SLES: SELF-LOCALIZED ENTITY SUPERVISION FOR LONG-HORIZON DYNAMIC MEMORY IN VIDEO WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon simulation of a persistent world requires maintaining the state of dynamic entities even while they lie outside the field of view, recovering each with its identity and evolving state intact upon re-entry; we refer to this capability as long-horizon dynamic memory. Despite rapid advances in generation quality and controllability, existing models frequently fail to recover entities that have left the field of view, or do so with a different identity or implausible dynamics. We study this problem under full-context bidirectional generation, where the complete observation history remains accessible, and find that failures persist, indicating that information utilization, rather than access, is the bottleneck. Probing self-attention reveals an observability–controllability gap: entity retrieval is encoded in reproducible attention structure, yet directly steering this routing does not reliably improve generation and can degrade identity consistency. We therefore use attention only to localize intervention. In this paper, we propose **S**elf-**L**ocalized **E**ntity **S**upervision (SLES), which derives a lightweight spatial entity support from text prior and reuses it to up-weight entity-aware supervision during training and to localize a Langevin-inspired overshooting sampler that concentrates stochastic correction on the entity region during inference, requiring no external memory, architectural modification, or target-side masks. We further introduce a phase-aware protocol separating reappearance, subject consistency, and temporal recovery. Across backbones and ablations, self-derived support approaches ground-truth mask supervision for identity recovery, while localized overshooting improves post-re-entry dynamics with limited collateral disturbance.[1]

1 INTRODUCTION

Recent video world models have advanced rapidly from short, passive video synthesis toward controllable interaction and increasingly long generation(Qian et al., 2026; Gao et al., 2026; Zhang et al., 2026c; Zhu et al., 2026; Team et al., 2026; Huang et al., 2026; Sun et al., 2025). However, controllability and duration alone do not guarantee that the generated world preserves its evolving state. Camera motion may temporarily move a dynamic entity outside the field of view; when it later re-enters, a persistent world model should recover the same entity with a consistent evolving state. We refer to this capability as long-horizon dynamic memory (Chen et al., 2026b; Zhang et al., 2026b; Chen et al., 2026a).

Existing approaches to long-horizon memory mainly improve information retention by extending temporal context or explicitly storing and retrieving historical views, geometry, or scene state (Xiao et al., 2026; Duan et al., 2026; Wu et al., 2026b). In contrast, dynamic memory, where an entity leaves observation and must later be recovered together with its evolving state, remains comparatively underexplored (Chen et al., 2026b; Duan et al., 2026; Xue et al., 2026; Zhang et al., 2025; 2026a). Our experiments show this challenge appears under 10-second exit–re-entry trajectories: many contemporary video world models fail to preserve entity identity or post-return dynamics.

Characterizing this failure first requires an evaluation that distinguishes its different forms. Existing memory metrics largely emphasize re-detection, reconstruction, visual consistency, or motion

[1] Code and evaluation protocol are provided in the supplementary materials.

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.

smoothness (Zhang et al., 2026b; Chen et al., 2026a;b), but they can assign similar scores to qualitatively different failure modes. We therefore introduce a phase-aware protocol that separately evaluates reappearance, subject consistency, and temporal recovery.

Using this evaluation, our pilot study reveals that long-horizon dynamic-memory failure stems not from insufficient context: even in full-context bidirectional generation, where observation history remains accessible throughout denoising, entity recovery is still imperfect. We therefore examine how the model uses retained information. Analysis of full-sequence self-attention reveals structured retrieval of historical entity evidence, and stronger entity recovery accompanies stronger entity-focused attention routing across paired checkpoints. Directly biasing this routing during training or inference, however, does not reliably improve generation and can degrade subject consistency. This exposes an observability–controllability gap: attention is an observable readout of entity-information utilization, but not a reliable control target.

This gap motivates an indirect approach: rather than steering attention itself, we localize where supervision and correction act. We first increase the flow-matching weight inside the ground-truth entity region and observe substantial gains, indicating that localized supervision improves the utilization of retained entity information. To remove the need for per-sample target masks, we then use text cross-attention as a semantic prior to derive the entity region from the model itself. The resulting self-localized supervision approaches ground-truth-mask supervision, without an external memory module or target-side annotations beyond calibration. We further extend the same localization principle to inference. Inspired by Langevin correction along the rectified-flow trajectory, we propose a region overshooting sampler that concentrates stochastic correction on the predicted entity region, improving long-horizon consistency. Together, these components form our Self-Localized Entity Supervision (SLES) framework.

We evaluate SLES across diverse dynamic-memory settings against contemporary video world models and memory-oriented approaches. Extensive controlled interventions, support-source ablations, sampling studies, and qualitative analyses consistently corroborate the effectiveness of spatially grounded supervision and localized correction. Our contributions are three-fold:

- We systematically characterize long-horizon dynamic memory from both evaluation and information-utilization perspectives, introducing a phase-aware evaluation protocol and exposing an observability–controllability gap in entity-focused self-attention.
- We introduce Self-Localized Entity Supervision, which derives entity regions from the model's own text-conditioned signals and uses them to localize both training supervision and inference-time stochastic correction, without an external memory module.
- We conduct extensive comparisons and controlled ablations, showing that entity-aware supervision mainly improves subject recovery, whereas localized sampling mainly improves temporal consistency with limited disturbance to the rest of the scene.

