EONWEAVE: WEAVING PERSISTENT EVIDENCE INTO CONSISTENT LONG-HORIZON WORLD GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon world generation requires a persistent model of the world, not merely short-range temporal continuity: a model must faithfully recover previously observed identities, scenes, and world states after they have disappeared from view for extended periods. We term this capability *revisitation consistency*. Existing autoregressive generators struggle in this regime because short-term conditions progressively accumulate errors, while historical evidence is either underutilized, redundantly stored, or inefficiently retrieved. To address these coupled limitations, we introduce **EonWeave**, a unified *trust–preserve–align* framework for integrating persistent evidence into long-horizon generation. **ARiSE** first calibrates source reliability by selectively re-noising the short-term condition while preserving clean historical evidence, preventing memory bypass and promoting adaptive reliance on persistent context. Building on this reliability asymmetry, **ReCAP** curates a fixed-capacity Persistent Scene Atlas through pose–appearance diversity-gain selection and block-adaptive projection, retaining complementary world-state evidence under bounded memory. **RaDAR** then converts the curated history into query-aligned references through noise-adaptive appearance–geometry routing and retrieval-to-reference distillation, enabling efficient utilization of heterogeneous historical evidence. Extensive experiments demonstrate state-of-the-art generation quality and revisitation consistency, with **EonWeave**'s PSNR advantage widening from approximately 3 dB in single-shot generation to 6 dB under extended autoregressive rollouts. These results establish **EonWeave** as a substantive advance toward consistent long-horizon world modeling.

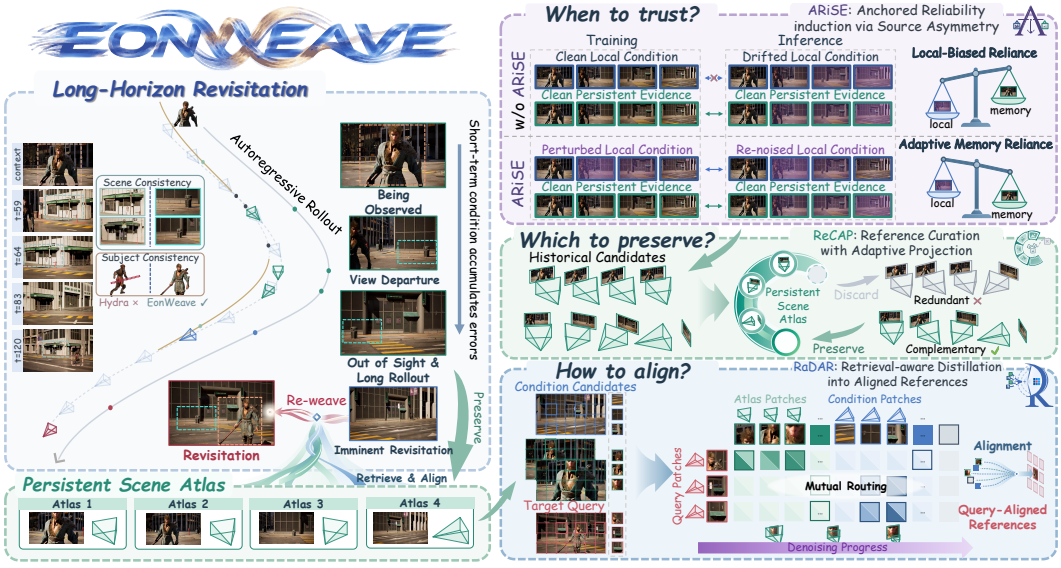


Figure 1: **Overview of EonWeave for revisitation-consistent long-horizon world generation.** EonWeave integrates persistent history through a *trust–preserve–align* pipeline: ARiSE calibrates reliance on historical evidence, ReCAP maintains a fixed-capacity Persistent Scene Atlas, and RaDAR distills retrieved evidence into a fixed-size, query-aligned reference grid, enabling faithful recovery of previously observed content after prolonged absence.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

World models aim to predict evolving environments, supporting embodied intelligence, interactive simulation, and long-form visual generation. Although video diffusion and flow-based models have improved visual quality and motion realism, most single-pass models remain constrained to short temporal windows because longer contexts incur considerable memory overhead (Yu et al., 2025). Long videos therefore rely on segment-level autoregressive rollout (Huang et al., 2025), where recursively reused generations accumulate error. A more fundamental challenge arises when camera motion, occlusion, or object dynamics hide previously observed content for extended intervals. When it reappears, the model must recover its established identity, appearance, spatial configuration, and plausible state despite the recent condition containing little direct evidence and substantial autoregressive drift. We term this capability *revisitation consistency*: preserving persistent world states across long observation gaps and viewpoint changes rather than merely frame-to-frame continuity.

