EDITED ENDPOINTS AS PARTIAL GOALS FOR IMAGE-TO-VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Edited endpoint images make the target of an image-to-video transition explicit and can improve transition completion, but also preserve context in the source image's layout. Full-frame endpoint conditioning turns this preserved context into requirements on the video's final state, potentially inducing temporal artifacts. We measure them with Hard-Cut (HC) and Background Motion Reversal (BMR), for abrupt transitions and late background reversals, alongside Dynamic Degree (DD) for motion reduction, and observe model-specific artifacts across five video models spanning training-free and learned endpoint conditioning. We study Partial, a selective conditioning scheme that constrains only an edit-relevant region estimated from image-editor attention and leaves the remaining terminal content to the video model. Across three backbones on TC-Bench-I2V, Partial reduces temporal artifacts relative to full-frame conditioning, with no significant difference in Transition Completion Ratio (TCR); on Wan2.2-Turbo, uniformly weakening conditioning does not reproduce these reductions. A donor-context intervention, which replaces only the context outside the edit-relevant region with the final frame of a Partial video under the same full-frame conditioning, lowers HC and BMR on HunyuanVideo-1.5 and lowers BMR while restoring DD on Wan2.2-Turbo, with no significant difference in TCR, so constraint extent alone does not account for these artifacts. Together, these findings suggest that edited endpoints should serve as partial future goals, although edit relevance does not guarantee future validity.

1 INTRODUCTION

Given a source image, "change the bus to purple" specifies a target appearance but not a trajectory. A video may miss or undo the change before its last frame (Feng et al., 2025). Editing the source image makes the target explicit, as edited keyframes already do in video editing and interpolation (Ku et al., 2024; Ouyang et al., 2024; Liu et al., 2025; Ma et al., 2024; Li et al., 2026).

An edited endpoint mixes two kinds of information: the purple bus specifies the target appearance, while the unchanged road, buildings, and sign preserve the editor's input in the source image's layout. When an edited frame shares its source frame's time, as in edit-then-propagate video editing, this alignment makes it a reliable anchor. As an endpoint, the same context is imposed at a later time: full-frame terminal conditioning requires the road, buildings, and sign to return to their source-time positions at the video's end. We call this promotion of preserved context to spatial future requirements *spatial over-specification*.

Consistent with this concern, full-frame endpoint conditioning can improve completion while reducing movement or increasing discontinuity and background motion reversal, under both training-free projection and learned endpoint conditioning (Tables 1 and C8). We study *whether terminal content affects temporal artifacts beyond constraint extent*.

Using established masked conditioning (Lugmayr et al., 2022; Avrahami et al., 2023), we test content at fixed full-frame support and vary support area, placement, and global mixing weight. Partial constrains the terminal frame to the edited endpoint only within fixed edit-relevant support estimated from image-editor attention, retaining the video's prediction elsewhere (Fig. 1).

A comparison between full-frame endpoint conditioning and Partial changes both the terminal content being imposed and the spatial extent of the constraint. To test content at fixed extent, we run a donor-

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.

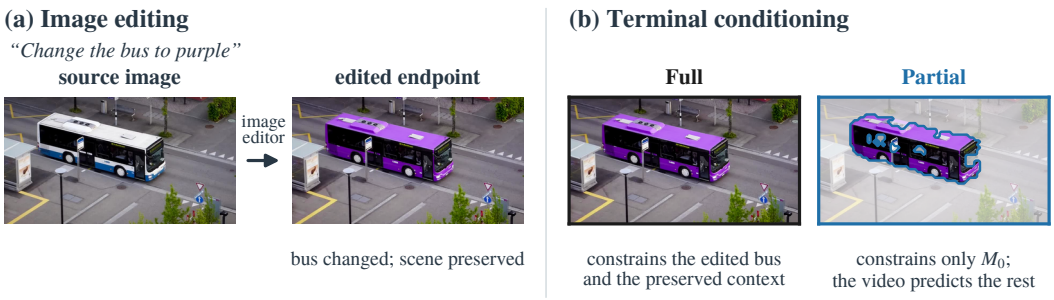


Figure 1: **Edited endpoints preserve context that need not specify the video's future.** Full-frame endpoint conditioning constrains both requested appearance and preserved context. Partial constrains only estimated edit-relevant support M_0 from image-editor attention (blue outline); washed-out positions retain the video prediction.

context intervention: inside the estimated edit support M_0, we keep the original edited endpoint; outside M_0, we substitute the final frame of a paired Partial rollout, the donor, and constrain the entire terminal frame with the same full-frame projection.

Across 103 TC-Bench-I2V prompts with five paired seeds per condition, we report three findings:

1. Terminal content matters beyond constraint extent. With full-frame support unchanged, the donor-context intervention lowers Hard-Cut (HC) and Background Motion Reversal (BMR) scores on HunyuanVideo-1.5, and lowers BMR while restoring Dynamic Degree (DD) on Wan2.2-Turbo (Fig. 3).

2. Support area and placement yield distinct completion–temporal trade-offs. On Wan2.2-Turbo, edit-relevant placement favors completion at matched area; shrinking support lowers HC, whereas enlarging it raises BMR. Tested global softening does not reproduce Partial's artifact reductions.

3. Selective constraints offer practical gains with residual costs. Across three evaluated backbone–conditioning settings, Partial mitigates temporal artifacts relative to full-frame endpoint conditioning at comparable Transition Completion Ratio (TCR), and improves completion over Endpoint-free on Wan2.2-Turbo and SANA-Video. Edit relevance need not imply future-valid position or visibility.

2 Related Work

Frame guidance and masked conditioning. Frame Guidance optimizes video latents using selected decoded frames, including edited-frame inputs and binary-masked L_2 guidance (Jang et al., 2026). Our projection follows RePaint's mixing of generated and noise-matched known pixels and Blended Latent Diffusion's mixing of generated and source-image latents (Lugmayr et al., 2022; Avrahami et al., 2023). We use this established mixing principle to analyze terminal constraints, rather than introduce a new guidance operator.

Edit localization and regional control. Prompt-to-Prompt connects prompt words with image layout through cross-attention (Hertz et al., 2023). DiffEdit derives masks from contrasts between prompt-conditioned noise predictions, whereas InstDiffEdit derives attention masks during editing (Couairon et al., 2023; Zou et al., 2024). MasaCtrl and MAG-Edit use masks for attention control and local latent optimization, respectively (Cao et al., 2023; Mao et al., 2024). EditCtrl combines local masked and global temporal context for video editing (Litman et al., 2026). For video generation from a single image, we use image-editor attention to estimate edit-relevant terminal support.

Edited keyframes, target states, and inbetweening. Prior work uses edited or specified frames as appearance anchors to propagate or as states to connect, typically imposing the whole frame at its time index. We instead impose an edited frame at a different time from its source, with no trajectory relating the two, and ask which parts of its spatial content remain valid future requirements. Edit-then-propagate video editing edits the first frame with an image editor and carries it along a source video that supplies motion (Ku et al., 2024; Ouyang et al., 2024; Liu et al., 2025). TC-Bench's SDXL+SEINE baseline and MaskINT edit keyframes and interpolate between them (Feng et al., 2025; Podell et al., 2024; Chen et al., 2024; Ma et al., 2024). Inbetweening methods connect specified

start and end states through fine-tuned or training-free bidirectional generation (Wang et al., 2025; Feng et al., 2024; Yang et al., 2025a;b). Wan2.1-FLF2V, Many-for-Many, and OmniWeaving instead learn first–last-frame or flexible frame conditioning (Wan et al., 2025; Li et al., 2026; Pan et al., 2026). VACE, Many-for-Many, and VideoCanvas support spatially masked video conditioning (Jiang et al., 2025; Li et al., 2026; Cai et al., 2026), offering learned approaches to partial goals.

Temporal artifacts of conditioning. Conditional image leakage links limited I2V motion to source-image over-reliance at large diffusion timesteps (Zhao et al., 2024). InfinityStory's supplementary ablations report restricted natural motion from fixing each shot's keyframe and rushed transitions toward prescribed positions (Elmoghany et al., 2026). DCARL addresses keyframe duplication and boundary discontinuity with noisy conditioning and boundary consistency (Ouyang et al., 2026). TIC-FT diagnoses abrupt content and noise-level transitions under direct concatenation and uses graded-noise buffer frames (Kim et al., 2025). Building on these observations, our donor-context intervention tests whether replacing editor-preserved terminal context changes temporal artifacts under unchanged full-frame support and projection (Section 5.2).

3 Problem Formulation and Method

3.1 Problem Setting

Given a source image I_0 and a transition prompt, an image editor produces an endpoint I_E from an edit prompt that embeds the transition prompt and asks for the frame after the transition completes (Appendix A). A frozen video model generates $V = (I_0, \ldots, I_T)$ from the source and the same transition prompt under three conditions: Endpoint-free, full-frame endpoint conditioning (Full), and partial endpoint conditioning (Partial). Within each controlled comparison, conditions share the source and prompt; endpoint-conditioned conditions share I_E except in Section 5.2.

3.2 Edited Endpoints and Partial Goals

An idealized decomposition makes the mismatch explicit:

$$I_E \approx M^* \odot I_T^{\text{edit}} + (1 - M^*) \odot I_0, \tag{1}$$

where M^* is an idealized edit-relevance mask and I_T^{edit} denotes the requested target appearance. This decomposition represents an editor that changes requested content while preserving context in the source layout. The second term adds no appearance beyond I_0, which already conditions the video; as a terminal requirement, it only fixes where source content must be at time T. Only the first term carries information that I_0 lacks. Full treats both terms as future requirements. Yet preserved context may change position or visibility as the scene evolves, so forcing it back can impose constraints beyond the edit prompt: spatial over-specification. Our FLUX.2 endpoints follow this pattern: outside the estimated support M_0 (Section 3.3), they differ from I_0 by a mean absolute RGB difference of 0.056, versus 0.232 inside (Appendix E.4).

Edit relevance and future validity. *Edit relevance* associates content with the requested change; *future validity* concerns its appropriateness as a future constraint. Partial uses the former to select support, without estimating the latter. Neither M^* nor its estimate below is a future-validity mask.

3.3 Edit-Relevant Support Estimation

We estimate binary support M_0 from the editor's text–image attention to object and edit phrases preselected from each prompt and fixed across conditions. We normalize and average their attention maps, then apply a two-component Gaussian mixture with an Ising spatial prior (GMM–Ising; Appendix A; Fig. A1).

The object and edit phrases connect the requested change to spatial regions through the editor's attention, so M_0 approximates M^* in Eq. (1) without a separate segmentation model.

3.4 Full, Soft, and Partial Conditioning

Partial imposes the endpoint only on M_0, keeping the content that I_0 lacks, and retains the video prediction elsewhere because future context is unknown.

Weighted terminal projection. Let $z_E = E(I_E)$ encode the endpoint, u index spatial latent positions, and $\hat{z}_T^{(k)}$ denote the clean prediction of the terminal latent frame at sampling step k. All conditions share one projection with a spatial weight w:

$$\hat{z}_T^{(k)}(u) \leftarrow w(u)\, z_E(u) + [1 - w(u)]\, \hat{z}_T^{(k)}(u), \tag{2}$$

with $w \equiv 1$ for Full, $w \equiv \alpha \in \{0.75, 0.5, 0.25\}$ for Soft, and $w = M_0$ for Partial, after transformation to each latent grid. Soft weakens every terminal position uniformly; α is a mixing weight, not a calibrated effective total constraint strength. Support stays fixed in terminal spatial coordinates throughout sampling and does not track motion.

Backbone implementations. Wan2.2 applies the same w to the terminal model input, with token timesteps scaled by $1 - w$. SANA Partial converts the projected clean prediction back to its solver's flow, while SANA Full clamps the terminal state. Hunyuan uses a global timestep, so it projects toward the endpoint noised to the current level after each scheduler update. Appendix A specifies encoding, mask transforms, and noise handling. Learned endpoint-conditioning models are additional observational baselines, not implementations of this projection rule.

4 Main Experiments

4.1 Experimental Setup and Metrics. We evaluate 103 TC-Bench-I2V prompts with five paired seeds per condition (515 videos) (Feng et al., 2025), using FLUX.2 endpoints. The primary backbones are Wan2.2-Turbo (Wan2.2), SANA-Video 2B (SANA; Chen et al., 2026), and HunyuanVideo-1.5 (Hunyuan; Wu et al., 2025a). Wan2.2-Turbo is quanhaol's community-distilled Wan2.2-TI2V-5B-Turbo checkpoint.[1] Appendix A gives sampler configurations.

Image-editing prior and comparison logic. FLUX.2-klein-base-9B edits the source image into a plausible final frame after the requested transition has completed (Appendix A.2). We generate one endpoint per prompt and reuse it for Full and Partial. Endpoint-free receives the source image and transition prompt; Full additionally constrains the entire terminal frame to the edited endpoint, whereas Partial constrains only M_0. Thus Endpoint-free versus Full tests the completion benefit and temporal artifacts of adding an explicit image-editing goal, while Full versus Partial tests selective use of that same goal. Partial requires no video-model fine-tuning.

Related baselines. Table 1 also summarizes selected comparisons from Table G1. Within the Wan2.2 family, we also include the non-distilled Wan2.2 Base and Wan2.2 with Qwen-Image-Edit endpoints (Wan et al., 2025; Wu et al., 2025b). OmniWeaving, built on HunyuanVideo-1.5, and MfM provide learned endpoint-conditioning baselines (Pan et al., 2026; Li et al., 2026). The three controlled blocks compare conditions within each backbone; the additional methods use their native configurations, with edit-type breakdowns and timing in Table G1.

Completion and temporal diagnostics. We report Transition Completion Ratio (TCR) and TC-Score with a Qwen2.5-VL-32B-Instruct judge (Bai et al., 2025). TCR conjoins transition and object-consistency assertions, excluding "other objects" as in TC-Bench (Feng et al., 2025). TC-Score excludes transition-object consistency from its VLM term and adds adjacent-frame CLIP consistency (Radford et al., 2021), which also varies with motion, so we use TCR as the primary completion measure.