2 Related Work

Long-horizon world models and dynamic memory. Recent video generators increasingly support camera, action, and history context conditioning for interactive and long-horizon world simulation (Qian et al., 2026; Gao et al., 2026; Zhang et al., 2026c; Zhu et al., 2026; Team et al., 2026; Huang et al., 2026; Sun et al., 2025). Long-term consistency is commonly improved by extending temporal context or explicitly preserving historical views, geometry, or scene state (Xiao et al., 2026; Li et al., 2025; Wu et al., 2026b; Huang et al., 2025; Wu et al., 2026a). More recent work begins to consider dynamic entities that leave the current observation and later return, including hybrid retrieval, out-of-sight dynamics, and object permanence (Chen et al., 2026b; Duan et al., 2026; Xue et al., 2026). Our work focuses on this less explored dynamic-memory setting and studies why entity persistence can still fail even when the relevant historical context remains available.

Evaluation of long-horizon memory. Existing benchmarks primarily evaluate long-term static memory through entity recall, re-detection, reconstruction, visual consistency, and motion-related metrics (Zhang et al., 2026b; Chen et al., 2026a;b). While informative, these metrics can conflate distinct failures in dynamic memory: an entity may fail to return, reappear with the wrong identity, or exhibit incorrect temporal evolution. We therefore introduce a phase-aware evaluation that

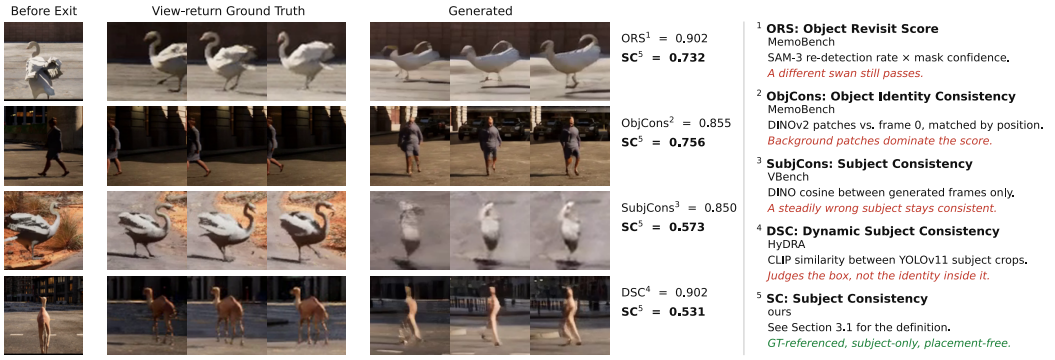


Figure 1: **Existing long-horizon memory metrics accept wrong-identity re-entries.** Each row shows one clip in which the subject leaves the view and later returns: left to right, the frame before exit, the ground-truth re-entry, and the generated re-entry, in which a *different* instance reappears. Object Revisit Score and Object Identity Consistency (MemoBench (Chen et al., 2026a)), Subject Consistency (VBench (Huang et al., 2024)), and Dynamic Subject Consistency (HyDRA (Chen et al., 2026b)) score all four failures high (0.85–0.90), whereas our Subject Consistency (SC) is the only score that drops (0.53–0.76).

separately measures reappearance, identity fidelity, and temporal recovery, offering a more precise decomposition of dynamic-memory behavior than a single consistency metric.

Internal correspondence and generative control. Video diffusion models exhibit correspondence structure in their internal representations: Track4Gen (Jeong et al., 2025) uses point-tracking supervision, DiffTrack (Nam et al., 2026) identifies temporal correspondence in query–key similarities, and MVTrack4Gen (Lee et al., 2026) extends this to camera-conditioned multi-view generation. We ask a complementary question: whether retrieval of previously observed entities remains detectable after out-of-view intervals, and whether it can be directly controlled. We find that retrieval is clearly observable, whereas explicit attention steering yields inconsistent gains, motivating our use of self-localized spatial supervision and region-localized sampling instead of direct manipulation of the retrieval mechanism.

3 Long-Horizon Dynamic Memory

3.1 Phase-Aware Long-horizon Dynamic Memory Evaluation

Existing long-horizon memory metrics capture useful aspects of video quality but do not disentangle distinct dynamic-memory failures. As illustrated in Fig. 1, none of them compares the returned subject against the ground-truth subject alone. Object Revisit Score (ORS) from MemoBench (Chen et al., 2026a) prompts SAM3 (Carion et al., 2026) for open-vocabulary re-detection and thus tests whether *an* object matching the description is present, not whether the *same* instance returns; Object Identity Consistency (ObjCons), also from MemoBench, matches whole-frame DINOv2 (Oquab et al., 2023) patches against the first frame by position, so a stable background dominates the score; Subject Consistency from VBench (Huang et al., 2024) compares generated frames only and rewards an incorrect yet temporally stable subject; and Dynamic Subject Consistency from HyDRA (Chen et al., 2026b) measures CLIP (Radford et al., 2021) similarity between YOLOv11 (Khanam & Hussain, 2024)-detected subject crops, coupling localization with appearance rather than isolating identity. On the four wrong-identity re-entries in Fig. 1, these metrics remain high (0.85–0.90), whereas our Subject Consistency (SC), defined below, is the only score that drops (0.53–0.76). Consequently, these metrics alone cannot determine whether a subject reappears, preserves its identity, and recovers appropriate temporal dynamics. Additional examples are provided in Appendix A.2.