Existing memory-augmented approaches retain additional frames, compress historical representations, or retrieve sparse tokens (Chen et al., 2026a; Xu et al., 2026; Peng et al., 2026), yet three challenges remain. First, clean adjacent conditions dominate conventional training and bypass persistent memory, whereas these conditions become increasingly unreliable during autoregressive inference. Second, relevant evidence may lie arbitrarily far in the past while memory must remain bounded, requiring selective preservation of distinctive world states. Third, retrieved evidence from different times, viewpoints, and spatial locations must be aligned with the current query rather than directly appended as heterogeneous tokens; high-noise diffusion timesteps further complicate correspondence as appearance degrades, increasing reliance on geometry. Revisitation-consistent generation therefore requires jointly determining *when* remote history should be trusted, *which* evidence should be preserved, and *how* it should be aligned with the current generation state.

To this end, we propose **EonWeave** (Fig. 1), which couples the short-term condition with a fixed-capacity *Persistent Scene Atlas* through a unified *trust–preserve–align* pipeline. To determine *when* persistent evidence should be trusted, **Anchored Reliability Induction via Source Asymmetry** (ARiSE) selectively re-noises the short-term condition while preserving clean history, preventing memory bypass and inducing reliability-aware source utilization. At inference, the generated condition is re-noised before short-term reuse while its uncorrupted counterpart remains available for storage. To determine *which* evidence should persist, **Reference Curation with Adaptive Projection** (ReCAP) maintains a bounded atlas whose *Atlas Records* retain visual latents, camera poses, and temporal coordinates. A pose–appearance diversity-gain criterion preserves observations that expand scene coverage, while block-adaptive projections adapt retained records to different generator stages. Finally, **Retrieval-aware Distillation into Aligned References** (RaDAR) determines *how* persistent evidence supports the query. Its *Noise-Adaptive Mutual Routing* balances appearance and geometric correspondence across denoising timesteps and suppresses ambiguous matches, after which retrieved evidence is distilled into one key–value reference per query patch, yielding fixed-size, spatially aligned context. Together, ARiSE, ReCAP, and RaDAR convert remote history into reliable, bounded, and query-aligned generative context, enabling established world states to be recovered after prolonged absence and remain coherent throughout autoregressive generation.

Extensive experiments under single-shot and autoregressive settings demonstrate state-of-the-art (SOTA) generation quality and revisitation consistency. On HM-World, EonWeave improves PSNR by 3 dB when generating 77 frames from 77 conditioning frames. This advantage further widens as the generation horizon extends, reaching approximately **6** dB on Mira-World over a 330-frame rollout from only 33 observed frames. Our main contributions are summarized as follows:

- We propose **EonWeave**, a unified framework for revisitation-consistent long-horizon world generation that frames persistent-history use as a *trust–preserve–align* process, allowing previously observed world states to be recovered after prolonged absence.
- We introduce **ARiSE**, an asymmetric reliability-induction strategy that addresses the train–test source-reliability mismatch. By selectively re-noising the short-term condition while preserving clean historical evidence, ARiSE prevents memory bypass and induces adaptive reliance on persistent context without auxiliary source-selection supervision.
- We develop **ReCAP**, which maintains a fixed-capacity *Persistent Scene Atlas* through pose–appearance diversity-gain curation and block-adaptive projection, preserving complementary world-state evidence under bounded memory throughout rollout.

- We present **RaDAR**, which combines appearance and camera geometry through noise-adaptive mutual routing and distills retrieved patches into a fixed-size, query-aligned reference grid, converting heterogeneous history into structured generative context while bounding attention complexity and suppressing unreliable evidence.
- Extensive experiments on HM-World and Mira-World establish substantial advantages in both generation quality and revisitation consistency, with the performance margin further widening under extended autoregressive rollouts. This sustained and horizon-amplified performance separation substantiates the *trust–preserve–align* design as a principled framework for exploiting persistent evidence across long temporal gaps, establishing **EonWeave** as a substantive advance toward revisitation-consistent long-horizon world modeling.

2 Related Work

2.1 Long-Horizon Video World Models

Video diffusion and flow-based models operate within finite temporal windows, motivating causal and autoregressive extensions for long-horizon generation. Prior work expands temporal coverage through causal prediction, frame-context packing, variable-length conditioning, and self-generated or self-augmented histories (Chen et al., 2024; Zhang et al., 2025; Song et al., 2025; Huang et al., 2025; Shen et al., 2026). However, when previously observed content disappears for extended intervals, its state is progressively diluted in the evolving context, causing identity drift and inconsistent recovery upon return. EonWeave instead targets *revisitation consistency* by persistently preserving historical world states and selectively reactivating them when local context becomes insufficient.