Temporal diagnostics use a common 8 FPS grid. Dynamic Degree (DD) detects substantial movement using VBench's RAFT-based decision (Huang et al., 2024; Teed & Deng, 2020). Abrupt changes and camera drift can raise flow scores (Ge et al., 2026; Minar et al., 2026), so DD does not assess motion appropriateness. Hard-Cut (HC) measures the largest drop of the adjacent-frame similarity s_t below its median, where ϕ is the CLIP image encoder and I_t the t-th frame on this grid:

$$\mathrm{HC} = \mathrm{median}_t\, s_t - \min_t s_t, \qquad s_t = \cos\big(\phi(I_t), \phi(I_{t+1})\big). \tag{3}$$

BMR measures the fraction of late along-axis motion mass opposing an early reference direction in the region outside the original estimated support M_0, which approximates the background and stays fixed across conditions, including shifted/random conditioning supports. Let $F_t(x)$ denote WAFT

[1] Official releases: FLUX.2-klein-base-9B and Wan2.2-TI2V-5B-Turbo.

Table 1: **Image-editing goals: completion and temporal trade-offs.** TC-Bench-I2V, 103 prompts $\times$ five paired seeds. Endpoint-free uses the source image and prompt only. Full and Partial share the same FLUX.2-klein-base-9B edited endpoint; their constraints cover the entire terminal frame and M_0, respectively. Related baselines use the endpoint prior named in each row.

	Transition completion			Temporal diagnostics		
Condition / endpoint prior	TCR ↑	ΔTCR (pp) ↑	TC-Score ↑	HC ↓	DD ↑	BMR ↓
Wan2.2-Turbo						
Endpoint-free	33.79	–	0.7841	**4.57**	**79.4**	**0.360**
Full (FLUX.2)	41.94	+8.16*	**0.8352**	6.47	66.8	0.599
Partial (FLUX.2)	**42.91**	+9.13*	0.8187	5.76	76.7	0.387
Δ (Partial − Full)	+0.97	–	−0.0164	−0.72	+9.90	−0.213
SANA-Video						
Endpoint-free	23.50	–	0.7530	**8.07**	**42.7**	**0.122**
Full (FLUX.2)	**35.73**	+12.23*	**0.7944**	9.77	32.8	0.420
Partial (FLUX.2)	35.34	+11.84*	0.7935	8.80	42.3	0.135
Δ (Partial − Full)	−0.39	–	−0.0008	−0.97	+9.51	−0.285
HunyuanVideo-1.5						
Endpoint-free	40.97	–	0.8718	**3.47**	51.5	**0.305**
Full (FLUX.2)	**46.60**	+5.63	0.8588	12.96	**52.6**	0.585
Partial (FLUX.2)	44.47	+3.50	**0.8804**	5.69	51.5	0.333
Δ (Partial − Full)	−2.14	–	+0.0216	−7.27	−1.17	−0.252
Related I2V and endpoint baselines (Table G1)						
Wan2.2 Base (FLUX.2)	45.24	+14.17*	0.8475	7.18	57.9	0.497
Wan2.2-Turbo (Qwen-Image-Edit)	35.15	+1.36	0.8001	7.11	64.1	0.617
OmniWeaving (FLUX.2)	43.30	+3.88	0.8647	5.65	36.9	0.487
MfM (FLUX.2)	43.88	+11.26*	0.8869	4.13	28.2	0.387

ΔTCR is the paired difference from Endpoint-free; Δ (Partial − Full) rows give paired Partial − Full differences in each column. *Significant; intervals in Tables C1 and C6. TCR/DD are percentages (differences in pp); HC is multiplied by 100. HC, BMR, and DD are temporal diagnostics: HC and BMR measure discontinuity and background reversal directly, while DD measures movement, so a drop relative to Endpoint-free indicates suppressed motion.

flow (Wang & Deng, 2026) and L the number of flow pairs; the median is taken componentwise. The magnitude-weighted vector b_t, reference axis $\hat{b}$, and reversal score are

$$b_t = \frac{\sum_{x \notin M_0} \|F_t(x)\| F_t(x)}{\sum_{x \notin M_0} \|F_t(x)\|}, \qquad \hat{b} = \frac{b_{\text{ref}}}{\|b_{\text{ref}}\|}, \quad b_{\text{ref}} = \text{median}_{t<L/2}\, b_t,$$
$$\text{BMR} = \frac{\sum_{t \geq L/2} \max(0, -\langle b_t, \hat{b} \rangle)}{\sum_{t \geq L/2} |\langle b_t, \hat{b} \rangle|}. \tag{4}$$

BMR lies in $[0, 1]$ and depends on the estimated mask and flow. Section 5.6 gives track-based checks.

Uncertainty. We report 95% confidence intervals (CIs) from 100,000 paired prompt-bootstrap draws over 103 prompts (Appendix C). A paired difference is significant when its pointwise CI excludes zero; otherwise we call the conditions comparable, which does not imply equivalence.

4.2 Endpoint-Free vs. Full Conditioning. For Endpoint-free→Full (Table 1), TCR rises significantly by 8.16 percentage points (pp) on Wan2.2 and 12.23 pp on SANA. DD decreases by 12.62 and 9.90 pp, respectively, while HC and BMR increase on both. Full supplies useful target information while inducing temporal artifacts.

Hunyuan's principal costs are discontinuity and reversal: HC rises from 3.47 to 12.96 and BMR from 0.305 to 0.585, while DD is 51.5% for Endpoint-free and 52.6% for Full. Thus, Full's temporal artifacts differ across settings. These observations are consistent with spatial over-specification: outside M_0, the first–last-frame difference is 0.13–0.23 for Endpoint-free videos but 0.06–0.08 under Full, which returns preserved context near its source-time layout (Appendix E.4).

Beyond these backbones, FLUX.2 Full conditioning significantly raises TCR on Wan2.2 Base (31.07% to 45.24%, +14.17 pp) and MfM (32.62% to 43.88%, +11.26 pp), whereas the gains of OmniWeaving (39.42% to 43.30%, +3.88 pp) and of Wan2.2 with Qwen-Image-Edit endpoints

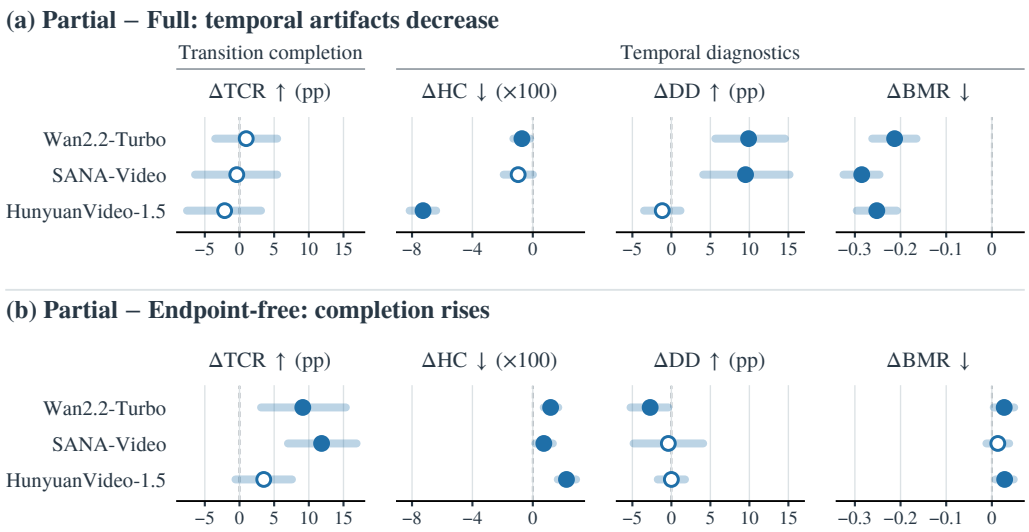


Figure 2: **Partial reduces temporal artifacts relative to Full.** Paired differences with 95% CIs over 103 prompts and five seeds; open markers are not significant. Tables 1 and C1 give levels and intervals.

(33.79% to 35.15%, +1.36 pp) are not significant (Table C6). An edited target can aid completion across conditioning approaches, but the magnitude and certainty of the gain vary across models and editors. Full's temporal artifacts also persist under learned endpoint conditioning: Full sharply lowers DD by 20.19 pp for MfM and 17.67 pp for OmniWeaving, and raises OmniWeaving's BMR (Table C8).

4.3 Partial vs. Endpoint-Free and Full. Relative to Full, Partial mitigates temporal artifacts at comparable TCR. Across Wan2.2, SANA, and Hunyuan, TCR stays within 2.14 pp of Full, while HC, DD, and BMR each move toward their Endpoint-free levels, and every significant temporal difference favors Partial (Fig. 2; Tables 1 and C1). Outside M_0, Partial leaves the first–last-frame difference of preserved context within 0.02 of Endpoint-free (Appendix E.4).

The mitigation is not a return to Endpoint-free. Partial keeps a significant TCR gain over Endpoint-free on Wan2.2 and SANA (+9.13 and +11.84 pp), and on Wan2.2, global softening at comparable TCR does not reproduce these reductions (Section 5.3). Some temporal artifacts remain relative to Endpoint-free (Table 1). Edit-type breakdowns are in Table C7.

5 Controlled Analysis

Section 5.1 localizes Hunyuan's Full discontinuity at the terminal boundary, and Section 5.2 tests whether terminal content, not only constraint extent, drives Full's temporal artifacts. We run this donor-context intervention on Hunyuan and on Wan2.2, whose Full costs are mainly suppressed motion and background reversal, to test whether the effect extends beyond a boundary discontinuity. On Hunyuan, where the discontinuity arises at this boundary, cross-seed donors test how much of the effect depends on matching the recipient trajectory. We use Wan2.2 for global softening, support area, and placement controls because its shared conditioning path applies the same weight ($w = 1$, α, or M_0) to model input and clean-prediction projection, allowing these factors to vary within one mechanism.

5.1 Localization of Terminal Discontinuity. Unlike Wan2.2 and SANA, Hunyuan Full sharply increases HC without reducing DD (Section 4.2), and the abrupt change is concentrated at the terminal boundary: the adjacent-frame pair entering Hunyuan's terminal latent slot is the minimum-similarity pair in 95.9% of Full videos, versus 1.7% for Endpoint-free (Appendix E.1). Full therefore preserves ordinary motion while introducing a localized discontinuity where the endpoint is imposed. If this discontinuity depends on what the endpoint imposes, not only on where, replacing terminal content at fixed support should reduce it.

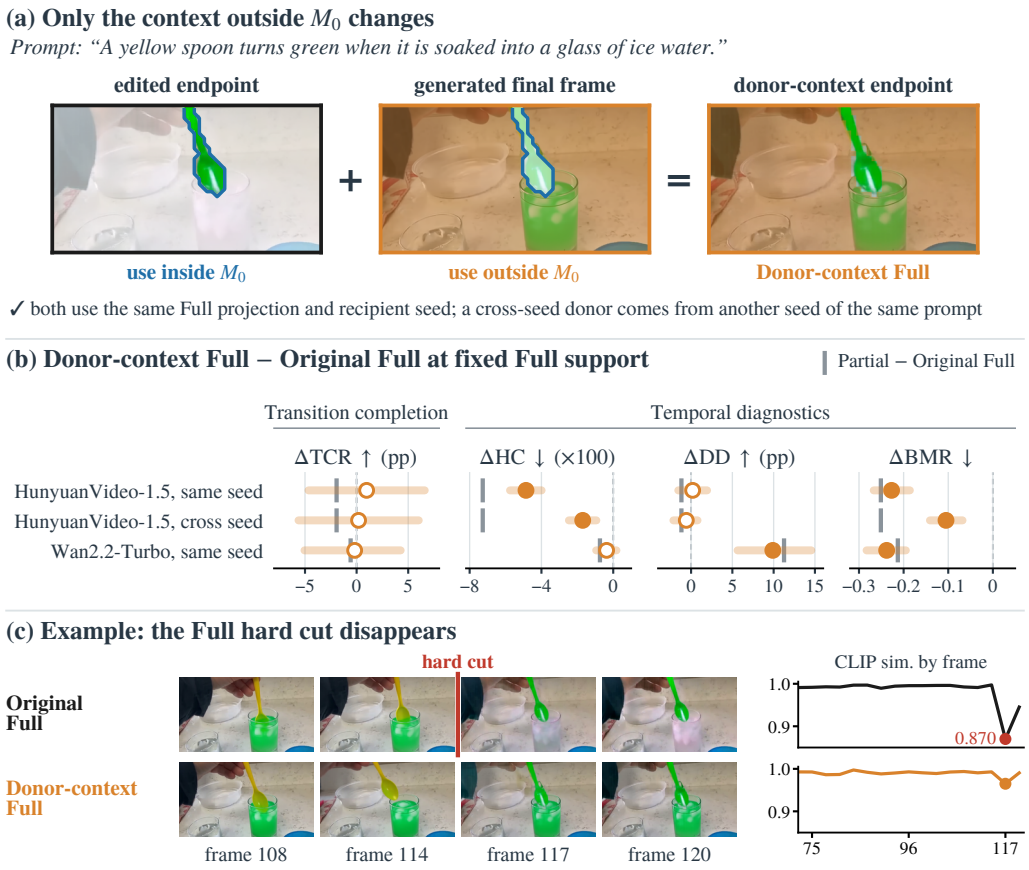


Figure 3: **Terminal content, not only constraint extent, affects temporal artifacts.** (a) The donor-context endpoint keeps the edited endpoint inside M_0 (blue outline) and uses the final frame of a paired Partial rollout outside it (orange tint). Original and Donor-context Full share the Full projection and the recipient seed; a cross-seed donor comes from another seed of the same prompt. (b) Donor-context minus Original Full with paired 95% CIs over 103 prompts and five seeds; open markers are not significant; gray ticks mark Partial minus Original Full. (c) HunyuanVideo-1.5, prompt 28, seed 22004: native frames and consecutive-frame CLIP similarity; the Original Full hard cut between frames 114 and 117 is absent under Donor-context Full.