To address this limitation, we evaluate dynamic memory along three complementary axes: *reappearance* (Rea-Score), *subject consistency* (SC), and *temporal recovery* (TCM). All evaluated methods use the same ground-truth re-entry boundary, so the three axes are measured over comparable phases. Generated subject regions are obtained with a frozen tracking–redetection pipeline, where appear-

ance cues are used only for candidate association, not for existence decisions. Full implementation details are provided in Appendix A.1.

Reappearance Score (Rea-Score). For each sample i, let $n_G^{(i)}$ denote the number of valid generated-subject frames after the re-entry boundary, and define $r_i = \delta[n_G^{(i)} \geq 5]$[2]. The Rea-Score is then computed as $\text{Rea-Score} = \frac{1}{N}\sum_{i=1}^{N} r_i$.

Subject Consistency (SC). For samples with a valid reappearance ($r_i = 1$), let $\mathbf{f}_G$ and $\mathbf{f}_T$ denote foreground descriptors aggregated over the generated and target post-entry segments. We define $\text{SC} = \cos(\mathbf{f}_G, \mathbf{f}_T)$, which measures whether the returned subject matches the target *instance* rather than an object of the same category. The comparison is ground-truth referenced, restricted to the subject foreground, and requires no spatial correspondence, so a stable background, a wrong but temporally stable subject, or correct localization alone cannot inflate the score.

Temporal Consistency Match (TCM). SC does not assess whether post-entry temporal behavior is recovered. For a foreground descriptor sequence $X = \{\mathbf{f}_t\}$, we measure temporal stability over lags $\ell = 1, \ldots, 8$ as

$$C(X) = \frac{1}{L}\sum_{\ell=1}^{L} \frac{1}{|\mathcal{P}_\ell|} \sum_{(t,t+\ell)\in\mathcal{P}_\ell} \cos(\mathbf{f}_t, \mathbf{f}_{t+\ell}), \qquad L = 8. \tag{1}$$

Here, $\mathcal{P}_\ell$ denotes the set of valid frame pairs $(t, t+\ell)$ within the sequence. Since higher temporal stability is not necessarily preferable—a nearly frozen subject can also achieve a high $C(X)$—we compare the stability of the generated and target sequences:

$$\text{TCM} = \frac{\min(C(G), C(T))}{\max(C(G), C(T))}, \tag{2}$$

where G and T denote the generated and target foreground descriptor sequences. TCM approaches 1 when both sequences exhibit similar descriptor-level temporal stability, and decreases when the generation is either more static or more variable than the target. Unlike SC, TCM can be evaluated whenever temporal consistency is estimable and does not require successful reappearance.

Complementarity. SC and TCM measure different failure modes: SC compares generated and target subject identity, whereas TCM compares their within-sequence temporal behavior. They are correlated but non-redundant; approximately one third of paired samples fall into the high-SC/low-TCM or low-SC/high-TCM regimes. Appendix A.3 provides sample-level statistics, controlled intervention analyses, and ranking changes across methods.

3.2 Understanding Information Utilization in Dynamic Memory

Entity retrieval is observable. Prior work (Chen et al., 2026b) and our pilot experiments show that full history alone is insufficient for effective long-range memory: failures persist even when history remains available, indicating that utilization, not access, is the bottleneck. Probing full-sequence self-attention via the relative attention density from pre-exit subject queries to conditioning-subject versus background keys, we find reproducible layer–head retrieval structures that are consistently stronger in the better-recovery checkpoint across all 22 paired scenes. Entity-supervised checkpoints also show higher average retrieval, which closely tracks TCM along that trajectory. However, this association vanishes in the baseline trajectory and at the scene level, indicating that attention is an informative readout of learned retrieval rather than a per-scene quality predictor.

Observable retrieval is not directly controllable. We next test whether this retrieval pattern can serve as a direct control target. Explicitly biasing context-key attention toward historical subject tokens during training or inference yields no reliable improvement and can degrade subject consistency, whereas dose-matched random perturbations have little effect. Training-time bias is also budget-dependent: it helps early but reverses at the full budget. Attention thus reveals a meaningful internal retrieval pattern without providing a reliable scalar control knob, motivating the indirect localized supervision and sampling in Sec. 4.See Appendix B for details.

[2] $\delta[\cdot]$ denotes the discrete delta: it equals 1 when the enclosed condition holds and 0 otherwise, i.e., the indicator function.