2.2 Persistent Memory and Historical Retrieval

Memory-augmented video models preserve history through frame retrieval, pose-aware memory, compressed latents, or sparse spatiotemporal retrieval. Context-as-Memory retrieves frames by field-of-view overlap, WorldMem associates memories with poses and timestamps, and HyDRA retrieves compressed spatiotemporal tokens (Yu et al., 2025; Xiao et al., 2025; Chen et al., 2026a); related work explores surfel-indexed memory, latent compression, implicit context compression, and time-aware positional encoding (Li et al., 2025; Hong et al., 2025; Peng et al., 2026; Xu et al., 2026). Memory utilization, bounded retention, and query-aligned readout are typically treated separately. EonWeave unifies them through *trust–preserve–align*: ARiSE encourages reliance on persistent evidence when the short-term condition is degraded, ReCAP curates a fixed-capacity pose–appearance atlas, and RaDAR distills retrieved candidates into a fixed-size, query-aligned reference grid.

3 Method

3.1 Problem Formulation

We consider *revisitation-consistent long-horizon world generation*. Given an initial video segment V_0, the model autoregressively generates future segments $\{V_n\}_{n=1}^{N}$, producing $V_{0:N} = V_0 \,\|\, V_1 \,\|\, \cdots \,\|\, V_N$, where $\|$ denotes temporal concatenation. During long rollouts, previously observed content may remain absent because of camera motion, occlusion, or object dynamics; when such content reappears, the model should recover its established identity, appearance, and spatial structure while remaining compatible with the intervening world evolution. At step n, the latest segment $C_n \equiv V_n$ provides the short-term condition for local motion continuity, while earlier observations are summarized by a fixed-capacity *Persistent Scene Atlas* $\mathcal{A}_n$, containing only history preceding C_n and initialized with $\mathcal{A}_0 = \emptyset$. The autoregressive generation and atlas update are defined as:

$$V_{n+1} \sim p_\theta(V_{n+1} \mid C_n, \mathcal{A}_n)\,, \qquad \mathcal{A}_{n+1} = \mathcal{W}(\mathcal{A}_n, V_n; K)\,, \qquad |\mathcal{A}_{n+1}| \le K, \tag{1}$$

where $\mathcal{W}$ is a capacity-controlled write operator and K denotes the maximum number of stored *Atlas Records*. Once V_n has served as the active condition, it becomes eligible for atlas insertion, avoiding redundant storage of the current short-term context. Accordingly, we have:

$$p_\theta(V_{1:N} \mid V_0) = \prod_{n=0}^{N-1} p_\theta(V_{n+1} \mid C_n, \mathcal{A}_n)\,. \tag{2}$$

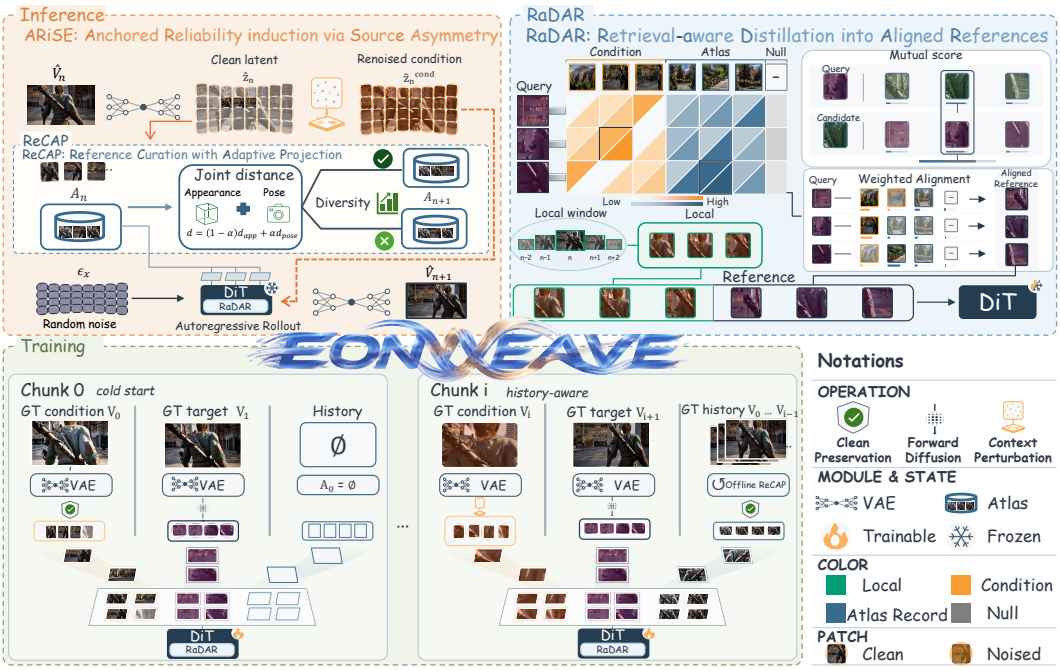


Figure 2: **Overview of EonWeave.** EonWeave integrates **ARiSE**, **ReCAP**, and **RaDAR** into a unified *trust–preserve–align* pipeline for exploiting persistent historical evidence and enabling revisitation-consistent long-horizon world generation through query-aligned generative context.

