	Peak	Opt.
NOVA d48w1024 (0.646B) (Deng et al., 2025): 60 s			
IID	**1434.9**	**29.17**	**326.3**
SourceValue	1441.9	29.23	329.1
VideoMAR 480p (1.394B) (Yu et al., 2025a): 49 frames			
IID	**244.2**	**22.43**	**3308.0**
SourceValue	245.8	22.49	3352.0
HMA base-cont (1.062B) (Wang et al., 2025a): 8 predicted frames			
IID	**24.1**	**13.12**	**294.5**
SourceValue	24.3	13.15	298.0
Common Wan2.1-T2V-14B (Team Wan et al., 2025) (480p, 16 fps; method-specific budgets)			
Ring Forcing (Xue et al., 2026)	11308.67	61.26	**3306.67**
MemRoPE (Kim et al., 2026)	1469.67	**43.83**	5154.67
LongLive (Yang et al., 2026)	1425.67	44.17	5154.67
World-R1-Large (Wang et al., 2026a)	14425.33	66.12	6257.33
Stable Video Infinity (Li et al., 2026)	9975.00	57.81	3828.00
Self Forcing (Huang et al., 2025)	**1291.00**	45.23	4568.00
Video-T1 (chunk tree) (Liu et al., 2025)	43082.33	59.23	3692.00
IID	10813.33	58.21	3692.00
SourceValue	10994.33	58.44	3754.67

porary buffers; it grows by 0.06 GiB for NOVA. The added time, 7.0 s per 60-second NOVA video, comes from the head-averaged read and the FP32 noise store.

For Wan, SourceValue takes 10,994.33 s against IID's 10,813.33 s, an increase of 181.00 s (1.67%). Peak memory rises from 58.21 to 58.44 GiB, and adaptation from 3,692.00 to 3,754.67 GPU-hours (1.70%). The 20 historical frame-equivalents plus three pending latent frames require $23 \times 60 \times 104 \times 16 \times 4 = 9{,}185{,}280$ source bytes, or 8.76 MiB. World-R1 retains one previous three-frame source block (1.14 MiB); the other Wan rows keep no cross-block source record. Ordinary transient noise is included in Peak. Costs include each method's teachers, geometric components, candidates and verifier where used; Opt. includes the LongLive checkpoint inherited by MemRoPE and the IID checkpoint inherited by Video-T1, and excludes base pretraining and hyperparameter search. Wan inference times use the same outputs in Tables 8 and 14, with seconds divided by 60 before rounding.

E Object readouts and conditional moments

This appendix defines the object readout used in Section 5.4 (Section E.1), relates its feature distance to human identity votes (Section E.2), derives the bounds that accompany Table 6 (Section E.3), and describes a controlled Gaussian replay of the analysis (Section E.4).

E.1 Appearance features and visibility

The readout segments the described entity with SAM 3 (Carion et al., 2026) at resolution 1008 and confidence threshold 0.5, maps the masks to the original image and thresholds them at 0.5. Within each decoded chunk it selects the largest valid mask, breaking ties by confidence, earlier frame and bounding-box order, crops the tight box, sets the background outside the mask to white, and resizes the crop to 256×256. DINOv3 ViT-B/16 pretrained on LVD-1689M (Siméoni et al., 2025) supplies a unit-normalized 768-dimensional CLS feature ϕ after its native input normalization. Each branch computes its own mask and feature. An empty mask or a nonfinite feature sets the validity indicator $b = 0$, and the readout is $Y = b\,\phi$ with $\|\phi\| = 1$ when valid.