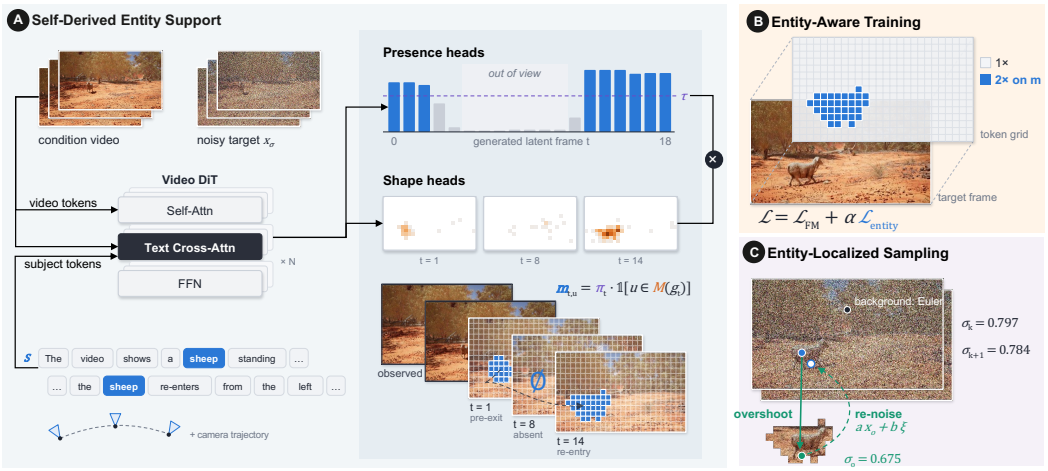


Figure 2: **Overview of the proposed Self-Localized Entity Supervision and Sampling.** **(A)** Fixed text cross-attention heads provide a self-localized entity support: presence heads determine subject visibility, shape heads localize it; absent frames receive empty support. **(B)** It adds one extra unit of flow-matching weight on subject latents, preserving the baseline objective elsewhere. **(C)** At inference, it localizes overshooting and re-noising to the subject region; background, absent frames, condition latents, and steps above the correction gate follow the base Euler update.

4 Self-Localized Entity Supervision

Sec. 3.2 shows that entity-related attention is informative but not a reliable control target. We therefore use attention only to localize where to intervene. As illustrated in Fig. 2, a binary entity support is read from text cross-attention and reused in two places: it up-weights flow-matching supervision on subject latents during training, and localizes an overshooting-based stochastic correction during sampling. No self-attention logits or backbone modules are modified. Oracle-mask supervision confirms that localized supervision is useful (Appendix C.1); the remaining challenge is to obtain the support without a target-side mask.

4.1 Self-Localized Entity Support

Let $\mathcal{S}$ denote the caption tokens referring to the target subject. For video token u in latent frame t, we read the text cross-attention mass assigned to $\mathcal{S}$ from head (ℓ, a):

$$s_{t,u}^{(\ell,a)} = \sum_{j \in \mathcal{S}} A_{t,u,j}^{(\ell,a)}. \tag{3}$$

A one-time calibration selects fixed presence heads $\mathcal{H}_p$ and shape heads $\mathcal{H}_g$. Their pooled fields are

$$p_{t,u} = \frac{1}{|\mathcal{H}_p|} \sum_{(\ell,a) \in \mathcal{H}_p} s_{t,u}^{(\ell,a)}, \qquad g_{t,u} = \frac{1}{|\mathcal{H}_g|} \sum_{(\ell,a) \in \mathcal{H}_g} s_{t,u}^{(\ell,a)}. \tag{4}$$

Presence and localization are separated because an absent frame can still contain a spatial maximum. We normalize p over the generated segment and define

$$z_t = \max_u \frac{p_{t,u} - \mu_p}{\sigma_p + \epsilon}, \qquad \hat{z}_t = \text{Median}_3(z)_t, \qquad \pi_t = \delta[\hat{z}_t > \tau]. \tag{5}$$

Let $\mathcal{M}(g_t)$ extract the spatial region around the dominant peak of g_t. The binary entity support is

$$m_{t,u} = \pi_t \, \delta[u \in \mathcal{M}(g_t)] \in \{0, 1\}. \tag{6}$$

Thus, absent frames receive zero support by construction. Head selection, threshold calibration, and the exact region-extraction rule are given in Appendix C.2.

The readout is noise-dependent. During training, support is read at

$$\sigma_{\text{read}} = \min(\sigma, 0.8). \tag{7}$$

For samples with $\sigma > 0.8$, an auxiliary copy is re-noised with the same noise realization only for support extraction, while the flow-matching objective remains evaluated at the original σ. At inference, support is read from the current sampler forward, and localized correction is disabled for $\sigma > 0.92$. The training read cap and inference correction gate are calibrated separately; details are provided in Appendix C.3.

4.2 Entity-Aware Training

Let $r_n = \hat{v}_\theta(x_\sigma, \sigma, y)_n - (\epsilon - x_0)_n$ denote the flow-matching residual on generated latent element n, and let $\tilde{m}_n$ be the entity support mapped to the same latent grid. We optimize

$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N} \|r_n\|^2 + \alpha \frac{1}{N}\sum_{n=1}^{N} \tilde{m}_n \|r_n\|^2, \qquad \alpha = 1. \tag{8}$$

Both terms use the full generated-tensor size N. Supported elements therefore receive weight $1 + \alpha = 2$, while all other elements retain the baseline weight 1. The support re-weights supervision rather than cropping the objective; if no reliable support is detected, the additional term vanishes exactly. Optimization details are provided in Appendix C.4.