5.2 Donor Context under Fixed Full Support. A Full–Partial comparison changes both imposed content and constraint extent, so we vary terminal content alone. Original Full (O) conditions on the endpoint latent z_E. A paired Partial rollout (P) with the same source, prompt, and seed supplies its final RGB frame I_T^P. Donor-context Full (D) combines original content inside estimated support with donor context outside:

$$z_D = M_0 \odot z_E + (1 - M_0) \odot E(I_T^P), \tag{5}$$

where E is the backbone's terminal-frame encoding, also used for z_E (Appendix D). Here M_0 selects the source of terminal content, not projection support. Within each backbone, O and D constrain all terminal spatial positions with the same Full support and conditioning procedure, sampler, source, prompt, and paired noises. The donor requires an extra rollout; it is not ground-truth future context and need not fit the new Full rollout.

Terminal content affects temporal artifacts at fixed extent. On Hunyuan, D significantly lowers HC from 12.96 to 8.10 and BMR from 0.585 to 0.358 (relative reductions of 37.5% and 38.8%), while TCR and DD remain comparable. On Wan2.2, D significantly lowers BMR by 0.238 and raises DD by 9.90 pp, while TCR and HC remain comparable; TC-Score decreases slightly (Fig. 3b; Appendix D.5). Replacing terminal context at fixed Full support thus affects more than Hunyuan's boundary discontinuity.

Cross-seed donors. To test whether compatibility with the same seed alone explains the donor effect, we take each Hunyuan donor from another seed of the same prompt under an assignment

Table 2: Selected Wan2.2-Turbo controls: paired comparisons with main Partial (TCR 42.91%). A is original support area. All conditions and intervals: Tables C2–C5.

	Transition completion	Temporal diagnostics		
Variant	ΔTCR ↑	ΔHC ↓	ΔDD ↑	ΔBMR ↓
Global softening: Partial − Soft				
Soft ($\alpha = 0.75$)	$+0.58$	-1.49^*	$+7.38^*$	-0.224^*
Spatial controls: Variant − Partial				
$0.5A$	-3.50	-0.53^*	$+0.39$	-0.017^*
$1.5A$	-0.97	$+0.00$	$+0.19$	$+0.036^*$
Shifted M_0	-6.60^*	-0.44^*	$+0.58$	$+0.041^*$
Random A #1	-5.24^*	-0.04	-1.36	$+0.028^*$

fixed before generation, keeping the recipient seed and its noises (Appendix D.2). Cross-seed donors still significantly lower HC by 1.70 and BMR by 0.105, with TCR and DD comparable, but these reductions are significantly smaller than with same-seed donors (Fig. 3b). The improvement therefore does not depend entirely on same-seed pairing.

Interpretation. Constraint extent alone does not account for the observed temporal artifacts; terminal content also matters. Donor position/visibility, generated-image properties, and trajectory compatibility change jointly. The larger effect with same-seed donors is consistent with a contribution from compatibility with the recipient trajectory, but the individual contributions remain unidentified. Appendix D gives secondary donor-relative and native-latent contrasts. We next examine which original endpoint constraints to retain without a donor continuation.

5.3 Global Softening vs. Spatial Selection. Global softening tests whether weakening every terminal constraint can reproduce Partial's reductions in temporal artifacts; spatial controls separately vary how many and which positions are constrained. All comparisons are paired with the main Partial, and all conditions retain the original edited endpoint.

At $\alpha = 0.75$, TCR is comparable ($+0.58$ pp), but Partial yields significantly higher DD and lower HC and BMR (Table 2). The same directions hold at $\alpha = 0.5$ and 0.25 (Tables C2 and C3). The tested global softening therefore does not reproduce these reductions.

5.4 Support Area Analysis. For original support area $A = \sum_u M_0(u)$, we test $0.5A$, $0.75A$, $1.25A$, and $1.5A$. Distance-based erosion retains interior cells, while expansion adds nearby exterior cells. This varies constraint extent around the same edit-relevant region, with the endpoint and conditioning rule fixed.

Reducing area to $0.5A$ or $0.75A$ lowers HC, and $0.5A$ also lowers BMR, while TCR remains comparable. Enlarging to $1.25A$ or $1.5A$ increases BMR, consistent with constraining more preserved context; $1.25A$ also lowers TCR (Tables 2, C4, and C5). No tested area is best on every metric.

5.5 Support Placement Analysis. To test whether edit-relevant placement matters beyond area, we relocate M_0 away from its original position while preserving its cell count, and compare three random supports targeting area A. Construction details are in Appendix A.4.

At matched area, shifted support lowers TCR by 6.60 pp relative to edit-relevant placement, and all three random supports also lower it significantly (Tables 2 and C4). Shifted support lowers HC while increasing BMR, and DD remains comparable for shifted and random supports. Completion-favorable placement thus need not minimize temporal diagnostics.

5.6 Diagnostic Cross-Checks. Abrupt transitions. Against blinded Qwen3.6-35B-A3B judgments, HC has AUROC 0.777 (95% CI $[0.70, 0.84]$) on 1,619 evaluable videos from nine generation conditions, excluding Hunyuan.[2] The judge sees all native frames with method identity hidden and identifies abrupt changes lacking a continuous transition. On the Wan2.2 validation subset, judged event incidence is 0.0% for Endpoint-free, 8.3% for Full, and 5.0% for Partial, supporting the direction of the metric comparison. Protocol and sensitivity checks are reported in Appendix E.

Reversal and return. On Wan2.2, CoTracker3-based Track Reversal Ratio (TRR) and Back-Tracking Ratio (BTR) cross-check BMR (Karaev et al., 2025). TRR measures late motion opposing an early

[2]Official model card: Qwen3.6-35B-A3B.

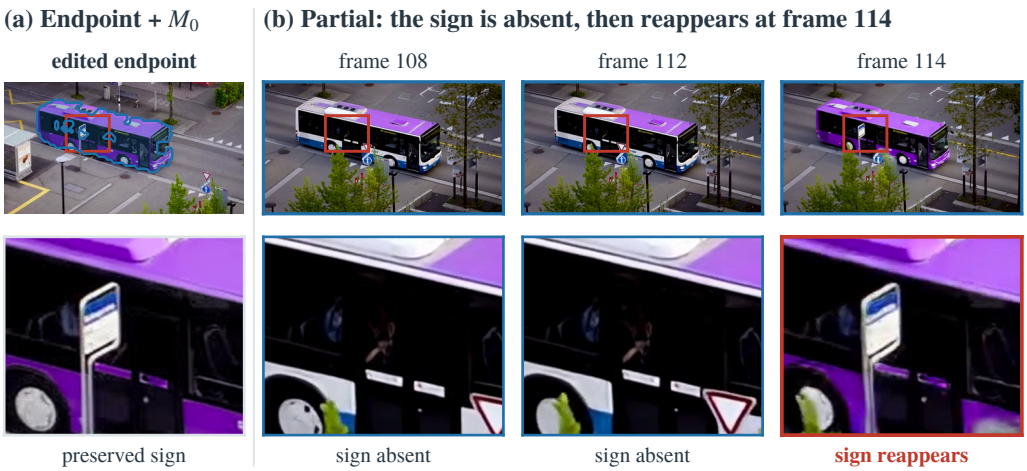


Figure 4: **Estimated edit-relevant support does not guarantee a future-valid terminal constraint.** Wan2.2-Turbo, purple-bus case 90001, seed 22001. (a) Saved M_0 outline (blue) and token-pooled weights (tint). (b) Partial: the boxed sign is absent at native frames 108 and 112, then reappears at 114. Identical regions are enlarged below.

direction along tracked trajectories; BTR measures return from the largest achieved displacement to the last visible position. On the common 398/515 prompt–seed pairs meeting tracking and motion criteria under all three conditions, all three diagnostics follow Full $>$ Partial $>$ Endpoint-free (Table E3). BMR's pooled Spearman correlations with TRR and BTR are 0.64 and 0.65: descriptive agreement among dependent, paired observations on this jointly measurable subset. Main BMR uses all 515 samples per condition (Appendix E).

6 Discussion and Limitations

Benefits and costs of partial goals. In the evaluated settings, imposing only edit-relevant support reduces discontinuity and background reversal and alleviates motion suppression relative to Full at comparable TCR, which global softening does not reproduce. The costs are equally specific: Partial does not raise TCR over Full, retains higher HC than Endpoint-free, raises BMR as support grows, and depends on edit-relevant placement (Sections 4.3–5.5).

Future validity. Edit relevance is a useful support heuristic, not a future-validity estimator. In Fig. 4, an endpoint-preserved sign reappears under Partial after being absent in earlier frames, illustrating how retained position and visibility constraints can become inappropriate.

Generalization and evaluation scope. We primarily use FLUX.2 endpoints across three backbone–conditioning settings, as FLUX.2 scored highest on TCR and TC-Score among tested editors (Table A1; Fig. A2). Donor-context tests cover Hunyuan and Wan2.2; cross-seed donors only Hunyuan, and spatial controls only Wan2.2. Partial remains untested with learned endpoint conditioners. Diagnostics measure specific temporal failures rather than overall video quality.

7 Conclusion

Edited endpoints preserve unedited context in the source image's layout, and full-frame conditioning imposes that context as a future requirement, a spatial over-specification. Measured with HC, BMR, and DD, the resulting temporal artifacts appear across five video models spanning training-free and learned endpoint conditioning, in forms that differ by model. Replacing only the preserved context under unchanged Full support reduces these artifacts on two backbones, partly even with donors from other seeds, so constraint extent alone does not account for them. Imposing only edit-relevant support accordingly mitigates them at comparable TCR, which global softening does not reproduce, while support area and placement trade completion against temporal diagnostics. Residual failures expose the gap between edit relevance and future validity: beyond how strongly to condition on an edited endpoint, the design question is which parts of its content should constrain the future.

AI USE STATEMENT

In this work, we used generative AI tools for three purposes: research execution, specifically assisting with method implementation and object/edit phrase annotation; retrieval and discovery of related work; and polishing writing, limited to sentence-level editing such as correcting typos and grammar and revising wording and style. We have reviewed all AI-assisted work: the authors made the final design decisions, reviewed and tested AI-assisted code, read retrieved papers in their original sources before citing them, and checked all edited text. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All models, benchmarks, and judges we use are publicly available. Appendix A specifies endpoint generation, attention aggregation, GMM–Ising support estimation, per-backbone sampler configurations (Table A2), mask transforms, and terminal projections; Appendix A.4 describes the support-size and placement controls. Appendix B defines every metric, sampling rule, threshold, and judge setting, and Table B1 lists each evaluation population. Appendix C gives the paired prompt bootstrap, including the number of draws and the random seed; Appendix D details the donor-context intervention; and Appendices E.2 and E.3 give the judge-agreement and track protocols. All generations use seeds 22001–22005. The supplementary material contains anonymized code for support estimation, terminal projection on each backbone, evaluation, and bootstrap analysis, together with the FLUX.2 edited endpoints, object/edit phrase annotations, estimated supports, and per-video metric values and judge outputs. Generated videos are not included; they can be regenerated with the released code and seeds, up to the sampler noise noted in the supplementary `VALIDATION.md`.

REFERENCES

Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended Latent Diffusion. *ACM Transactions on Graphics*, 42(4):149:1–149:11, 2023. doi: 10.1145/3592450. URL https://doi.org/10.1145/3592450.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, et al. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Minghong Cai, Qiulin Wang, Zongli Ye, Wenze Liu, Quande Liu, Weicai Ye, Xintao Wang, Pengfei Wan, Kun Gai, and Xiangyu Yue. VideoCanvas: Unified Video Completion from Arbitrary Spatiotemporal Patches via In-Context Conditioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, 2026. URL https://arxiv.org/abs/2510.08555.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22560–22570, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Cao_MasaCtrl_Tuning-Free_Mutual_Self-Attention_Control_for_Consistent_Image_Synthesis_and_ICCV_2023_paper.html.

Junsong Chen, Yuyang Zhao, Jincheng Yu, Ruihang Chu, Junyu Chen, Shuai Yang, Xianbang Wang, Yicheng Pan, Daquan Zhou, Huan Ling, Haozhe Liu, Hongwei Yi, Hao Zhang, Muyang Li, Yukang Chen, Han Cai, Sanja Fidler, Ping Luo, Song Han, and Enze Xie. SANA-Video: Efficient Video Generation with Block Linear Diffusion Transformer. In *International Conference on Learning Representations*, pp. 38973–38997, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/41b93c59da0d0f835907fd661d419db2-Paper-Conference.pdf.

Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. SEINE: Short-to-Long Video Diffusion Model for Generative Transition and Prediction. In *International Conference on Learning Representations*, pp. 52208–52222, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/e54e6eef11f87a874bf1e4551fc6d04e-Paper-Conference.pdf.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. DiffEdit: Diffusion-based semantic image editing with mask guidance. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=3lge0p5o-M-.

Mohamed Elmoghany, Liangbing Zhao, Xiaoqian Shen, Subhojyoti Mukherjee, Yang Zhou, Gang Wu, Viet Dac Lai, Seunghyun Yoon, Ryan Rossi, Abdullah Rashwan, Puneet Mathur, Varun Manjunatha, Daksh Dangi, Chien Nguyen, Nedim Lipka, Trung Bui, Krishna Kumar Singh, Ruiyi Zhang, Xiaolei Huang, Jaemin Cho, Yu Wang, Namyong Park, Zhengzhong Tu, Hongjie Chen, Hoda Eldardiry, Nesreen Ahmed, Thien Nguyen, Dinesh Manocha, Mohamed Elhoseiny, and Franck Dernoncourt. InfinityStory: Unlimited Video Generation with World Consistency and Character-Aware Shot Transitions. *arXiv preprint arXiv:2603.03646*, 2026. URL https://arxiv.org/abs/2603.03646.