The objective is to preserve local temporal continuity and persistent world states without retaining the complete generation history, requiring the model to determine *when* remote evidence should be trusted over a degraded short-term condition, *which* historical observations should be preserved under bounded capacity, and *how* to align them with the current generation state, as shown in Fig. 2.

3.2 ARiSE: Anchored Reliability Induction via Source Asymmetry

The short-term condition C_n captures immediate dynamics, whereas the Persistent Scene Atlas $\mathcal{A}_n$ preserves earlier world states. Conventional training, however, presents both sources from ground-truth video, allowing the temporally adjacent C_n to dominate prediction and the atlas pathway to be bypassed (Yu et al., 2025). During autoregressive inference, this reliability structure reverses: the reused condition accumulates generation errors while historical evidence can remain comparatively stable, creating a source-reliability mismatch precisely when revisited content must be recovered.

3.2.1 Asymmetric Reliability Training

To induce reliability-aware source arbitration, *Anchored Reliability Induction via Source Asymmetry (ARiSE)* selectively degrades the short-term condition C_n while preserving $\mathcal{A}_n$ as a reliable historical anchor. Let z_n^{cond} denote the latent of C_n. For $n > 0$, the training condition is constructed as:

$$\bar{z}_n^{\text{cond}} = (1 - b_n\sigma_c)z_n^{\text{cond}} + b_n\sigma_c\epsilon_c,\ b_n \sim \text{Bernoulli}(p_{\text{aug}}),\ \sigma_c \sim \mathcal{U}(0, \sigma_{\max}),\ \epsilon_c \sim \mathcal{N}(0, I), \quad (3)$$

while the initial observed condition remains clean, i.e., $\bar{z}_0^{\text{cond}} = z_0^{\text{cond}}$. Thus, ARiSE exposes the model to a continuum of short-term reliability levels while leaving historical evidence uncorrupted. The training atlas $\mathcal{A}_n^{\text{clean}}$ is constructed from clean ground-truth history as a reliable historical anchor. Let z_{n+1}^{tgt} denote the clean latent of the target segment. Under the linear Flow Matching path,

$$\begin{gathered} x_t = (1-t)z_{n+1}^{\text{tgt}} + t\epsilon_x, \qquad v^\star = \epsilon_x - z_{n+1}^{\text{tgt}}, \qquad t \sim \mathcal{U}(0,1), \quad \epsilon_x \sim \mathcal{N}(0, I), \\ \mathcal{L}_{\text{ARiSE}} = \mathbb{E}\left[w(t)\left\|v^\star - v_\theta\left(x_t, t, \bar{z}_n^{\text{cond}}, \mathcal{A}_n^{\text{clean}}\right)\right\|_2^2\right], \end{gathered} \quad (4)$$

where $w(t)$ follows the timestep weighting of the base Flow Matching objective. As the reliability of $\bar{z}_n^{\text{cond}}$ decreases, information weakened in the local condition remains recoverable from $\mathcal{A}_n^{\text{clean}}$. The

standard generation loss therefore acquires a direct learning signal for persistent evidence utilization, inducing continuous source arbitration without auxiliary supervision or explicit source selection.

3.2.2 INFERENCE-TIME RELIABILITY ALIGNMENT

During autoregressive inference, each generated segment becomes the short-term condition for the next rollout step and may progressively accumulate prediction errors. ARiSE therefore maps its reliability back into the range encountered during training by re-noising the generated latent $\hat{z}_n$:

$$\tilde{z}_n^{\mathrm{cond}} = (1-\sigma^\star)\hat{z}_n + \sigma^\star \epsilon_n, \qquad \epsilon_n \sim \mathcal{N}(0, I), \qquad 0 \le \sigma^\star \le \sigma_{\max}, \tag{5}$$

where $\tilde{z}_n^{\mathrm{cond}}$ conditions the next generation step, while the uncorrupted $\hat{z}_n$ is retained for subsequent atlas curation. This asymmetric reuse attenuates the influence of accumulated errors in short-term conditioning while preserving persistent evidence, maintaining local dynamics through the condition and long-term scene state through the atlas. This inference-time reliability alignment limits error propagation and strengthens recovery of previously established world states upon revisitation.