4.3 Region-Localized Overshooting Sampling

Entity-aware training improves subject recovery, while post-re-entry temporal evolution can still deviate from the target dynamics. We therefore reuse the same self-localized support to introduce a localized stochastic correction during inference.

A prescribed family of time marginals can be realized by both deterministic probability-flow ODEs and stochastic dynamics (Song et al., 2020; Albergo et al., 2025). This suggests that the sampling path can be perturbed stochastically while retaining the statistics expected at the next noise level. For the rectified-flow interpolation $x_\sigma = (1-\sigma)x_0 + \sigma\epsilon$, $\epsilon \sim \mathcal{N}(0, I)$, consider a descending schedule $\sigma_k > \sigma_{k+1}$ with $\Delta_k = \sigma_k - \sigma_{k+1} > 0$.

Localized overshooting. Let $m_n^{(k)} \in \{0, 1\}$ denote the self-localized entity support for latent element n at sampling step k. We assign each element a support-dependent overshoot level

$$\sigma_{o,n} = \max\Big\{0,\ \sigma_{k+1} - c\, m_n^{(k)} \Delta_k\Big\}, \tag{9}$$

where $c > 0$ controls the correction strength. Supported elements satisfy $\sigma_{o,n} < \sigma_{k+1}$, extrapolating toward a cleaner state, while unsupported elements satisfy $\sigma_{o,n} = \sigma_{k+1}$. Using the velocity already evaluated at (x_k, σ_k), we advance each element to its overshoot level:

$$x_{o,n} = x_{k,n} + \hat{v}_\theta(x_k, \sigma_k, y)_n (\sigma_{o,n} - \sigma_k). \tag{10}$$

Moment-preserving re-noising. The overshot state is then mapped back to the scheduled level σ_{k+1} through

$$\begin{aligned} x_{k+1,n} &= a_n x_{o,n} + b_n \xi_n, \qquad \xi_n \sim \mathcal{N}(0, I), \\ a_n &= \frac{1-\sigma_{k+1}}{1-\sigma_{o,n}}, \qquad b_n = \sqrt{\max\big\{\sigma_{k+1}^2 - (a_n \sigma_{o,n})^2,\ 0\big\}}. \end{aligned} \tag{11}$$

If the overshot state follows the ideal interpolation at $\sigma_{o,n}$, these coefficients satisfy

$$a_n(1 - \sigma_{o,n}) = 1 - \sigma_{k+1}, \qquad (a_n \sigma_{o,n})^2 + b_n^2 = \sigma_{k+1}^2, \tag{12}$$

thereby recovering the signal coefficient and noise variance associated with the scheduled marginal $\rho_{\sigma_{k+1}}$. The stochastic correction therefore changes the path between two sampling levels while preserving the signal–noise statistics expected at the next model evaluation.

Exact reduction to the base sampler. For $m_n^{(k)} = 0$, $(\sigma_{o,n}, a_n, b_n) = (\sigma_{k+1}, 1, 0)$, so Eqs. 10–11 reduce exactly to the base Euler update $x_{k+1,n} = x_{k,n} + \hat{v}_\theta(x_k, \sigma_k, y)_n(\sigma_{k+1} - \sigma_k)$. Thus, unsupported latents receive the unmodified Euler update at that step.

We set $m_n^{(k)} = 0$ for background elements, absence frames, conditioning latents, and steps with $\sigma_k > \sigma_{\text{samp}}^{\max} = 0.92$. Conditioning latents remain re-pinned to the observed video after each update.

Our final sampler uses $c = 8$. Unlike uniform correction, it perturbs only latent elements supported by the self-localized entity region. The support is read from attention produced by the same sampler forward used to predict velocity, requiring no additional model evaluation. Full sampler settings and computational-cost details are provided in Appendices C.5–C.6.

5 EXPERIMENTS

We conduct controlled experiments on Wan2.2-TI2V-5B (Wan et al., 2025) and LTX2.5-22B (HaCohen et al., 2025), training only the self- and cross-attention modules; the complete optimization configuration is given in Appendix C.4. Each 153-frame sequence comprises 77 context and 76 generated frames. We construct a trajectory-disjoint held-out evaluation pool by grouping all camera views of each subject–trajectory pair and holding each group out together.; Appendix E.3 reports the complete 345-scene results. We compare Baseline, SLES-Train, and SLES-Full (Table 1), averaged over two seeds, and a **GT-Mask Reference** (single seed) that replaces self-localized support with the ground-truth entity mask under the same recipe, measuring the gap to precise spatial supervision rather than a universal upper bound. Released models are evaluated under our standardized full-context protocol; Appendix D.3 audits the differences among public conditioning interfaces.

5.1 MAIN RESULTS

Table 1: **Performance comparison with recent video world models and memory-oriented methods.** Rea-Score, SC, and TCM are existing consistency metrics whose limitations for dynamic entity re-entry are discussed in Sec3.1 3.1; Cam. Gain measures camera-displacement amplitude (1 is ideal). SLES-Train adds self-localized entity supervision to Baseline, and SLES-Full further applies localized overshooting at inference. Released-model results are protocol-specific. Bold/underlined denote best/second-best results. † Reproduced with official code on our training set.