Haiwen Feng, Zheng Ding, Zhihao Xia, Simon Niklaus, Victoria Abrevaya, Michael J. Black, and Xuaner Zhang. Explorative Inbetweening of Time and Space. In *European Conference on Computer Vision*, pp. 378–395, 2024. doi: 10.1007/978-3-031-73229-4_22. URL https://doi.org/10.1007/978-3-031-73229-4_22.

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/.

Xingtong Ge, Yi Zhang, Yushi Huang, Dailan He, Xiahong Wang, Bingqi Ma, Guanglu Song, Yu Liu, and Jun Zhang. SALT: Self-Consistent Distribution Matching with Cache-Aware Training for Few-Step Video Generation. In *European Conference on Computer Vision*, 2026. URL https://eccv.ecva.net/virtual/2026/poster/3473.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-Prompt Image Editing with Cross-Attention Control. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=_CDixzkzeyb.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_VBench_Comprehensive_Benchmark_Suite_for_Video_Generative_Models_CVPR_2024_paper.html.

Sangwon Jang, Taekyung Ki, Jaehyeong Jo, Jaehong Yoon, Soo Ye Kim, Zhe Lin, and Sung Ju Hwang. Frame Guidance: Training-Free Guidance for Frame-Level Control in Video Diffusion Models. In *International Conference on Learning Representations*, pp. 60977–61003, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/636f1f78f7096caf126afaa27deb7b53-Paper-Conference.pdf.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-One Video Creation and Editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17191–17202, 2025. URL https://arxiv.org/abs/2503.07598.

Zhangqi Jiang, Zheng Sun, Xianfang Zeng, Yufeng Yang, Xuanyang Zhang, Yongliang Wu, Wei Cheng, Gang Yu, Xu Yang, and Bihan Wen. GEditBench v2: A Human-Aligned Benchmark for General Image Editing. *arXiv preprint arXiv:2603.28547*, 2026. URL https://arxiv.org/abs/2603.28547.

Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker3: Simpler and Better Point Tracking by Pseudo-Labelling Real Videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6013–6022, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Karaev_CoTracker3_Simpler_and_Better_Point_Tracking_by_Pseudo-Labelling_Real_Videos_ICCV_2025_paper.html.

Kinam Kim, Junha Hyung, and Jaegul Choo. Temporal In-Context Fine-Tuning with Temporal Reasoning for Versatile Control of Video Diffusion Models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 74491–74518, 2025. doi: 10.52202/085713-2501. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/6c03ebd7344bcc0c12c4ba8aa8487c00-Paper-Conference.pdf.

Max Ku, Cong Wei, Weiming Ren, Harry Yang, and Wenhu Chen. AnyV2V: A Tuning-Free Framework For Any Video-to-Video Editing Tasks. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=RFrJCkw2oa.

Ruibin Li, Tao Yang, Yangming Shi, Weiguo Feng, Shilei Wen, Bingyue Peng, and Lei Zhang. Many-for-Many: Unify the Training of Multiple Video and Image Generation and Manipulation Tasks. In *International Conference on Learning Representations*, pp. 59233–59269, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/60f9118a849e8e9a0c67e2a36ad80ebf-Paper-Conference.pdf.

Yehonathan Litman, Shikun Liu, Dario Seyb, Nicholas Milef, Yang Zhou, Carl Marshall, Shubham Tulsiani, and Caleb Leak. EditCtrl: Disentangled Local and Global Control for Real-Time Generative Video Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8965–8975, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Litman_EditCtrl_Disentangled_Local_and_Global_Control_for_Real-Time_Generative_Video_CVPR_2026_paper.html.

Shaoteng Liu, Tianyu Wang, Jui-Hsien Wang, Qing Liu, Zhifei Zhang, Joon-Young Lee, Yijun Li, Bei Yu, Zhe Lin, Soo Ye Kim, and Jiaya Jia. Generative Video Propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17712–17722, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Liu_Generative_Video_Propagation_CVPR_2025_paper.html.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver++: Fast Solver for Guided Sampling of Diffusion Probabilistic Models. *Machine Intelligence Research*, 22 (4):730–751, 2025. doi: 10.1007/s11633-025-1562-4.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting Using Denoising Diffusion Probabilistic Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11461–11471, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Lugmayr_RePaint_Inpainting_Using_Denoising_Diffusion_Probabilistic_Models_CVPR_2022_paper.html.

Haoyu Ma, Shahin Mahdizadehaghdam, Bichen Wu, Zhipeng Fan, Yuchao Gu, Wenliang Zhao, Lior Shapira, and Xiaohui Xie. MaskINT: Video Editing via Interpolative Non-autoregressive Masked Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7403–7412, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Ma_MaskINT_Video_Editing_via_Interpolative_Non-autoregressive_Masked_Transformers_CVPR_2024_paper.html.

Qi Mao, Lan Chen, Yuchao Gu, Zhen Fang, and Mike Zheng Shou. MAG-Edit: Localized Image Editing in Complex Scenarios via Mask-Based Attention-Adjusted Guidance. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 6842–6850, 2024. doi: 10.1145/3664647.3680830. URL https://doi.org/10.1145/3664647.3680830.

Matiur Rahman Minar, Seunghun Oh, GangHyeon Jeong, and Unsang Park. Steady-Forcing: Balancing Spatial Persistence and Motion Continuity in Long-Horizon Nature Video Diffusion. *arXiv preprint arXiv:2606.14732*, 2026. URL https://arxiv.org/abs/2606.14732.

Junyi Ouyang, Wenbin Teng, Gonglin Chen, Yajie Zhao, and Haiwei Chen. DCARL: A Divide-and-Conquer Framework for Autoregressive Long-Trajectory Video Generation. In *European Conference on Computer Vision*, 2026. URL https://eccv.ecva.net/virtual/2026/poster/4282.

Wenqi Ouyang, Yi Dong, Lei Yang, Jianlou Si, and Xingang Pan. I2VEdit: First-Frame-Guided Video Editing via Image-to-Video Diffusion Models. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024. URL https://doi.org/10.1145/3680528.3687656.

Kaihang Pan, Qi Tian, Jianwei Zhang, Weijie Kong, Jiangfeng Xiong, Yanxin Long, et al. OmniWeaving: Towards Unified Video Generation with Free-form Composition and Reasoning. *arXiv preprint arXiv:2603.24458*, 2026. URL https://arxiv.org/abs/2603.24458.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *International Conference on Learning Representations*, pp. 1862–1874, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/081b08068e4733ae3e7ad019fe8d172f-Paper-Conference.pdf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Zachary Teed and Jia Deng. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In *European Conference on Computer Vision*, 2020. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/html/3526_ECCV_2020_paper.php.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, et al. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Xiaojuan Wang, Boyang Zhou, Brian Curless, Ira Kemelmacher-Shlizerman, Aleksander Holynski, and Steve Seitz. Generative Inbetweening: Adapting Image-to-Video Models for Keyframe Interpolation. In *International Conference on Learning Representations*, pp. 30594–30609, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/4bbdef62653d8088717640e7660a1ebb-Paper-Conference.pdf.

Yihan Wang and Jia Deng. WAFT: Warping-Alone Field Transforms for Optical Flow. In *International Conference on Learning Representations*, pp. 157389–157402, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/ff2853ed92a6e6983ecccd9c922a42d2-Paper-Conference.pdf.

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, et al. HunyuanVideo 1.5 Technical Report. *arXiv preprint arXiv:2511.18870*, 2025a. URL https://arxiv.org/abs/2511.18870.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, et al. Qwen-Image Technical Report. *arXiv preprint arXiv:2508.02324*, 2025b. URL https://arxiv.org/abs/2508.02324.

Jay Zhangjie Wu, Xuanchi Ren, Tianchang Shen, Tianshi Cao, Kai He, Yifan Lu, Ruiyuan Gao, Enze Xie, Shiyi Lan, Jose M. Alvarez, Jun Gao, Sanja Fidler, Zian Wang, and Huan Ling. ChronoEdit: Towards Temporal Reasoning for In-Context Image Editing and World Simulation. In *International Conference on Learning Representations*, pp. 9886–9899, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/10b99813d5992673da90a499b6bcd5f0-Paper-Conference.pdf.

Serin Yang, Taesung Kwon, and Jong Chul Ye. ViBiDSampler: Enhancing Video Interpolation Using Bidirectional Diffusion Sampler. In *International Conference on Learning Representations*, pp. 3441–3459, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/098a3d992f0d2128bab619b3c654e1c3-Paper-Conference.pdf.

Zuhao Yang, Jiahui Zhang, Yingchen Yu, Shijian Lu, and Song Bai. Versatile Transition Generation with Image-to-Video Diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16981–16990, 2025b. URL https://openaccess.thecvf.com/content/ICCV2025/html/Yang_Versatile_Transition_Generation_with_Image-to-Video_Diffusion_ICCV_2025_paper.html.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

Min Zhao, Hongzhou Zhu, Chendong Xiang, Kaiwen Zheng, Chongxuan Li, and Jun Zhu. Identifying and Solving Conditional Image Leakage in Image-to-Video Diffusion Model. In *Advances in Neural Information Processing Systems*, volume 37, pp. 30300–30326, 2024. doi: 10.52202/079017-0954. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/35cb54b887e7aafe74829677cce6c5c6-Paper-Conference.pdf.

Siyu Zou, Jiji Tang, Yiyi Zhou, Jing He, Chaoyi Zhao, Rongsheng Zhang, Zhipeng Hu, and Xiaoshuai Sun. Towards Efficient Diffusion-Based Image Editing with Instant Attention Masks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7864–7872, 2024. doi: 10.1609/aaai.v38i7.28622. URL https://ojs.aaai.org/index.php/AAAI/article/view/28622.

Appendix

Contents

A Implementation Details 16
A.1 Endpoint and Support Construction 16
A.2 Choice of Endpoint Editor 16
A.3 Video Configurations and Projection 18
A.4 Support-Size and Placement Controls 19
B Metric Definitions and Evaluation Populations 19
B.1 Transition Completion and TC-Score 19
B.2 Dynamic Degree and Hard-Cut Score 20
B.3 Background Motion Reversal 20
B.4 Evaluation Populations 20
C Paired Uncertainty and Additional Results. 20
C.1 Resampling Protocol 20
C.2 Global-Softening and Spatial-Control Results 21
C.3 Additional Endpoint Comparisons 21
D Donor-Context Intervention and Score Decomposition 24
D.1 Construction, Pairing, and Scope 24
D.2 Cross-Seed Donors on Hunyuan 24
D.3 Secondary Donor and Representation Contrasts 24
D.4 Descriptive Closure of the Full–Partial Gap 25
D.5 TC-Score Decomposition and Its Relation to HC 26
E Diagnostics and Cross-Checks 26
E.1 Hunyuan Discontinuity Localization 26
E.2 Blinded Model-Judge Agreement for HC 27
E.3 Track Reversal Ratio and Back-Tracking Ratio 27
E.4 Source-Time Layout of Preserved Context 28
F Residual Failures and Exploratory Support Release 29
G Broader I2V Completion Comparisons 30

A Implementation Details

A.1 Endpoint and Support Construction

Endpoint generation. We generate one endpoint per prompt with FLUX.2-klein-base-9B, seed 42, 50 editing steps, bfloat16 inference, and default guidance. We center-crop the source to the target aspect ratio and resize it with Lanczos to 1280×704. We use the following final-frame prompt:

> Use the input image as the starting frame of a video. The following sentence describes a transition that happens over the course of the video: "{transition}"
>
> Generate a plausible final frame of that video: the moment right after the transition described above has fully completed.
>
> The frame should look like a later moment from the same continuous shot.
>
> Preserve scene continuity, identity, camera framing, and all persistent background/content, changing only what the transition describes.

Attention aggregation. We average attention over the annotated object and edit phrases to obtain a_o and a_e. For endpoint-image tokens, text-to-image and image-to-text attention receive equal weight; we average all heads, 32 transformer blocks, and 50 editing steps in the conditional pass. Reference-image tokens are excluded. We combine the normalized maps as

$$a(u) = \tfrac{1}{2}\left[\mathcal{N}(a_o)(u) + \mathcal{N}(a_e)(u)\right], \qquad \mathcal{N}(a)(u) = \frac{a(u) - \min_v a(v)}{\max(\max_v a(v) - \min_v a(v), 10^{-12})}. \tag{A1}$$

We estimate support on the 44×80 latent grid (height $\times$ width) using a two-component Gaussian mixture with shared variance:

$$p(a(u)) = \sum_{j \in \{0,1\}} \pi_j \mathcal{N}(a(u); \mu_j, \sigma^2), \qquad \mu_0 < \mu_1. \tag{A2}$$

We initialize the means at 25% and 75% of the observed value range, with equal mixture weights and empirical variance. Variance floors are 10^{-6} at initialization and 10^{-8} during updates. Expectation–maximization runs for up to 300 iterations, stopping when both means satisfy $|\mu_j^{\text{new}} - \mu_j| \le 10^{-12} + 10^{-5}|\mu_j|$. The high-mean component represents edit-relevant evidence. We encourage spatial coherence with

$$\mathcal{E}(M) = \sum_u \left[\frac{(a(u) - \mu_{M(u)})^2}{2\sigma^2} - \log \pi_{M(u)}\right] + \beta \sum_{(u,v)\in\mathcal{E}_4} \mathbf{1}[M(u) \neq M(v)]. \tag{A3}$$

Here $\mathcal{E}_4$ is the four-neighbor grid and $\beta = 2$. We initialize labels by unary cost and apply up to 12 synchronous conditional-mode updates with edge padding, stopping when labels no longer change. Figure A1 shows the resulting supports.