3.3 RECAP: REFERENCE CURATION WITH ADAPTIVE PROJECTION

ARiSE determines when persistent evidence should be trusted; ReCAP determines which evidence should survive under bounded memory. Retaining the complete history is increasingly redundant and expensive, whereas indiscriminate compression risks discarding rare states required for later revisitation. We therefore introduce *Reference Curation with Adaptive Projection (ReCAP)*, which maintains a fixed-capacity *Persistent Scene Atlas* and adapts its retained observations to the generator's internal representation spaces. Each *Atlas Record* is defined as $a_i = (Z_i, \pi_i, u_i)$, where Z_i is a variational autoencoder (VAE) latent, π_i its camera pose, and u_i its global latent-frame index. The atlas contains only observations preceding C_n, with $\mathcal{A}_0 = \emptyset$ and $|\mathcal{A}_n| \le K$. Training atlases are constructed from ground-truth history, while inference updates them online, both based on pose–appearance diversity; detailed procedures are provided in Algorithms 1–3 in Appendix B.

3.3.1 POSE–APPEARANCE ATLAS CURATION

ReCAP retains observations that expand both visual and geometric coverage. The distance between two Atlas Records combines appearance and viewpoint diversity and is defined as

$$d(a_i, a_j) = (1-\alpha)\underbrace{(1-\cos(Z_i, Z_j))}_{\text{appearance diversity}} + \alpha \underbrace{\frac{1}{2}\left[\frac{\theta_{ij}}{\pi} + \min\left(\frac{\|\mathbf{t}_i - \mathbf{t}_j\|_2}{s}, 1\right)\right]}_{\text{viewpoint diversity}}, \tag{6}$$

where θ_{ij} is the geodesic rotation angle between the cameras, $\mathbf{t}_i$ and $\mathbf{t}_j$ are their translations, s is the translation normalization scale, and α balances appearance and viewpoint diversity. Atlas coverage is quantified by nearest-neighbor dispersion over the retained records under a fixed budget,

$$D(\mathcal{A}) = \sum_{a_i \in \mathcal{A}} \min_{\substack{a_j \in \mathcal{A} \\ j \neq i}} d(a_i, a_j). \tag{7}$$

A new record a_{new} is inserted directly before capacity is reached. Once the atlas is full, ReCAP evaluates the gain induced by replacing each record to expand the current pose–appearance coverage:

$$\gamma_i = D(\mathcal{A} \setminus \{a_i\} \cup \{a_{\mathrm{new}}\}) - D(\mathcal{A}), \qquad i^\star = \arg\max_i \gamma_i. \tag{8}$$

Replacement occurs only when $\gamma_{i^\star} > 0$, admitting new observations only when they expand pose–appearance coverage and thereby promoting complementary world states under a fixed budget.

3.3.2 BLOCK-ADAPTIVE PROJECTION

Stored VAE latents and intermediate generator states occupy different representation spaces. ReCAP therefore adapts each Atlas Record independently at transformer block ℓ:

$$\mathbf{k}_i^{(\ell)} = W_{\mathrm{atlas}}^{(\ell)} \mathcal{S}(Z_i) + W_{\mathrm{cam}}^{(\ell)} e(\pi_i), \qquad \mathbf{v}_i^{(\ell)} = W_{\mathrm{atlas}}^{(\ell)} \mathcal{S}(Z_i), \tag{9}$$

where $\mathcal{S}(\cdot)$ tokenizes the stored latent and $e(\pi_i)$ encodes its camera pose. The key incorporates viewpoint information for correspondence estimation, while the value preserves visual content; block-specific projections allow the same historical record to support different representation levels. ReCAP thus couples bounded pose–appearance curation with block-adaptive readout, providing a compact persistent write–read interface that makes retained evidence consumable by the generator.

3.4 RaDAR: Retrieval-aware Distillation into Aligned References

ReCAP determines which observations remain available; RaDAR determines how they support the current generation state. Existing sparse-retrieval methods directly append selected historical tokens to attention (Chen et al., 2026a), causing context size to scale with retrieval fan-out while mixing evidence across times, viewpoints, and locations. We therefore introduce *Retrieval-aware Distillation into Aligned References (RaDAR)*, which treats retrieval only as evidence selection and distills the condition and atlas evidence for each query patch into a single key–value reference. The resulting reference grid is aligned with the current spatial layout and remains fixed in size.