	Dynamic Memory			Camera	Existing Metrics		
Method	Rea-Score ↑	SC ↑	TCM ↑	Cam. Gain → 1	ObjCons ↑	SubjCons ↑	DSC ↑
Released world models							
Matrix-Game 3.5 (Qian et al., 2026)	90.6	.7548	.9144	.638	**.6732**	**.8156**	.8459
EchoWM (Zhang et al., 2026c)	91.1	.7495	.8796	.421	<u>.6551</u>	.7622	.8199
SANA-WM (Zhu et al., 2026)	<u>97.8</u>	.7734	.8979	.345	.6493	.7982	.8178
HY-WorldPlay 1.5 (Sun et al., 2025)	86.1	.7780	.9107	.534	.6122	<u>.8020</u>	.8490
LingBot-World-V2 (Gao et al., 2026)	96.6	.8115	<u>.9232</u>	.634	.6402	.7761	.8361
Memory-oriented method							
HyDRA (Chen et al., 2026b)†	78.9	.7465	.9071	.668	.6193	.7669	.8486
Controlled comparison							
Baseline	96.1	.8204	.9154	**1.064**	.6536	.7658	.8720
SLES-Train	**99.2**	**.8380**	.9231	1.093	.6520	.7670	**.8757**
SLES-Full	97.2	<u>.8378</u>	**.9318**	<u>1.091</u>	.6513	.7659	<u>.8750</u>

Training and sampling improve complementary axes. As shown in Table 1, Baseline achieves Rea-Score$= 96.1\%$ and SC$= .8204$; adding self-localized entity supervision raises them to 99.2% and .8380 (a two-seed mean SC gain of $+.0176$), numerically matching the GT-Mask Reference (SC .8380 vs. .8379; Rea-Score 99.2% vs. 98.9%). This does not imply that the self-localized support reproduces pixel-accurate segmentation; rather, it is sufficiently localized to provide an effective subject-focused training signal. Adding region-localized sampling then changes SC only marginally ($.8380 \rightarrow .8378$) while increasing TCM from .9231 to .9318: its primary contribution is post-reentry temporal behavior rather than subject identity. Rea-Score decreases from 99.2% to 97.2%, but this direction is not stable across sampling seeds, so we do not interpret it as a systematic trade-off.

SLES leads on all dynamic-memory axes. Under our standardized full-context protocol, different systems excel on different aspects of long-horizon generation: high reappearance or conventional consistency does not necessarily imply high SC or TCM. Within this comparison, SLES-Train attains the highest Rea-Score and SC and SLES-Full the highest TCM, while their shared backbone enters as Baseline; the gains are therefore attributable to our training and sampling components

rather than to the backbone itself. Camera execution does not offer an alternative explanation: **Cam. Gain**—the ratio of accumulated generated-to-target horizontal displacement, where 1 indicates matched amplitude—is nearly identical across the controlled configurations (1.06–1.09) while SC and TCM vary substantially. Conventional metrics respond differently across components, reflecting their distinct diagnostic scopes (Sec. 3.1); a complementary camera-following analysis on the HyDRA reproduction ladder is provided in Appendix D.5.

Qualitative results mirror the metrics. Fig. 3 visualizes subject and temporal recovery. Several released models regenerate a semantically plausible object after re-entry but change its appearance, whereas SLES-Full preserves the characteristic subject identity across subsequent frames. The temporal profiles further show that SLES-Full follows the target descriptor-similarity curve more closely across temporal lags, while Baseline and SLES-Train exhibit mismatched temporal stability. These two views correspond directly to SC and TCM, respectively.

Table 2: **Training-time entity-support controls.** Training-time controls test whether gains require spatially meaningful entity support; ΔSC is the paired SC difference against the Baseline.

Support	Rea-Score ↑	SC ↑	TCM ↑	ΔSC
None	96.1	.8186	.9120	–
Scrambled	98.3	.8274	.9162	+.0108
Self-localized	98.9	**.8388**	.9193	**+.0212**
GT mask	98.9	.8379	**.9239**	+.0207

Table 3: **Inference-time localization and backbone transfer.** (a) Inference controls: correction location and full-frame fidelity cost. (b) SLES-Train/Full on Wan2.2 and LTX-2.5, recalibrating only the per-backbone entity-support readout; training and sampling unchanged.

(a) Inference-time localization

Region	SC ↑	TCM ↑	LPIPS ↓	SSIM ↑
None	.8388	.9193	.3049	.4659
Inverted	.8401	.9177	.3048	.4659
Random	.8443	.9214	**.3045**	.4656
Self	.8440	**.9300**	.3051	**.4658**
Uniform	**.8496**	.9342	.3200	.4544

(b) Transfer across backbones

	Wan2.2		LTX-2.5	
Configuration	SC ↑	TCM ↑	SC ↑	TCM ↑
Baseline	.8186	.9120	.7134	.8866
SLES-Train	.8388	.9193	.7397	.8919
SLES-Full	**.8440**	**.9300**	**.7409**	**.9018**