A.2 Choice of Endpoint Editor

Partial needs only an edited endpoint and a support estimated from the editor's text–image attention (Appendix A), so the primary editor must be open-weight with accessible attention. We use FLUX.2-klein-base-9B for its editing quality, completion, and layout preservation (Table A1).

Editing quality. On GEditBench v2, FLUX.2 [klein] 9B ranks first among open-weight editors with an Elo of 1,039, on par with Qwen-Image-Edit-2511 at 1,038 (Jiang et al., 2026). That evaluation uses the 4-step distilled checkpoint; we use its undistilled base, whose 50-step sampler provides the attention we aggregate.

Completion and temporal diagnostics. Under Full conditioning on Wan2.2, FLUX.2 endpoints give the highest TCR of the three editors with the fewest parameters. Paired TCR differences are comparable, $+1.75$ $[-3.69, +7.38]$ pp versus ChronoEdit (Wu et al., 2026) and $+6.80$ $[-0.39, +13.98]$ pp versus Qwen-Image-Edit (Wu et al., 2025b), but only FLUX.2 endpoints significantly raise TCR over Endpoint-free (Table C6). TC-Score is significantly higher

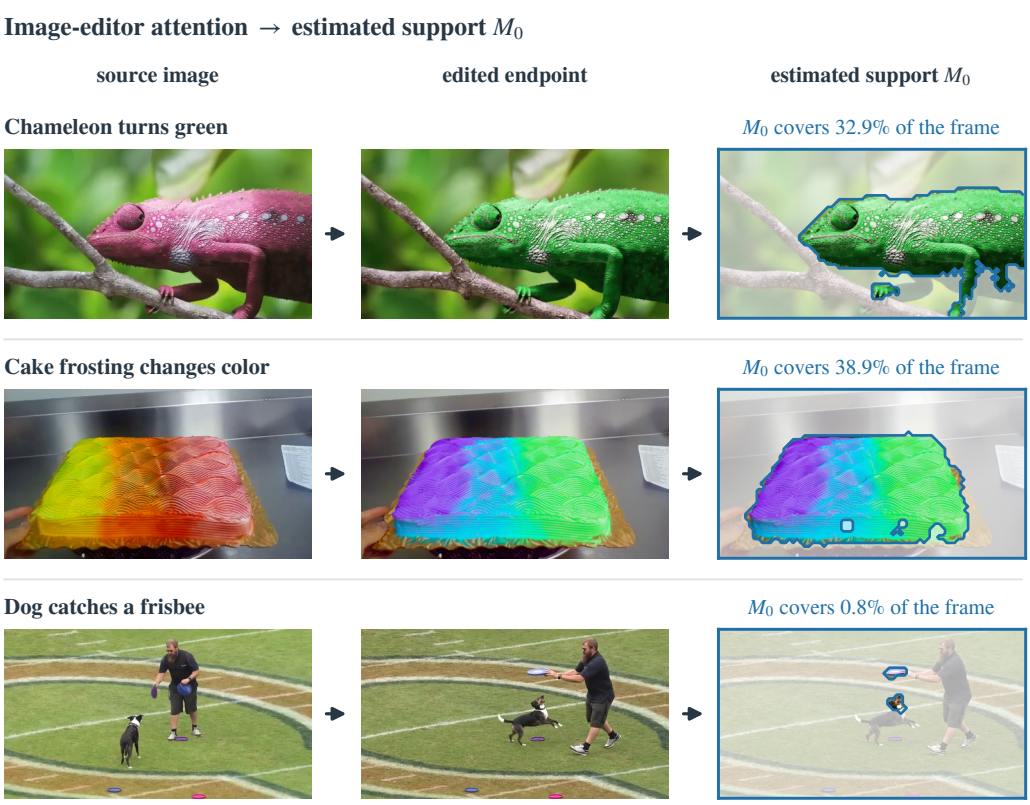


Figure A1: **Estimated edit support.** Source images, edited endpoints, and GMM–Ising masks for TC-Bench prompts 0, 5, and 82. Blue marks support; percentages give constrained area before token pooling.

than with Qwen-Image-Edit endpoints, $+0.035$ $[+0.018, +0.052]$, and comparable with ChronoEdit endpoints, $+0.014$ $[-0.001, +0.029]$. Full's temporal artifacts are not specific to FLUX.2: relative to Endpoint-free, all three editors significantly raise HC, and Qwen-Image-Edit endpoints also change BMR by $+0.256$ $[+0.207, +0.307]$ and DD by -15.34 $[-20.19, -10.68]$ pp. HC, DD, and BMR are comparable between FLUX.2 and Qwen-Image-Edit endpoints.

Layout preservation. FLUX.2 endpoints differ from the source least outside M_0 and most inside it. The outside difference is lower than for Qwen-Image-Edit, -0.022 $[-0.035, -0.008]$, and ChronoEdit, -0.038 $[-0.056, -0.023]$. Because M_0 is estimated from FLUX.2 attention, we also measure Qwen-Image-Edit endpoints outside the support estimated from its own attention; they still differ more, 0.080 versus 0.056, a paired difference of -0.024 $[-0.038, -0.010]$. FLUX.2 endpoints thus most closely follow Eq. (1), making Full's temporal artifacts less likely to stem from unrequested changes outside the edit.

Qualitative examples. In Figure A2, FLUX.2 and Qwen-Image-Edit recolor the chameleon while keeping its pose and texture, whereas ChronoEdit also changes its appearance. For the sunset, FLUX.2 and ChronoEdit lower the sun while keeping the skyline, whereas Qwen-Image-Edit returns a four-panel collage. For the meteor, FLUX.2 extends the streak toward the lower right, Qwen-Image-Edit leaves the source nearly unchanged, and ChronoEdit renders a spray of streaks near the upper left.

Other editors. Support estimation uses only the endpoint and the attention to the annotated object and edit phrases. Open-weight diffusion and flow editors expose this attention in joint or cross-attention layers, so we expect the procedure to extend to them without changing the video model. Adapting our attention recorder to Qwen-Image-Edit's token layout, with the same GMM–Ising rule, yields supports for all 103 prompts, covering 6.3% of the frame on average versus 8.9% for FLUX.2.

Table A1: Endpoint editors under Full conditioning on Wan2.2-Turbo (103 prompts × five paired seeds). Out and In give the mean absolute RGB difference between the endpoint and the source image outside and inside the FLUX.2 support M_0. ChronoEdit uses the original transition prompt and the last frame of a five-frame generation (Appendix G).

		Full on Wan2.2-Turbo					Endpoint vs. source	
Editor	Params	TCR ↑	TC-Score ↑	HC ↓	DD ↑	BMR ↓	Out	In
FLUX.2-klein-base-9B	9B	41.94	0.8352	6.47	66.8	0.599	0.056	0.232
Qwen-Image-Edit-2511	20B	35.15	0.8001	7.11	64.1	0.617	0.077	0.161
ChronoEdit-14B	14B	40.19	0.8214	6.13	–	–	0.094	0.192
Endpoint-free	–	33.79	0.7841	4.57	79.4	0.360	–	–

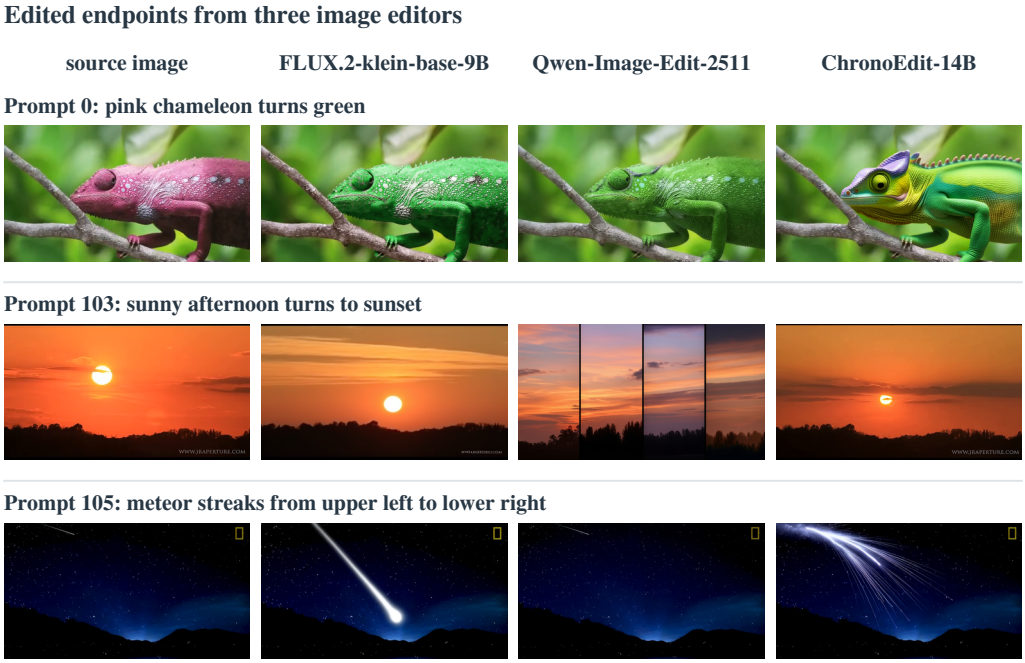


Figure A2: **Endpoints from three editors.** Source images and edited endpoints for TC-Bench prompts 0, 103, and 105. FLUX.2 and Qwen-Image-Edit use the final-frame prompt in Appendix A; ChronoEdit uses the original transition prompt to match its formulation of editing as video generation.

We did not evaluate Partial with these supports; editors without attention access, such as closed APIs, would need supports from another source.

A.3 Video Configurations and Projection

We use seeds 22001–22005 and share each edited endpoint between Full and Partial. Wan2.2 uses the distilled TI2V-5B checkpoint; SANA uses the 2B 720p checkpoint with guidance scale 6.0.

Wan2.2 average-pools M_0 in 2×2 blocks from the 44×80 latent grid to a 22×40 token grid and repeats the weights over the corresponding latent cells. SANA area-pools to its 22×40 latent/token grid. Pooled weights can be fractional; they are broadcast over latent channels, with token timesteps scaled by $1-w$. At each step, the terminal state x enters the Wan2.2 model as $x_{\text{in}} = w \odot z_E + (1-w) \odot x$, and the clean prediction is projected with Eq. (2) before the sampling update; both blends use the same w, including $w = \alpha$ for Soft.

SANA Partial keeps solver state x separate from blended model input y. At noise level σ, it projects the clean prediction $y - \sigma v$ with M_0 and converts it back to flow relative to x for the native DPM-Solver++ update (Lu et al., 2025). SANA Full instead clamps the terminal state at initialization and after each scheduler step, with zero terminal token timesteps. With $M_0 \equiv 1$, the Partial update

Table A2: Primary generation configurations. Spatial dimensions are width by height. HunyuanVideo-1.5 selects an aspect-ratio-compatible resolution using its 480p configuration.

Backbone	Steps	Frames	FPS	Resolution	Terminal condition
Wan2.2-Turbo	4	121	24	1280×704	Input/clean projection
SANA-Video 2B	50	81	16	1280×704	See wrappers below
HunyuanVideo-1.5	8	121	24	480p	Noise-matched projection

reduces to this clamp: from identical initial noise, the two implementations' final latents differ by at most 1.2×10^{-7} in two verification videos.

For Hunyuan, we resize and center-crop the source and mask to the selected 480p aspect ratio, then pool and threshold the mask at 0.5. We repeat the endpoint over five frames, encode it with the video VAE, and select the second temporal latent slot. Endpoint noise ϵ_E is sampled once from a seed derived from the video seed and held fixed. At noise level σ_k, masked projection targets

$$z_E^{(k)} = (1 - \sigma_k) z_E + \sigma_k \epsilon_E, \tag{A4}$$

at initialization and after each scheduler update, ending with clean projection at zero noise.

Spatial controls are paired with the main Partial. Additional completion comparisons use learned endpoint conditioning in MfM and OmniWeaving, and 50 sampling steps for Wan2.2 Base.

A.4 Support-Size and Placement Controls

For $A = \sum_u M_0(u)$, size controls target rounded areas $0.5A$, $0.75A$, $1.25A$, and $1.5A$. Euclidean distance transforms rank cells: erosion retains the deepest interior cells, while expansion adds the nearest exterior cells. Seeded random values break ties.

Shifted support preserves area while changing location. We sample in-bounds, even-coordinate translations with IoU below 0.1, or use the lowest-overlap candidate. When translation cannot preserve area, we clip and regrow the mask to its original cell count; this occurs for 12 of 103 prompts.

Random support selects $\mathrm{round}(A/4)$ cells on the 22×40 token grid and expands each to a 2×2 latent block. Area therefore matches M_0 to within two latent cells. We evaluate three independent random supports per prompt. These supports remain fixed during sampling; random release in Appendix F instead shrinks support over denoising steps.

B Metric Definitions and Evaluation Populations

Temporal sampling and preprocessing. For N native frames at FPS f, TC-Score uses unique indices $\mathrm{round}(k(N-1)/15)$, $k = 0, \ldots, 15$. HC samples $\mathrm{round}(kf/8)$; DD and BMR take every $\max(1, \mathrm{round}(f/8))$-th frame. All three temporal diagnostics thus use 8 FPS for the 16/24 FPS main configurations. CLIP uses its standard 224-pixel image preprocessing; flow uses decoded resolution. Track diagnostics use native 24 FPS frames resized to 640×352.

B.1 Transition Completion and TC-Score

We evaluate 50 attribute-binding, 38 object-binding, and 15 background-shift prompts using the benchmark assertions and Qwen2.5-VL-32B-Instruct. TCR requires Yes for every assertion except those about "other objects."