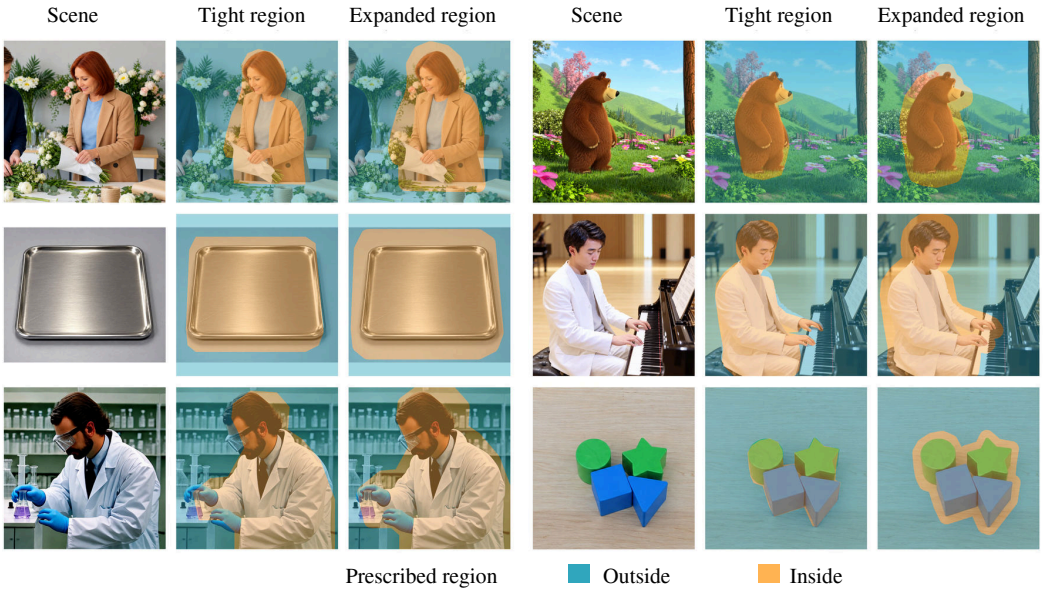


Figure 21: Region boundaries change what a comparison includes. Six cases: flower-shop character, plate, scientist, bear, pianist and blocks. Each triplet shows the scene, a tight region and an expanded region (orange inside, cyan outside); compare the newly included boundary and nearby background. These prescribed overlays illustrate spatial support and differ from the SAM 3 masks used for the object features.

The readout therefore mixes visibility and appearance. With $\pi_{\text{vis}} = \Pr(b = 1 \mid C)$ and $\mu_{\text{vis}} = \mathbb{E}[\phi \mid C, b = 1]$,

$$\operatorname{tr}\operatorname{Cov}(b\,\phi \mid C) = \pi_{\text{vis}}\ \operatorname{tr}\operatorname{Cov}(\phi \mid C, b = 1) + \pi_{\text{vis}}(1 - \pi_{\text{vis}})\|\mu_{\text{vis}}\|^2, \tag{15}$$

so mask failures enter the variance. The decoder also gives pixel features a path to the history beyond the sampler's conditioning C. With the notation of Appendix A.1, the discrepancy difference under the natural generated-history distribution is

$$D^{\text{IID}} - D^{\text{SV}} = 2\big(K^{\text{SV}}_{k,h} - K^{\text{IID}}_{k,h}\big) + \big(G^{\text{IID}} - G^{\text{SV}}\big) + \big(V^{\text{IID}}_{k} - V^{\text{SV}}_{k}\big), \tag{16}$$

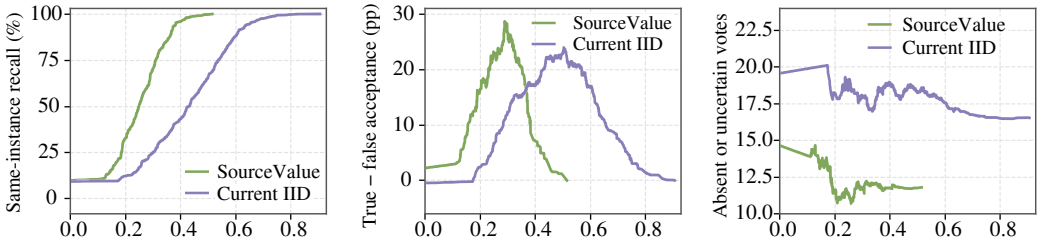
where all terms are averaged over the same generated-history distribution and $K^{\text{IID}}_{k,h}$ collects the dependence that reaches an IID-noise output through the decoder. This term explains the positive IID cross-covariance in Table 6.

NarraStream contributes 1,046 entity descriptions and 3,984 first-to-later relations over 14 endpoint combinations. NarrLV anchors 737 object nouns from 357 narratives at the generated first frame; three noun-free narratives enter task scoring only. Current-appearance measurements include all applicable queries, and historical comparisons require an earlier appearance. Figure 21 shows how the choice of region boundary changes what a visual comparison includes.

E.2 Feature distance and identity votes

Figure 22 relates the squared feature distance $\|Y_k - Y_h\|^2$ to five blinded votes per relation. The 3,984 NarraStream relations, ten rollouts and three seeds yield 119,520 evaluations per noise rule; the IID reference pools two independent noise branches at the same checkpoint and prefix (239,040 evaluations). Votes distinguish same, different, absent and undetermined instances and allow for changes in pose, viewpoint and illumination. Let $n_c(\tau)$ count the category-c votes at squared distance at most τ and $n_c = n_c(\infty)$, for $c \in \{s, d, a, u\}$ (same, different, absent, undetermined). Panels (a)–(c) report the recall $n_s(\tau)/n_s$, the separation $n_s(\tau)/n_s - n_d(\tau)/n_d$, and the ambiguity $\{n_a(\tau) + n_u(\tau)\}/\sum_c n_c(\tau)$. All pairs and votes enter at their observed frequencies, including failed-mask zeros, and the curves connect 379 SourceValue and 757 pooled-IID thresholds.