5.2 Ablations and Backbone Transfer

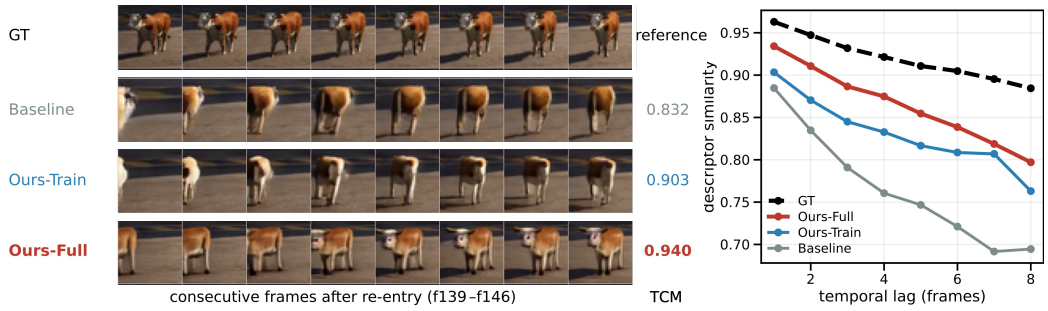
Supervision gains require spatial alignment. The scrambled-support control (Table 2) matches the number of additionally weighted tokens per frame but randomizes their locations, isolating the contribution of spatial alignment. It retains part of the improvement over Baseline, yet self-localized support improves SC by a further $+.0113$ (paired Wilcoxon, $p = 2.4 \times 10^{-4}$, $n = 176$), indicating that the gain arises from where supervision is applied rather than from its amount.

Sampling gains require localization. The inverted and random controls (Table 3a) preserve the correction mechanism and effective support size while altering only its spatial location. Self-localized sampling improves TCM by $+.0078$ over no sampling ($p = .013$), $+.0100$ over inverted support ($p = 1.5 \times 10^{-3}$), and $+.0082$ over random support ($p = 1.9 \times 10^{-3}$), with no significant SC change ($+.0029$, $p = .75$). Uniform correction attains marginally higher raw TCM (.9342 vs. .9300) but at a substantially higher full-frame cost: LPIPS increases by .0151 and SSIM decreases by .0114, versus $+.0002$ and a negligible change under localized sampling. Localization thus achieves selective temporal correction at minimal fidelity cost (Appendix E.1).

Ablations on backbones. With only the entity-support readout recalibrated per backbone (Table 3b), self-localized supervision on LTX-2.5 improves SC from .7134 to .7397 ($\Delta = +.0244$, $p = 8.7 \times 10^{-5}$, $n = 166$); the full method further reaches SC $= .7409$ and TCM $= .9018$, corresponding to improvements of $+.0239$ SC ($p = 8.5 \times 10^{-4}$) and $+.0173$ TCM ($p = 2.0 \times 10^{-4}$) over the LTX baseline. Full transfer results and calibration details are provided in Appendix E.4.



(a) Reappearance and subject consistency.



(b) Temporal recovery.

Figure 3: **Qualitative comparison results.** **(a)** Representative re-entry cases: baseline methods synthesize a visually plausible yet identity-inconsistent subject after re-entry, whereas SLES-Full preserves the identity and characteristic appearance of the original subject across subsequent frames. **(b)** Post-re-entry descriptor similarity as a function of temporal lag (1–8). Curves lying above the reference profile indicate excessive temporal stability (under-motion), whereas curves below it reflect greater temporal variation; SLES-Full most closely matches the target temporal profile.

6 Conclusion

We study long-horizon dynamic memory in video world models, where entities leave observation and later re-enter. Existing metrics conflate distinct failure modes, so we introduce a phase-aware evaluation of reappearance, subject consistency, and temporal consistency. Even with full history accessible, failures persist: self-attention reveals structured retrieval of historical subject information, but directly steering it does not reliably improve generation. We therefore propose Self-Localized Entity Supervision, which localizes both training supervision and inference-time stochastic correction using subject regions derived from the model's own text-conditioned signals. Experiments show that entity-aware training mainly improves subject recovery, while localized sampling improves temporal consistency with limited disturbance elsewhere, suggesting that long-horizon dynamic memory requires learning not only to retain historical information but also where and how to use it.

REFERENCES

Michael Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Haoyu Chen, Kaichen Zhou, Hang Hua, Kaile Zhang, Jingwen Qian, Wufei Ma, Haonan Chen, Chunjiang Liu, Yizhou Zhao, Xiaoyuan Wang, et al. Memobench: Benchmarking world modeling in dynamically changing environments. In *European Conference on Computer Vision*, pp. 658–676. Springer, 2026a.

Kaijin Chen, Dingkang Liang, Xin Zhou, Yikang Ding, Xiaoqiang Liu, Pengfei Wan, and Xiang Bai. Out of sight but not out of mind: Hybrid memory for dynamic video world models. *arXiv preprint arXiv:2603.25716*, 2026b.

Zicheng Duan, Jiatong Xia, Zeyu Zhang, Wenbo Zhang, Gengze Zhou, Chenhui Gou, Yefei He, Feng Chen, Xinyu Zhang, and Lingqiao Liu. Liveworld: Simulating out-of-sight dynamics in generative video world models. *arXiv preprint arXiv:2603.07145*, 2026.

Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, Ka Leong Cheng, Haojie Zhang, Jian Gao, Tianrui Feng, Yuzheng Liu, Yao Yao, Yinghao Xu, Xing Zhu, Yujun Shen, and Hao Ouyang. Infinite worlds with versatile interactions. *arXiv preprint arXiv:2607.07534*, 2026.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2025.

Xixi Hu, Keyang Xu, Bo Liu, Qiang Liu, and Hongliang Fei. Amo sampler: Enhancing text rendering with overshooting. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13157–13166. IEEE, 2025.

Junchao Huang, Xinting Hu, Boyao Han, Shaoshuai Shi, Zhuotao Tian, Tianyu He, and Li Jiang. Memory forcing: Spatio-temporal memory for consistent scene generation on minecraft. *arXiv preprint arXiv:2510.03198*, 2025.

Junchao Huang, Guian Fang, Shengju Qian, Xianghao Kong, Zhuoran Zhao, Wei Huang, Yihua Du, Zixin Zhang, Justin Cui, Yuchao Gu, et al. Solarwm: Open data and scalable training for long-horizon video world models. *arXiv preprint arXiv:2609.02886*, 2026.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Hyeonho Jeong, Chun-Hao P Huang, Jong Chul Ye, Niloy J Mitra, and Duygu Ceylan. Track4gen: Teaching video diffusion models to track points improves video generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7276–7287. IEEE, 2025.

Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024.

JoungBin Lee, Jaewoo Jung, Jongmin Lee, Tongmin Kim, Hyunsung Kim, Takuya Narihira, Kazumi Fukuda, Jahyeok Koo, Jisang Han, Yuki Mitsufuji, et al. Mvtrack4gen: Multi-view point tracking as geometric supervision for 4d video generation. *arXiv preprint arXiv:2606.26087*, 2026.

Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. Vmem: Consistent interactive video scene generation with surfel-indexed view memory. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25690–25699. IEEE, 2025.

Jisu Nam, Soowon Son, Dahyun Chung, Jiyoung Kim, Siyoon Jin, Junhwa Hur, and Seungryong Kim. Emergent temporal correspondences from video diffusion transformers. *Advances in Neural Information Processing Systems*, 38:123683–123727, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Runjia Qian, Zile Wang, Jihai Zhang, Kai Zou, Wei Yu, Jiaxing Li, Zexiang Liu, Yaokun Li, Fei Kang, Kaichen Huang, et al. Matrix-game 3.5: Enhancing real-time streaming interactive world models with patch memory. *arXiv preprint arXiv:2608.29910*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. *arXiv preprint arXiv:2512.14614*, 2025.

DreamX Team, Yancheng Bai, Rui Chen, Xiangxiang Chu, Rujing Dang, Hao Dou, Bingjie Gao, Qiwen Gu, Siyu Hong, Jiachen Lei, et al. Dreamx-world 1.0: A general-purpose interactive world model. *arXiv preprint arXiv:2606.16993*, 2026.

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

Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. In *International Conference on Learning Representations*, volume 2026, pp. 119186–119209, 2026a.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. *Advances in Neural Information Processing Systems*, 38:49371–49393, 2026b.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. *Advances in Neural Information Processing Systems*, 38:49632–49652, 2026.

Bowen Xue, Brandon Y Feng, Chenguo Lin, Yuchen Lin, Yujia Zeng, Lvmin Zhang, Maneesh Agrawala, Honglei Yan, and Panwang Pan. Ring forcing: Towards precise long-term memory for autoregressive video diffusion. In *European Conference on Computer Vision*, pp. 225–242. Springer, 2026.

Lvmin Zhang, Shengqu Cai, Muyang Li, Chong Zeng, Beijia Lu, Anyi Rao, Song Han, Gordon Wetzstein, and Maneesh Agrawala. Tinyhistory: Lightweight video history embeddings via two-stage context learning. *arXiv preprint arXiv:2512.23851*, 2025.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. *Advances in Neural Information Processing Systems*, 38:30546–30566, 2026a.

Shengjun Zhang, Zhang Zhang, Simin Huang, Zhenyu Tang, Hanyang Wang, Chensheng Dai, Min Chen, Yifan Li, Yuxin Li, Yingjie Chen, et al. Mbench: A comprehensive benchmark on memory capability for video world models. *arXiv preprint arXiv:2606.00793*, 2026b.

Songchun Zhang, Yaowei Li, Junhao Zhuang, Weiyang Jin, Haoyu Wang, Xin Lu, Yilang Sun, Shiyi Zhang, Haoran Li, Xiaoxiao Ma, Yuming Li, Yijun Liu, Yaofeng Su, Yanwen Ma, Haoyu Wu, Zihan Su, Yue Ma, Lvmin Zhang, Haoyang Huang, Zeyue Xue, Anyi Rao, and Nan Duan. Echowm: Open and enterable omnimodal world models. *arXiv preprint arXiv:2608.23189*, 2026c. URL https://arxiv.org/abs/2608.23189.

Haoyi Zhu, Haozhe Liu, Yuyang Zhao, Tian Ye, Junsong Chen, Jincheng Yu, Tong He, Song Han, and Enze Xie. Sana-wm: Efficient minute-scale world modeling with hybrid linear diffusion transformer. *arXiv preprint arXiv:2605.15178*, 2026.