TC-Score combines a VLM assertion average, excluding transition-object consistency, with mean adjacent-frame CLIP ViT-B/32 similarity on the 16-frame grid. We map CLIP similarity from $[0.9, 0.98]$ to $[0, 1]$ and clip it. The VLM and CLIP terms receive weights $2/3$ and $1/3$, respectively. Adjacent-frame CLIP similarity falls with larger motion and abrupt changes, so the CLIP term mixes completion with properties that DD and HC measure, and HC uses the same similarity (Appendix D.5). We therefore treat TCR as the primary completion measure and report TC-Score for comparison with TC-Bench.

B.2 Dynamic Degree and Hard-Cut Score

DD follows VBench. For each adjacent pair, $m(t)$ is the mean of the largest 5% of RAFT flow magnitudes. A video is dynamic if $m(t) > 6\min(H, W)/256$ for at least $\mathrm{round}(4N/16)$ pairs, where N is the sampled frame count. The main clips have 41 frames at 8 FPS and require 10 qualifying pairs. DD measures substantial movement, including camera motion.

HC is the difference between the median and minimum adjacent-frame CLIP cosine similarities (Eq. (3)), an absolute rather than proportional drop. We report $100\times$HC, including its confidence intervals. The minimum-similarity pair t_{HC} identifies the event location.

B.3 Background Motion Reversal

BMR uses WAFT flow at 8 FPS outside the FLUX.2-derived support. We keep this measurement mask fixed across all conditions, resize it with nearest-neighbor interpolation, and take a deterministic strided subset of at most 80,000 pixels. Equation (4) weights flow by magnitude to prevent static pixels from dominating the direction estimate.

The reference axis is the normalized componentwise median flow over the first half of the video; reversal is measured over the second half. The ratio is undefined for a zero reference or zero late along-axis travel. All main and control conditions have 515 finite values, without DD filtering. We cross-check BMR with track diagnostics in Appendix E.3.

B.4 Evaluation Populations

Table B1 summarizes the cohorts used in each analysis.

Table B1: Appendix populations and references. Counts refer to paired prompt–seed positions unless stated otherwise.

Analysis	Population and reference
Main backbone comparisons	103 prompts $\times$ 5 seeds per condition; Table 1.
Wan2.2-Turbo spatial controls	103 prompts $\times$ 5 seeds per support, paired with the main Wan2.2-Turbo Partial (TCR 42.91%).
HunyuanVideo-1.5 donor-context intervention	Original Full reuses the main cohort. Donor P regenerates main Partial with identical settings; TCR 44.66% versus 44.47% reflects re-judging variation.
HunyuanVideo-1.5 cross-seed donors	O, same-seed D, and P are reused; 515 new D_{cross} videos use a donor-seed assignment fixed before generation.
Wan2.2-Turbo donor-context intervention	O, D, N, and P are new 515-video runs sharing one noise stream per prompt–seed; O reproduces main Full (HC 6.49 versus 6.47, BMR 0.600 versus 0.599), and P reproduces main Partial's HC, DD, and BMR exactly; TCR 42.72% versus 42.91% reflects re-judging variation.
HC model-judge agreement	90 prompts $\times$ 2 samples $\times$ 9 conditions; 1,619 evaluable videos. HunyuanVideo-1.5 is excluded.
Track cross-check	Common 398 pairs measurable in all three Wan2.2-Turbo conditions; main BMR uses all 515.

C Paired Uncertainty and Additional Results

C.1 Resampling Protocol

For a metric y, let y_{isc} denote its value for prompt i, seed s, and condition c. The paired difference between target a and reference b is

$$\widehat{\Delta}_{a-b} = \frac{1}{103}\sum_{i=1}^{103} d_i, \qquad d_i = \frac{1}{5}\sum_{s=1}^{5}(y_{isa} - y_{isb}). \tag{C1}$$

Table C1: Paired mean differences. Each bootstrap draw resamples prompts and carries all five seeds and conditions together. The direction is stated in every row.

	Transition completion	Temporal diagnostics		
Difference	ΔTCR ↑	ΔHC ↓	ΔDD ↑	ΔBMR ↓
Wan2.2-Turbo Full–Endpoint-free	$+8.16$ $[+1.75, +14.56]$	$+1.90$ $[+1.27, +2.59]$	-12.62 $[-17.48, -8.16]$	$+0.239$ $[+0.189, +0.289]$
Wan2.2-Turbo Partial–Full	$+0.97$ $[-3.50, +5.44]$	-0.72 $[-1.29, -0.17]$	$+9.90$ $[+5.63, +14.56]$	-0.213 $[-0.262, -0.165]$
Wan2.2-Turbo Partial–Endpoint-free	$+9.13$ $[+3.11, +15.34]$	$+1.18$ $[+0.72, +1.69]$	-2.72 $[-5.24, -0.39]$	$+0.027$ $[+0.004, +0.049]$
SANA-Video Full–Endpoint-free	$+12.23$ $[+5.44, +19.22]$	$+1.70$ $[+0.67, +2.70]$	-9.90 $[-15.53, -4.27]$	$+0.298$ $[+0.254, +0.343]$
SANA-Video Partial–Full	-0.39 $[-6.41, +5.44]$	-0.97 $[-1.93, +0.01]$	$+9.51$ $[+4.08, +15.15]$	-0.285 $[-0.325, -0.246]$
SANA-Video Partial–Endpoint-free	$+11.84$ $[+6.99, +16.89]$	$+0.73$ $[+0.17, +1.33]$	-0.39 $[-4.85, +4.08]$	$+0.013$ $[-0.012, +0.038]$
HunyuanVideo-1.5 Full–Endpoint-free	$+5.63$ $[0.00, +11.46]$	$+9.50$ $[+8.53, +10.51]$	$+1.17$ $[-1.55, +3.88]$	$+0.279$ $[+0.233, +0.326]$
HunyuanVideo-1.5 Partial–Full	-2.14 $[-7.57, +3.11]$	-7.27 $[-8.16, -6.40]$	-1.17 $[-3.50, +1.17]$	-0.252 $[-0.296, -0.208]$
HunyuanVideo-1.5 Partial–Endpoint-free	$+3.50$ $[-0.58, +7.57]$	$+2.23$ $[+1.65, +2.87]$	$+0.00$ $[-1.75, +1.75]$	$+0.028$ $[+0.008, +0.048]$

Table C2: Global softening versus Partial on Wan2.2-Turbo. Partial has higher DD and lower HC/BMR than each tested soft setting; all direct paired temporal intervals exclude zero (Table C3).

	Transition completion		Temporal diagnostics		
Condition	TCR ↑	TC-Score ↑	HC ↓	DD ↑	BMR ↓
Full ($\alpha = 1$)	41.94	0.835	6.47	66.8	0.599
Soft ($\alpha = 0.75$)	42.33	0.828	7.25	69.3	0.611
Soft ($\alpha = 0.5$)	39.03	0.816	7.62	69.9	0.630
Soft ($\alpha = 0.25$)	39.22	0.800	8.48	72.2	0.614
Partial (M_0)	42.91	0.819	5.76	76.7	0.387

We resample 103 prompts with replacement, retaining each prompt's five seeds and paired conditions. We use 100,000 draws with NumPy PCG64 seed 20260919 and take the 2.5th/97.5th percentiles. TCR and temporal metrics share the same draws; subgroup estimates retain the selected edit type and recompute its denominator. TCR/DD differences are in pp, HC uses its displayed scale, and BMR its original scale.

Intervals are pointwise, unadjusted, and conditional on these five seeds. A CI spanning zero does not establish equivalence or non-inferiority. We compute differences and ratios from unrounded per-video values.

C.2 Global-Softening and Spatial-Control Results

Tables C1, C3, and C5 report direct paired differences for the main, softening, and spatial comparisons.

C.3 Additional Endpoint Comparisons

Additional endpoint-free versus Full comparisons use MfM, OmniWeaving, and Wan2.2 Base (Li et al., 2026; Pan et al., 2026; Wan et al., 2025).

Table C3: Direct Partial minus Soft paired differences.

	Transition completion	Temporal diagnostics		
Soft α	ΔTCR $\uparrow$	ΔHC $\downarrow$	ΔDD $\uparrow$	ΔBMR $\downarrow$
0.75	$+0.58$ $[-4.66, +5.63]$	-1.49 $[-2.23, -0.84]$	$+7.38$ $[+3.88, +11.07]$	-0.224 $[-0.272, -0.177]$
0.5	$+3.88$ $[-1.75, +9.51]$	-1.86 $[-2.67, -1.16]$	$+6.80$ $[+2.91, +10.87]$	-0.243 $[-0.292, -0.195]$
0.25	$+3.69$ $[-1.75, +9.13]$	-2.72 $[-3.60, -1.95]$	$+4.47$ $[+0.78, +8.35]$	-0.227 $[-0.273, -0.182]$

Table C4: Wan2.2-Turbo support-size and placement controls. Differences are paired with the main Partial.

	Transition completion		Temporal diagnostics		
Support	TCR $\uparrow$	ΔTCR $\uparrow$ [95% CI]	HC $\downarrow$	DD $\uparrow$	BMR $\downarrow$
Partial (M_0)	42.91	–	5.76	76.7	0.387
$0.5A$	39.42	-3.50 $[-8.16, +0.97]$	5.23	77.1	0.370
$0.75A$	42.72	-0.19 $[-4.08, +3.69]$	5.42	77.1	0.374
$1.25A$	38.45	-4.47 $[-8.54, -0.39]$	5.86	76.3	0.405
$1.5A$	41.94	-0.97 $[-4.85, +2.91]$	5.76	76.9	0.422
Shifted M_0	36.31	-6.60 $[-11.84, -1.55]$	5.31	77.3	0.428
Random A #1	37.67	-5.24 $[-9.90, -0.78]$	5.72	75.3	0.414
Random A #2	35.73	-7.18 $[-12.62, -1.75]$	5.80	75.0	0.418
Random A #3	36.70	-6.21 $[-11.26, -1.17]$	5.79	75.3	0.404

Full–Endpoint-free TCR gains are significant for Wan2.2, SANA, MfM, and Wan2.2 Base, and comparable for Wan2.2 with ChronoEdit or Qwen-Image-Edit endpoints, Hunyuan, and OmniWeaving (Table C6).

In Wan2.2's edit-type results, Partial–Endpoint-free intervals are positive for Attribute binding and Background shifts, while the interval for Object binding spans zero. All subgroup Partial–Full intervals include zero (Table C7).

Temporal artifacts across conditioning mechanisms. Table C8 extends the Full–Endpoint-free comparison to temporal diagnostics. Wan2.2, SANA, Hunyuan, and Wan2.2 Base impose the edited endpoint through training-free terminal projection, whereas MfM and OmniWeaving learn endpoint conditioning during training. Every Full setting incurs at least one significant temporal artifact, but its form differs. Training-free projection raises HC and BMR in all four settings. Learned conditioning leaves HC comparable but lowers DD by 20.19 pp for MfM and 17.67 pp for OmniWeaving; OmniWeaving also raises BMR by 0.198. Learned conditioning therefore does not remove the temporal artifacts of full-frame endpoints; it shifts them toward suppressed motion. Comparisons are paired within each model, whose frame counts and resolutions differ.

Table C5: Support-control differences relative to the main Partial in Table C4. Near-zero bounds retain enough digits to preserve their sign.

	Temporal diagnostics		
Support	ΔHC ↓	ΔDD ↑	ΔBMR ↓
$0.5A$	-0.53 $[-0.83, -0.25]$	$+0.39$ $[-1.36, +2.14]$	-0.017 $[-0.033, -0.001]$
$0.75A$	-0.34 $[-0.68, -0.02]$	$+0.39$ $[-1.17, +1.94]$	-0.013 $[-0.026, +0.001]$
$1.25A$	$+0.10$ $[-0.15, +0.36]$	-0.39 $[-2.33, +1.36]$	$+0.019$ $[+0.001, +0.036]$
$1.5A$	$+0.00$ $[-0.28, +0.29]$	$+0.19$ $[-1.36, +1.94]$	$+0.036$ $[+0.017, +0.056]$
Shifted M_0	-0.44 $[-0.86, -0.03]$	$+0.58$ $[-1.94, +3.11]$	$+0.041$ $[+0.018, +0.066]$
Random A #1	-0.04 $[-0.41, +0.33]$	-1.36 $[-3.50, +0.78]$	$+0.028$ $[+0.002, +0.054]$
Random A #2	$+0.05$ $[-0.35, +0.45]$	-1.75 $[-3.88, +0.39]$	$+0.031$ $[+0.008, +0.056]$
Random A #3	$+0.03$ $[-0.39, +0.45]$	-1.36 $[-3.69, +0.97]$	$+0.017$ $[-0.010, +0.045]$

Table C6: Full–Endpoint-free completion differences. All rows use FLUX.2 endpoints except the ChronoEdit and Qwen-Image-Edit rows.

	Transition completion		
Backbone	Endpoint-free TCR ↑	Full TCR ↑	Full–Endpoint-free [95% CI]
Wan2.2-Turbo	33.79	41.94	$+8.16$ $[+1.75, +14.56]$
Wan2.2-Turbo (ChronoEdit)	33.79	40.19	$+6.41$ $[-0.19, +13.01]$
Wan2.2-Turbo (Qwen-Image-Edit)	33.79	35.15	$+1.36$ $[-5.63, +8.54]$
SANA-Video	23.50	35.73	$+12.23$ $[+5.44, +19.22]$
HunyuanVideo-1.5	40.97	46.60	$+5.63$ $[0.00, +11.46]$
MfM	32.62	43.88	$+11.26$ $[+3.88, +18.64]$
OmniWeaving	39.42	43.30	$+3.88$ $[-2.52, +10.49]$
Wan2.2 Base	31.07	45.24	$+14.17$ $[+8.16, +20.39]$

Table C7: Wan2.2-Turbo completion by edit type: TCR means and paired differences. Partial uses the main fixed-support cohort.