SourceValue reaches same-instance votes at smaller distances, and its separation between same and different votes peaks at about 28 points near a squared distance of 0.3, above the IID peak of about 23 points near 0.5. Absent or undetermined votes make up about 12% of SourceValue votes when all pairs are accepted, against about 16.5% for IID. The feature distance thus tracks the judges' identity decisions, but only partially, which is why the primary endpoint uses human judgments and the features serve the mechanistic analysis.



(a) Retaining true matches (b) Identity separation (c) Ambiguity among accepted

Figure 22: Feature distance separates identity categories only partially. Squared SAM 3/DINOv3 object-feature distance on 3,984 NarraStream relations, ten rollouts and three seeds; IID pools two noise branches at the same checkpoint and prefix. (a) Same-instance recall; (b) same-minus-different acceptance; (c) absent or undetermined share among accepted votes. Curves keep all observed thresholds and mask failures.

E.3 Risk and reversible-pair bounds

Table 6 reports conditional moments, which require conditional means. On held-out trajectories, a predictor $\widehat{m}(C)$ satisfies

$$\mathbb{E}\|Y - \widehat{m}(C)\|^2 = \mathbb{E}\ \mathrm{tr}\,\mathrm{Cov}(Y \mid C) + \mathbb{E}\|\mathbb{E}[Y \mid C] - \widehat{m}(C)\|^2,$$

so the prediction risk bounds the conditional variance plus an approximation error. The predictor uses 256 sine/cosine random Fourier features of unit norm, a coefficient Frobenius norm of at most one, and ridge strengths $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. Five outer trajectory folds each hold out two of ten repeats, and four inner folds select the ridge strength. The target conditions on all non-noise inputs C, and the finite-feature approximation error stays in the bound.

Lower bounds come from perturbing the noise without changing C. Each source coordinate is generated by applying the inverse standard normal CDF to $(w + 0.5)/2^{32}$ for a uniform 32-bit word w and rounding to FP32. A proposal flips one fixed low-order bit in a uniformly chosen word and is accepted only if the conditioning C is exactly unchanged. This symmetric proposal defines a conditionally reversible rejection kernel $\mathcal{T}_0$, whose lazy version $\mathcal{T} = (I + \mathcal{T}_0)/2$ has spectrum in $[0, 1]$. For centered f, g with variances V_f, V_g and covariance K_{fg}, define $\mathcal{E}_f = \langle f, (I - \mathcal{T})f\rangle$ and $\mathcal{E}_{fg} = \langle f, (I - \mathcal{T})g\rangle$. Then

$$0 \le \mathcal{E}_f \le V_f, \qquad |K_{fg} - \mathcal{E}_{fg}| \le \sqrt{(V_f - \mathcal{E}_f)(V_g - \mathcal{E}_g)}, \tag{17}$$

where the second inequality is Cauchy–Schwarz for the positive operator $\mathcal{T}$. Each nonlazy proposal contributes $\|f - f^\star\|^2/4$ and $(f - f^\star)^\top(g - g^\star)/4$, with the lazy self-loop averaged analytically; rejections and unchanged features contribute zeros. Held-out risks and $2K_{fg} = V_f + V_g - V_{f-g}$ complete the outer intervals.

Accepted perturbations leave the historical object feature unchanged, so its variance lower bound is zero, and the averaged upper bounds are 0.707 for NarraStream and 0.713 for NarrLV. The simultaneous bounds on the discrepancy gain are $[-0.207, 0.556]$ and $[-0.230, 0.496]$. The point estimates in Table 6 show the pattern predicted by Equation (1): in units of 10^{-2}, the NarraStream discrepancy falls by 17.5, twice the increase in cross-covariance is 17.6, and the current variance changes by 0.1; on NarrLV the discrepancy and covariance terms are 13.3 and 13.0. The joint intervals are wide because they combine finite-sample error with every value of the conditional moments that the risk and perturbation bounds allow.

E.4 Controlled Gaussian replay

The Gaussian comparisons use ten repetitions at each of the 14 occurrence-endpoint combinations, with source dimension $23{,}040\,h$ for an occurrence at unit h. Endpoint-specific first- and second-degree terms are signed and have wide intervals. At update 181, a bounded nonpolynomial response uses 324 anchors, ten independent Gaussian chains per repeat, $L = 0, \ldots, 180$, $r = 0.5^L$ and a shared IID reference. Its sign and sign-product terms produce an infinite Hermite tail; the discrepancy vanishes at $L = 0$, and Theorem 2 gives the strict advantage of a direct read for $L > 1$. Simultaneous bounds allocate error 0.05 across 60,675 intervals: weighted bounded means use Hoeffding radii, independent-pair decompositions handle shared-endpoint U-statistics, and full-dimensional Hermite bounds use nonnegative degree energies, a common $D(0)$ and an unrestricted higher-order tail.