3.4.1 Patch-Level Historical Routing

For clarity, rollout-step and transformer-block indices are omitted. Let the current hidden states contain P_q query patches, with patch p comprising N_p tokens $\{\mathbf{x}_q^{(p,n)}\}_{n=1}^{N_p}$. Each query accesses two complementary sources: an intra-segment local context $(\mathbf{K}_{\text{local}}, \mathbf{V}_{\text{local}})$ (Chen et al., 2026a), which preserves short-range temporal coherence, and historical candidates from $\mathcal{J}_{\text{cond}} \cup \mathcal{J}_{\text{atlas}}$, indexing patches from the latest condition C_n and patches projected from $\mathcal{A}_n$ by ReCAP, respectively. For candidate m with tokens $\{\mathbf{x}_k^{(m,n)}\}_{n=1}^{N_m}$, learned projections W_{rq} and W_{rk} map query and historical states into a routing space independent of the generative key–value projections:

$$\mathbf{r}_q^{(p)} = \frac{1}{N_p}\sum_{n=1}^{N_p} W_{rq}\mathbf{x}_q^{(p,n)}, \qquad \mathbf{r}_k^{(m)} = \frac{1}{N_m}\sum_{n=1}^{N_m} W_{rk}\mathbf{x}_k^{(m,n)}. \tag{10}$$

3.4.2 Noise-Adaptive Mutual Routing

Appearance correspondence becomes less reliable at highly corrupted denoising states, whereas camera geometry remains stable. *Noise-Adaptive Mutual Routing (NAMR)* adapts the contributions of appearance and geometry according to the normalized corruption level $\bar{t} \in [0,1]$:

$$s_{p,m} = \big(1-\lambda(t)\big)\cos\Big(\mathbf{r}_q^{(p)}, \mathbf{r}_k^{(m)}\Big) + \lambda(t)\cos\Big(\mathbf{g}_q^{(p)}, \mathbf{g}_k^{(m)}\Big), \qquad \lambda(t) = \text{sigmoid}(a\bar{t}+b), \tag{11}$$

where $\mathbf{g}_q^{(p)}$ and $\mathbf{g}_k^{(m)}$ encode camera-ray and patch-location geometry, and a, b are learned scalars. Thus, routing relies more strongly on geometry under severe corruption and progressively shifts toward appearance as denoising proceeds. NAMR further suppresses ambiguous correspondences through bidirectional normalization. With learned temperature τ and dustbin logit z,

$$\pi_{m|p} = \frac{\exp(s_{p,m}/\tau)}{\exp(z/\tau) + \sum_{j\in\mathcal{J}_{\text{cond}}\cup\mathcal{J}_{\text{atlas}}}\exp(s_{p,j}/\tau)}, \quad \pi_{\varnothing|p} = \frac{\exp(z/\tau)}{\exp(z/\tau) + \sum_{j\in\mathcal{J}_{\text{cond}}\cup\mathcal{J}_{\text{atlas}}}\exp(s_{p,j}/\tau)},$$
$$\pi_{p|m} = \frac{\exp(s_{p,m}/\tau)}{\sum_{i=1}^{P_q}\exp(s_{i,m}/\tau)}, \qquad \rho_{p,m} = \pi_{m|p}\pi_{p|m}, \qquad \rho_{p,\varnothing} = \pi_{\varnothing|p}. \tag{12}$$

Here, $\pi_{m|p}$ measures the preference of query p for candidate m, whereas $\pi_{p|m}$ measures how specifically m corresponds to p among all queries. Their product $\rho_{p,m}$ favors mutually distinctive matches, while $\rho_{p,\varnothing}$ permits abstention when no reliable historical correspondence exists.

3.4.3 Query-Aligned Reference Distillation

Condition and atlas candidates are selected independently for each query patch p:

$$\mathcal{I}_p = \text{TopIdx}_{r_c}(\{\rho_{p,m}\}_{m\in\mathcal{J}_{\text{cond}}}) \cup \text{TopIdx}_{r_a}(\{\rho_{p,m}\}_{m\in\mathcal{J}_{\text{atlas}}}), \tag{13}$$

where r_c and r_a denote the retrieval budgets for condition and atlas candidates, respectively; the atlas budget reserves a dedicated retrieval quota for persistent evidence, preventing locally proximate condition candidates from monopolizing the selected support. Let $\bar{\mathbf{k}}_m$ and $\bar{\mathbf{v}}_m$ denote the pooled generative key and value of candidate m, where spatiotemporal rotary position embeddings (RoPE) are applied to the source keys before pooling. For atlas candidates, the temporal coordinates are given by the true historical offsets maintained by ReCAP. The retained mutual scores and dustbin mass are then jointly normalized over the selected candidate support as:

$$\zeta_p = \rho_{p,\varnothing} + \sum_{j\in\mathcal{I}_p}\rho_{p,j}, \qquad \alpha_{p,m} = \frac{\rho_{p,m}}{\zeta_p}, \qquad \alpha_{p,\varnothing} = \frac{\rho_{p,\varnothing}}{\zeta_p}. \tag{14}$$