	Transition completion					
Edit type	Endpoint-free	Full	Partial	Full–Endpoint-free	Partial–Endpoint-free	Partial–Full
Attribute binding	31.60	48.40	47.60	$+16.80$ $[+7.27, +26.79]$	$+16.00$ $[+5.56, +26.67]$	-0.80 $[-6.55, +4.58]$
Background shifts	38.67	46.67	54.67	$+8.00$ $[-5.33, +23.33]$	$+16.00$ $[+3.08, +30.91]$	$+8.00$ $[0.00, +20.00]$
Object binding	34.74	31.58	32.11	-3.16 $[-13.02, +6.45]$	-2.63 $[-9.00, +3.89]$	$+0.53$ $[-8.12, +9.41]$
Overall	33.79	41.94	42.91	$+8.16$ $[+1.75, +14.56]$	$+9.13$ $[+3.11, +15.34]$	$+0.97$ $[-3.50, +5.44]$

Table C8: Full–Endpoint-free temporal differences by endpoint-conditioning mechanism. Paired prompt bootstrap as in Table C1; comparisons are paired within each model.

Model	ΔHC ↓	ΔDD ↑ (pp)	ΔBMR ↓
Training-free terminal projection			
Wan2.2-Turbo	$+1.90$ $[+1.27, +2.59]$	-12.62 $[-17.48, -8.16]$	$+0.239$ $[+0.189, +0.289]$
SANA-Video	$+1.70$ $[+0.67, +2.70]$	-9.90 $[-15.53, -4.27]$	$+0.298$ $[+0.254, +0.343]$
HunyuanVideo-1.5	$+9.50$ $[+8.53, +10.51]$	$+1.17$ $[-1.55, +3.88]$	$+0.279$ $[+0.233, +0.326]$
Wan2.2 Base	$+1.47$ $[+0.76, +2.24]$	-2.14 $[-6.99, +2.72]$	$+0.082$ $[+0.041, +0.123]$
Learned endpoint conditioning			
MfM	$+0.52$ $[-0.02, +1.12]$	-20.19 $[-27.57, -13.20]$	$+0.019$ $[-0.025, +0.063]$
OmniWeaving	-0.41 $[-0.90, +0.08]$	-17.67 $[-24.27, -11.07]$	$+0.198$ $[+0.149, +0.245]$

D Donor-Context Intervention and Score Decomposition

D.1 Construction, Pairing, and Scope

The conditions separate terminal content from projection support:

Table D1: Donor-context construction. O/D/D_{cross}/N all retain Full support. Content inside M_0 always comes from the original edited endpoint.

Condition	Inside M_0	Outside M_0	Support
O: Original Full	Original latent	Original latent	Full
D: Donor-context Full	Original latent	Re-encoded donor final RGB	Full
D_{cross}: Cross-seed donor	Original latent	Re-encoded final RGB of another seed's donor	Full
N: Native-latent Full	Original latent	Native donor tail latent	Full
P: Partial donor	Original latent	Video prediction retained	M_0

Generation and pairing. On Hunyuan, we use 480p, eight steps, and 121 frames at 24 FPS, with the native scheduler and noise-matched projection. Each of the 103 prompts uses seeds 22001–22005. Within each pair, source, prompt, edited endpoint, transformed M_0, and initial/endpoint noise are fixed.

P is a Partial rollout. We take its final RGB frame before compression and its native terminal latent. We encode the original endpoint and donor RGB through the same five-frame repetition procedure and select the second temporal slot. D uses Eq. (5); N uses the native donor slot instead. Both retain the original endpoint inside M_0.

O reuses the main Full videos; P, D, and N each contain 515 new videos. The donor P cohort is separate from main Partial (Table B1).

On Wan2.2, all four conditions are new 515-video runs with the main four-step sampler, 121 frames at 24 FPS, and one CUDA generator per prompt–seed that drives the initial noise and per-step re-noising of every condition. P is Partial with the main settings; we save its final RGB frame before compression and its native terminal latent. The endpoint and donor RGB use the same tail-slot encoding. M_0 enters as the pooled weight map of Wan2.2 Partial (2×2 average pooling to the 22×40 token grid, repeated to the 44×80 latent grid), so it selects content, while O, D, and N use Full support ($w = 1$).

Both backbones use the same BMR measurement mask for all conditions and all 515 pairs, without DD filtering. Confidence intervals follow Appendix C.

Interpretation. O→D changes non-edit terminal content under fixed Full support. The donor is generated, not ground-truth future context; position/visibility, image domain, and trajectory compatibility are not separately manipulated. D→N tests terminal representation. Partial itself ignores endpoint content outside M_0, so replacing that content does not change its projection.

D.2 Cross-Seed Donors on Hunyuan

For each prompt, a cyclic shift $k \in \{1, 2, 3, 4\}$ assigns donor seed index $(j + k) \bmod 5$ to recipient seed index j. Shifts are balanced across prompts (26/26/26/25) and were fixed before generation. The recipient keeps its source, prompt, seed, initial and endpoint noises, and transformed M_0; only the donor frame used outside M_0 changes. A pilot reproduced the existing Original Full videos and, with the donor seed set to the recipient seed, the same-seed D videos exactly. Cross-seed donors retain 35% ($[22, 45]$) of the same-seed HC reduction and 46% ($[33, 58]$) of the BMR reduction (Tables D2 and D3).

D.3 Secondary Donor and Representation Contrasts

On Hunyuan, donor-context Full has $+2.91$ $[0.00, +5.83]$ pp TCR relative to P, with higher HC and BMR; the completion interval includes zero. Replacing the re-encoded donor with its native slot further reduces HC (-0.94 $[-1.17, -0.71]$), while the BMR interval spans zero (Table D3). On Wan2.2, the native terminal latent leaves TCR, HC, and DD comparable to the re-encoded donor

Table D2: Donor-context condition means. Within each backbone, O/D/D_{cross}/N use Full support; P is the Partial donor cohort. All conditions share the BMR measurement mask.

	Transition completion		Temporal diagnostics		
Condition	TCR ↑	TC-Score ↑	HC ↓	DD ↑	BMR ↓
HunyuanVideo-1.5					
O — Original Full	46.60	0.859	12.96	52.62	0.585
D — Donor-context Full	47.57	0.873	8.10	52.82	0.358
D_{cross} — Cross-seed donor	46.80	0.869	11.27	52.04	0.480
N — Native-latent Full	45.63	0.879	7.16	52.23	0.354
P — Partial	44.66	0.882	5.69	51.46	0.333
Wan2.2-Turbo					
O — Original Full	43.30	0.834	6.49	65.44	0.600
D — Donor-context Full	43.11	0.821	6.12	75.34	0.362
N — Native-latent Full	42.72	0.820	6.26	74.95	0.382
P — Partial	42.72	0.819	5.76	76.70	0.387

Table D3: Paired donor-context differences. D–O and D_{cross}–O are the primary contrasts; the others are secondary. Appendix C specifies the uncertainty interpretation.

	Transition completion		Temporal diagnostics		
Difference	ΔTCR ↑	ΔTC-Score ↑	ΔHC ↓	ΔDD ↑	ΔBMR ↓
HunyuanVideo-1.5					
D–O	$+0.97$ $[-4.66, +6.60]$	$+0.014$ $[+0.003, +0.026]$	-4.86 $[-5.75, -3.99]$	$+0.19$ $[-1.55, +1.94]$	-0.227 $[-0.267, -0.186]$
D_{cross}–O	$+0.19$ $[-5.63, +6.02]$	$+0.010$ $[-0.000, +0.021]$	-1.70 $[-2.47, -0.94]$	-0.58 $[-2.14, +0.78]$	-0.105 $[-0.141, -0.068]$
D_{cross}–D	-0.78 $[-3.88, +2.33]$	-0.004 $[-0.009, +0.001]$	$+3.16$ $[+2.65, +3.71]$	-0.78 $[-2.14, +0.58]$	$+0.122$ $[+0.091, +0.154]$
N–O	-0.97 $[-6.60, +4.66]$	$+0.020$ $[+0.010, +0.030]$	-5.80 $[-6.68, -4.94]$	-0.39 $[-1.94, +1.17]$	-0.231 $[-0.272, -0.189]$
N–D	-1.94 $[-5.44, +1.36]$	$+0.005$ $[-0.000, +0.011]$	-0.94 $[-1.17, -0.71]$	-0.58 $[-1.55, +0.39]$	-0.004 $[-0.018, +0.009]$
O–P	$+1.94$ $[-3.50, +7.38]$	-0.023 $[-0.034, -0.012]$	$+7.27$ $[+6.39, +8.17]$	$+1.17$ $[-1.17, +3.50]$	$+0.251$ $[+0.208, +0.295]$
D–P	$+2.91$ $[0.00, +5.83]$	-0.008 $[-0.014, -0.003]$	$+2.41$ $[+2.07, +2.75]$	$+1.36$ $[-0.97, +3.50]$	$+0.025$ $[+0.004, +0.046]$
Wan2.2-Turbo					
D–O	-0.19 $[-5.05, +4.27]$	-0.013 $[-0.024, -0.002]$	-0.37 $[-0.95, +0.18]$	$+9.90$ $[+5.63, +14.56]$	-0.238 $[-0.283, -0.195]$
N–O	-0.58 $[-5.44, +4.08]$	-0.013 $[-0.026, -0.001]$	-0.24 $[-0.74, +0.24]$	$+9.51$ $[+4.85, +14.37]$	-0.218 $[-0.263, -0.174]$
N–D	-0.39 $[-2.91, +2.14]$	-0.000 $[-0.007, +0.006]$	$+0.14$ $[-0.14, +0.41]$	-0.39 $[-2.14, +1.36]$	$+0.020$ $[+0.001, +0.040]$
O–P	$+0.58$ $[-4.47, +5.83]$	$+0.015$ $[+0.003, +0.027]$	$+0.74$ $[+0.22, +1.29]$	-11.26 $[-16.31, -6.60]$	$+0.213$ $[+0.170, +0.258]$
D–P	$+0.39$ $[-3.11, +4.08]$	$+0.002$ $[-0.005, +0.009]$	$+0.36$ $[+0.05, +0.69]$	-1.36 $[-3.88, +1.17]$	-0.025 $[-0.049, -0.001]$

and raises BMR ($+0.020$ $[+0.001, +0.040]$). Relative to P, D has comparable TCR and DD, slightly lower BMR, and slightly higher HC (Table D3).

D.4 Descriptive Closure of the Full–Partial Gap

For an artifact metric m, the two summaries use different denominators:

$$\text{Relative reduction} = \frac{\bar{m}_O - \bar{m}_D}{\bar{m}_O}, \qquad R_m = \text{Gap closure} = \frac{\bar{m}_O - \bar{m}_D}{\bar{m}_O - \bar{m}_P}. \tag{D1}$$

O→D reduces HC by 37.5% and BMR by 38.8% relative to O. Using donor P as reference, it closes 66.83% and 90.16% of the Full–Partial gaps. Residual D–P differences are $+2.41$ $[+2.07, +2.75]$ for HC and $+0.025$ $[+0.004, +0.046]$ for BMR. On Wan2.2, with the same-run P as reference, D closes 112% ($[100, 125]$) of the BMR gap, reaching P's BMR; the HC closure is uncertain (51%,

$[-58, 93]$) because the Full–Partial HC gap is small. Gap closure summarizes these mean differences; it is not a causal mediation estimate.

D.5 TC-Score Decomposition and Its Relation to HC

For TC-Score $= \frac{2}{3}V + \frac{1}{3}C$, we bootstrap the VLM term V and mapped CLIP term C over the same paired prompts. Their weighted O→D contributions are $+0.00298$ and $+0.01150$, totaling $+0.01448$ (Table D4). The increase is mainly in CLIP consistency; the VLM interval includes zero. On Wan2.2, the weighted contributions are -0.00258 and -0.01031, totaling -0.01289: the decrease is mainly in CLIP consistency, the VLM interval includes zero, and it accompanies a 9.90 pp DD increase. TC-Score and HC both use adjacent-frame CLIP similarity, so their agreement is not independent evidence of completion.

Table D4: O→D decomposition of TC-Score on each backbone. Weighted VLM and CLIP terms sum to TC-Score.

Component	O mean	D mean	Δ [95% CI]
HunyuanVideo-1.5			
VLM	0.86545	0.86991	$+0.00447$ $[-0.01200, +0.02020]$
Mapped CLIP	0.84545	0.87995	$+0.03451$ $[+0.02604, +0.04341]$
$\frac{2}{3}$ VLM	0.57696	0.57994	$+0.00298$ $[-0.00800, +0.01347]$
$\frac{1}{3}$ mapped CLIP	0.28182	0.29332	$+0.01150$ $[+0.00868, +0.01447]$
TC-Score ↑	0.85878	0.87326	$+0.01448$ $[+0.00293, +0.02575]$
Wan2.2-Turbo			
VLM	0.84855	0.84468	-0.00387 $[-0.01737, +0.00941]$
Mapped CLIP	0.80368	0.77274	-0.03093 $[-0.04778, -0.01484]$
$\frac{2}{3}$ VLM	0.56570	0.56312	-0.00258 $[-0.01158, +0.00627]$
$\frac{1}{3}$ mapped CLIP	0.26789	0.25758	-0.01031 $[-0.01593, -0.00495]$
TC-Score ↑	0.83359	0.82070	-0.01289 $[-0.02386, -0.00233]$

E Diagnostics and Cross-Checks

E.1 Hunyuan Discontinuity Localization

Table E1: HunyuanVideo-1.5 Dynamic Degree after removing HC-neighbor pairs or the three largest-flow pairs. The original motion threshold and count requirement remain fixed. HC-neighbor removal drops two pairs at a sequence edge and three otherwise.

Condition	Original DD ↑	Remove $t_{\mathrm{HC}} \pm 1$	Remove top-3 flow
Endpoint-free	51.5	48.3	46.0
Full	52.6	50.5	47.0
Partial	51.5	49.9	46.6

Table E2: HunyuanVideo-1.5 HC localization and RAFT magnitude. t_{HC} is each video's minimum-similarity pair. The last two columns average RAFT top-5% magnitudes at that pair and at the remaining pairs; they do not use a fixed pair 38. The second column reports the fraction with $t_{\mathrm{HC}} = 38$.

Condition	$t_{\mathrm{HC}} = 38$	$m(t_{\mathrm{HC}})$	Other pairs
Endpoint-free	1.7%	19.5 px	11.3 px
Full	95.9%	88.9 px	11.4 px
Partial	51.8%	18.3 px	11.4 px

Let t_{HC} be each video's minimum-similarity adjacent-frame pair on the 8-FPS grid. We remove either t_{HC} and its neighbors or the three largest-flow pairs while keeping DD's threshold and qualifying-pair count fixed (Table E1). Both removals reduce DD similarly across Endpoint-free, Full, and Partial, so Full's maintained DD is not explained by a few extreme motion spikes. Pair 38 connects native frames 114 and 117, the onset of Hunyuan's terminal latent slot, and is t_{HC} in 95.9% of Full videos, versus 1.7% for Endpoint-free and 51.8% for Partial. At each video's own t_{HC}, mean RAFT top-5% magnitude is 88.9 px for Full, compared with 19.5 and 18.3 px for Endpoint-free and Partial; all remaining pairs average 11.3–11.4 px (Table E2).

E.2 Blinded Model-Judge Agreement for HC

Protocol. We use Qwen3.6-35B-A3B with temperature 0, top-p 1, thinking disabled, and a 1,536-token response limit. We evaluate 90 prompts outside the 13-prompt development subset, seeds 22001 and 22002, and nine conditions: Wan2.2 Endpoint-free, Full, Partial, adaptive release, and random release; MfM Full; OmniWeaving Full; and SANA Full and Partial. Hunyuan is excluded. The judge sees all native frames (81–121), resized on a 32-pixel grid within a 768×432 per-frame pixel budget. Method names, prompts, endpoints, masks, and metric values are hidden; no DD filter is used.

We ask whether any interval changes abruptly without a continuous transition. The judge provides visible evidence and event intervals before a PRESENT/ABSENT/UNCERTAIN label. Continuous fast motion and smooth fades are excluded, and first/last-frame differences alone are insufficient.

Results. Of 1,620 inputs, 1,619 receive valid judgments: 68 PRESENT and 1,551 ABSENT. HC achieves AUROC 0.777 (95% CI $[0.70, 0.84]$) and AUPRC 0.157, against 4.2% event prevalence. We use stepwise average precision and 1,000 prompt-cluster bootstrap draws (seed 20260917) for AUROC uncertainty. Wan2.2 event rates are 0.0% for Endpoint-free, 8.3% for Full, and 5.0% for Partial.

On 18 prompt–seed pairs across all nine conditions, native-frame judgments agree with judgments using 8 FPS input in 158/162 cases (97.5%) and with those from a repeat evaluation in 159/162 (98.1%).

E.3 Track Reversal Ratio and Back-Tracking Ratio

Track extraction. CoTracker3 tracks points at the native 24 FPS on videos resized to 640×352 (Karaev et al., 2025). We query a 24×24 grid at frames 0, 30, and 60, restricted to the region outside M_0 dilated by two latent cells.

Reference-axis construction. A track's early reference velocity v_i is the median of valid adjacent displacements during the first third of its remaining duration; at least six valid steps are required. Tracks starting before the video's midpoint with $\|v_i\| \geq 0.3$ pixels per frame provide the global reference

$$g = \frac{\sum_i \|v_i\| v_i}{\sum_i \|v_i\|}, \qquad u_g = g/\|g\|. \tag{E1}$$

At least ten such tracks and $\|g\| \geq 0.2$ pixels per frame are required.

Track applicability. Each subsequently retained track must have at least four valid late steps and an along-axis excursion of at least six pixels; at least ten retained tracks are required. These criteria define when the track diagnostics are applicable.

Trajectory diagnostics. Writing $q_{it} = \langle p_{it} - p_{i,t_i}, u_g \rangle$ for projected displacement, TRR measures the fraction of along-axis step mass that is negative in the last third of the video:

$$\mathrm{TRR} = \frac{\sum_{i,t \in \mathcal{L}_i} \max(0, -\Delta q_{it})}{\sum_{i,t \in \mathcal{L}_i} |\Delta q_{it}|}. \tag{E2}$$

Here $\mathcal{L}_i$ includes only valid late adjacent steps. BTR measures return from the largest achieved displacement to the last visible track position, which need not be the video's last frame. Write $q_{i,\max} = \max_t q_{it}$ and $q_{i,\min} = \min_t q_{it}$ over visible positions. The per-track ratio and its excursion-weighted aggregation are

$$e_i = q_{i,\max} - q_{i,\min}, \qquad r_i^{\mathrm{return}} = \frac{q_{i,\max} - q_{i,\mathrm{end}}}{e_i}, \tag{E3}$$

$$\mathrm{BTR} = \frac{\sum_i e_i r_i^{\mathrm{return}}}{\sum_i e_i}. \tag{E4}$$

We add 10^{-6} to normalization denominators for numerical stability. TRR and BTR use point trajectories and a last-third late window, complementing BMR's flow-based last-half measurement.

Table E3: Flow and track diagnostics on the common intersection of 398/515 Wan2.2-Turbo samples applicable under all three conditions. Every mean uses the same samples.

Condition	Common applicable	BMR ↓	TRR ↓	BTR ↓
Endpoint-free	398 / 515	0.386	0.385	0.332
Full	398 / 515	0.633	0.598	0.600
Partial	398 / 515	0.415	0.410	0.359

Table E4: Outside-M_0 first–last-frame RGB difference of preserved context: mean absolute RGB difference between each video's first and last frames. For reference, FLUX.2 endpoints differ from the source image by 0.056 $[0.045, 0.068]$ outside M_0 and 0.232 $[0.206, 0.258]$ inside. Intervals use the paired prompt bootstrap. Wan2.2-Turbo Partial videos were regenerated with identical settings for this measurement.

	First–last RGB difference			Difference from Endpoint-free	
Model	Endpoint-free	Full	Partial	Full	Partial
Wan2.2-Turbo	0.177	0.076	0.164	-0.102 $[-0.116, -0.087]$	-0.014 $[-0.019, -0.009]$
SANA-Video	0.233	0.058	0.216	-0.174 $[-0.195, -0.155]$	-0.017 $[-0.024, -0.010]$
HunyuanVideo-1.5	0.127	0.082	0.126	-0.045 $[-0.059, -0.031]$	-0.001 $[-0.002, +0.000]$

Common-population comparison. Track diagnostics apply to 500/515 Endpoint-free, 413/515 Full, and 492/515 Partial videos. We compare all three metrics on their common 398 prompt–seed pairs across 98 prompts (Table E3). Partial–Full differences in BMR/TRR/BTR are $-0.218/-0.188/-0.241$. Pooled Spearman correlations between BMR and TRR/BTR are 0.64/0.65 across the 1,194 paired, condition-specific observations. Main BMR results use all 515 videos per condition.

E.4 Source-Time Layout of Preserved Context

We measure the first–last-frame RGB difference of preserved context outside the estimated support. Images and frames are downsampled to width 640, and M_0 is resized using nearest-neighbor interpolation, as for BMR. The primary distance is the mean absolute RGB difference in $[0, 1]$ outside M_0; LPIPS-Alex is secondary (Zhang et al., 2018). For videos, we compare each video's first and last frames at its own resolution, avoiding resampling differences between pipelines. FLUX.2 endpoints differ from the source image less outside M_0 than inside in 93% of prompts, whereas this context differs more between the first and last frames of Endpoint-free videos than between the endpoint and the source in 91–99% of prompts. Under Full, the last frame stays close to the first frame outside M_0, and Partial leaves the difference within 0.02 of Endpoint-free (Table E4). LPIPS gives the same ordering.

Before computing these quantities, we also specified a dose–response test: whether Full's HC and BMR costs relative to Endpoint-free grow with the Endpoint-free first–last-frame difference across prompts, and whether Partial's reductions relative to Full do (Spearman correlation with a prompt bootstrap; for Full−Endpoint-free, this difference and the outcomes come from disjoint seeds). Only the two Hunyuan HC tests support it ($\rho = +0.28$ $[+0.08, +0.47]$ and -0.35 $[-0.50, -0.17]$); the other ten intervals include zero. We therefore do not claim that Full's cost scales with scene motion.

F Residual Failures and Exploratory Support Release



Figure F1: **A residual failure of fixed support.** (a) The edited endpoint preserves an occluding bus-stop sign; blue shows M_0, and the red box defines the crop. (b) Identical image-coordinate crops at all eight original frame indices. Red borders mark the sign's reappearance in Partial and random release; attention-ranked and adaptive release avoid it in this example. The release variants are exploratory. All rows use Wan2.2-Turbo, purple-bus case 90001 and seed 22001, as in Fig. 4; frame intervals are unequal.

Fixed support cannot follow moving content. Figure F1 shows an endpoint-preserved sign reappearing under Partial after disappearing earlier in the video.

Release rules. We explore shrinking support over Wan2.2's denoising timesteps $1000, 960, 889, 727$. Adaptive release starts from pooled M_0. At step k, we compute the channelwise Euclidean distance between the endpoint latent and each middle-frame clean prediction, average across frames, and pool spatially by 2×2 to obtain G_k. With τ_k the mean over tokens, we update $w_{k+1} = w_k \mathbf{1}[G_k < \tau_k]$ at the first three steps. Released tokens never re-enter.

Random and attention-ranked release match each sample's adaptive weight totals to whole-token precision. They retain a prefix of initially active tokens ranked randomly or by pooled FLUX.2 attention, respectively. The same w_k controls model-input blending, token timesteps, and clean-prediction projection.

Completion–discontinuity trade-off. On the 515-video Wan2.2 cohort, adaptive release yields TCR 38.45%, TC-Score 0.811, HC 4.89, and DD 76.3%, compared with fixed support's 42.91%, 0.819, 5.76, and 76.7%. Attention-ranked and adaptive release avoid the illustrated sign reappearance, while random release does not. The aggregate adaptive result trades lower HC for lower TCR; this example does not establish a general advantage over fixed support.

G BROADER I2V COMPLETION COMPARISONS

Table G1: Automatic evaluation results of I2V models on TC-Bench-I2V. FPS reports video-stage throughput.

Methods	Params	TC-Bench-I2V ↑								FPS ↑
		Attribute		Object		Background		Overall		
		TCR	TC-Score	TCR	TC-Score	TCR	TC-Score	TCR	TC-Score	
Image-to-Video Generation										
Wan2.2-Turbo Release'25.08	5B	31.60	0.7647	34.74	0.7842	38.67	0.8486	33.79	0.7841	2.270
Wan2.2 Base Release'25.07	5B	32.00	0.7935	30.00	0.7794	30.67	0.8442	31.07	0.7957	0.182
SANA-Video ICLR'26	2B	23.20	0.7729	22.11	0.7245	28.00	0.7590	23.50	0.7530	1.031
HunyuanVideo-1.5 arXiv'25.11 (480p, 8-step)	8.3B	52.80	0.9076	25.26	**0.8312**	41.33	0.8556	40.97	0.8718	0.615
MfM ICLR'26	8.1B	39.60	0.8449	24.74	0.7887	29.33	0.8365	32.62	0.8230	0.034
OmniWeaving arXiv'26.03	8.3B	42.80	0.8653	**35.79**	0.8034	37.33	0.8305	39.42	0.8374	0.027
Image-to-Video Generation w/ Image Editing										
FLUX.2 + Wan2.2-Turbo	14B	48.40	0.8466	31.58	0.8081	46.67	0.8655	41.94	0.8352	2.270
ChronoEdit + Wan2.2-Turbo	19B	40.80	0.8239	33.68	0.7891	**54.67**	0.8948	40.19	0.8214	2.266
Qwen-Image-Edit + Wan2.2-Turbo	25B	41.20	0.8096	30.53	0.7736	26.67	0.8357	35.15	0.8001	2.266
FLUX.2 + Wan2.2 Base	14B	54.80	0.8700	31.58	0.8034	48.00	0.8838	45.24	0.8475	0.186
FLUX.2 + SANA-Video	11B	44.40	0.8451	25.79	0.7200	32.00	0.8136	35.73	0.7944	1.029
FLUX.2 + HunyuanVideo-1.5	17.3B	**58.40**	0.9018	31.05	0.8069	46.67	0.8470	**46.60**	0.8588	0.565
FLUX.2 + MfM	17.1B	58.00	**0.9344**	25.79	0.8198	42.67	**0.8983**	43.88	**0.8869**	0.034
FLUX.2 + OmniWeaving	17.3B	51.60	0.8977	32.63	0.8200	42.67	0.8683	43.30	0.8647	0.029

We additionally compare ChronoEdit and Qwen-Image-Edit endpoints (Wu et al., 2026; 2025b) on the same 103 prompts and five seeds. Image-editing rows use Full conditioning. Tables C7 and C6 provide paired completion differences; Table 1 evaluates Partial.

Endpoint sources and parameter counts. The Params column reports generative-backbone parameter counts in billions, including the editor but excluding separate text/vision encoders and VAEs. FLUX.2 uses klein-base-9B and the final-frame prompt in Appendix A. ChronoEdit uses its 14B checkpoint, which formulates editing as video generation with the input and edited images as the first and last frames. We therefore supply the original transition prompt directly, without the final-frame template, and use the last of five generated frames as the endpoint. Thus both editor and prompt formulation change. Qwen-Image-Edit uses the 20B 2511 release[3] with the FLUX.2 final-frame prompt, so only the editor changes.

Native configurations and timing. MfM uses 97 frames at 1280×704 and OmniWeaving 81 at 720p. FPS is computed as the number of decoded frames divided by the median video-stage time, excluding endpoint generation. MfM includes batch-amortized startup; other methods use per-video timers. Native configurations and timing scopes vary across methods.

[3]Official model card: Qwen-Image-Edit-2511.