Table 1: **Single-shot generation results on HM-World.**

Method	PSNR↑	SSIM↑	LPIPS↓	Subj. Cons.↑	Backg. Cons.↑
ReCamMaster (Bai et al., 2025)	19.426	0.553	0.388	0.922	0.938
Self-Forcing (Huang et al., 2025)	17.054	0.476	0.467	0.859	0.907
DFoT (Song et al., 2025)	17.693	0.482	0.410	0.893	0.913
Context-as-Memory (Yu et al., 2025)	18.921	0.530	0.342	0.911	0.922
WorldMem (Xiao et al., 2025)	18.962	0.583	0.436	0.903	0.937
UCM (Xu et al., 2026)	15.475	0.441	0.537	0.873	0.931
HyDRA (Chen et al., 2026a)	19.201	0.587	0.383	0.913	0.932
EonWeave (Full)	21.669	0.659	0.305	0.924	0.937
EonWeave-DMD	22.463	0.683	0.300	0.917	0.939

The selected evidence is then distilled into a single reference pair for the current query patch p:

$$\mathbf{k}_{\text{ref}}^{(p)} = \sum_{m \in \mathcal{I}_p} \alpha_{p,m} \bar{\mathbf{k}}_m + \alpha_{p,\varnothing} \mathbf{k}_\varnothing, \qquad \mathbf{v}_{\text{ref}}^{(p)} = \sum_{m \in \mathcal{I}_p} \alpha_{p,m} \bar{\mathbf{v}}_m + \alpha_{p,\varnothing} \mathbf{v}_\varnothing, \tag{15}$$

where $(\mathbf{k}_\varnothing, \mathbf{v}_\varnothing)$ is a learnable null reference. Reliable candidates are distilled according to mutual-routing confidence, whereas uncertain evidence collapses toward the null state. Stacking one reference per query patch yields $\mathbf{K}_{\text{ref}}, \mathbf{V}_{\text{ref}} \in \mathbb{R}^{P_q \times d}$, which are combined with the local context:

$$\mathbf{K}_{\text{ctx}} = [\mathbf{K}_{\text{local}}; \mathbf{K}_{\text{ref}}], \qquad \mathbf{V}_{\text{ctx}} = [\mathbf{V}_{\text{local}}; \mathbf{V}_{\text{ref}}], \qquad \mathbf{Y} = \text{Attn}(\mathbf{Q}, \mathbf{K}_{\text{ctx}}, \mathbf{V}_{\text{ctx}}). \tag{16}$$

Regardless of how many historical candidates are examined or selected, the resulting historical context contains exactly P_q distilled references and therefore remains fixed in size, converting heterogeneous history into bounded, query-aligned evidence for revisitation-consistent generation.

4 Experiments

4.1 Experimental Setup

Datasets and Protocols. We evaluate EonWeave on two complementary benchmarks targeting controlled revisitation and long-horizon world consistency. **HM-World** (Chen et al., 2026a) is a synthetic benchmark at 480×832 with decoupled camera and subject trajectories. We adopt the HyDRA *single-shot* protocol (77 conditioning $\rightarrow$ 77 generated frames) and an *autoregressive* protocol with 33 conditioning frames followed by three recursively generated 33-frame segments. **Mira-World** (Ju et al., 2024) comprises 12,390 sequences of 330 frames each at 640×360, partitioned into 33-frame segments and autoregressively completed from the first segment, providing a substantially longer and more diverse stress test of sustained generation fidelity and persistent world-state consistency. More details are provided in Appendix G.1.

Baselines. We compare EonWeave with UCM (Xu et al., 2026), ReCamMaster (Bai et al., 2025), DFoT (Song et al., 2025), Context-as-Memory (Yu et al., 2025), WorldMem (Xiao et al., 2025), HyDRA (Chen et al., 2026a), and Self-Forcing (Huang et al., 2025), spanning camera-conditioned, context-based, memory-augmented, and autoregressive paradigms. We further construct **AR-HyDRA**, an autoregressive extension of HyDRA. All baselines are reproduced using official implementations and checkpoints under identical evaluation protocols for a fair comparison.

Metrics. We evaluate three complementary axes: *visual fidelity*, using frame-averaged PSNR, SSIM, and LPIPS; *revisitation consistency*, using VBench Subject and Background Consistency (Huang et al., 2024); and *long-horizon degradation*, using the normalized PSNR drift index $\Delta_{\text{Dft}}^{\text{Qal}}$. For each evaluation sequence, we compute the relative decrease in mean frame-wise PSNR from the first generated chunk to the final generated chunk, and then average these relative decreases across sequences. Lower values indicate less degradation.

Implementation Details. EonWeave is built on Wan2.1-T2V-1.3B. The Persistent Scene Atlas uses K=18 records with α=0.4; ARiSE uses p_{aug}=0.5, $\sigma_{\max}$=0.5, and $\sigma^\star$=0.15; and RaDAR uses 2×2 patches, r_c=r_a=4, and a 5-frame local window. We train for 15,000 steps on 32 NVIDIA A800 GPUs with a global batch size of 32, using AdamW with a learning rate of 10^{-4}. Distribution matching distillation (DMD) further yields the four-step variant **EonWeave-DMD** under a reduced sampling regime without changing the long-horizon generation protocol described above.

Table 2: **Autoregressive generation results on HM-World and Mira-World.**

Method	HM-World			Mira-World		
	PSNR↑	Subj. Cons.↑	$\Delta^{\text{Qal}}_{\text{Dft}}$ ↓	PSNR↑	Subj. Cons.↑	$\Delta^{\text{Qal}}_{\text{Dft}}$ ↓
ReCamMaster	20.220	0.951	0.243	19.449	0.950	0.395
Self-Forcing	17.017	0.893	0.253	19.118	0.902	0.359
WorldMem	19.248	0.959	0.259	15.867	0.948	0.581
UCM	16.622	0.952	0.255	16.873	0.934	0.491
AR-HyDRA	21.666	0.954	0.253	18.550	0.949	0.460
EonWeave (Full)	22.604	0.957	0.218	23.571	0.968	0.337
EonWeave-DMD	23.685	0.960	0.214	25.222	0.985	0.321

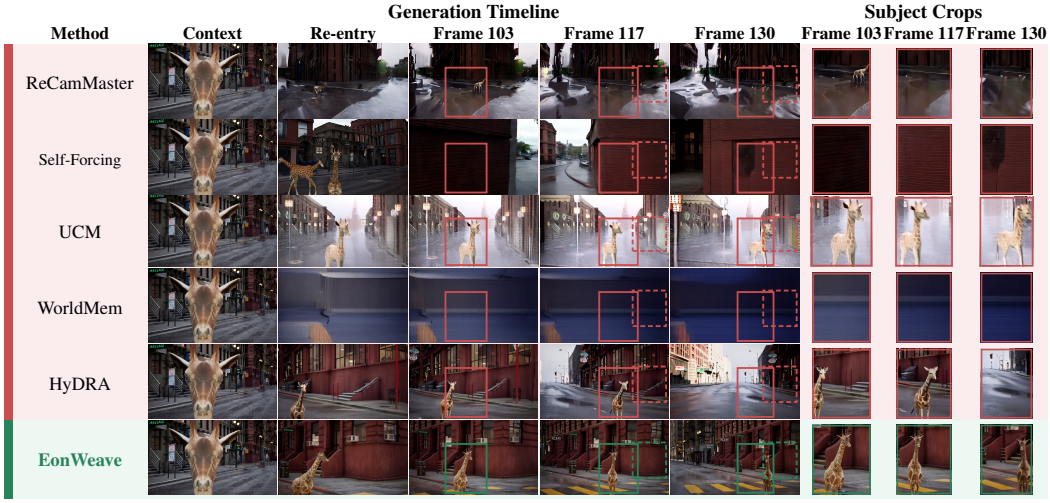


Figure 3: **Qualitative comparison on HM-World.** EonWeave better preserves subject identity and scene structure upon revisitation than existing methods. Crops highlight identity preservation. Solid and dashed boxes denote subject and scene regions, respectively. red: failure; green: consistency.

4.2 Main Results

Single-shot generation. Under the HM-World single-shot protocol, EonWeave substantially improves generation fidelity, yielding a 3.037 dB PSNR gain and a 0.096 SSIM improvement over prior methods, while reducing LPIPS by 12.3% (Tab. 1). The concurrent gains in pixel accuracy, structural similarity, and perceptual quality demonstrate a broad enhancement in conditional generation capability. This improvement is enabled by RaDAR's query-adaptive retrieval, which efficiently selects and aligns informative evidence from the observed condition with the current generation target, allowing EonWeave to exploit the available conditioning context more effectively. EonWeave further improves subject and background consistency, indicating more faithful recovery of previously observed world states under revisitation. Collectively, these results establish EonWeave's advantage in both high-fidelity generation and persistent world-state consistency.

Long-horizon autoregressive generation. Tab. 2 reveals an increasingly pronounced advantage of EonWeave under long-horizon autoregressive generation. Relative to AR-HyDRA, EonWeave improves PSNR by 2.02 dB on HM-World and 6.67 dB on the longer Mira-World benchmark, while increasing subject consistency by up to 0.036 and reducing $\Delta^{\text{Qal}}_{\text{Dft}}$ by as much as 30.3%. More importantly, this performance margin progressively widens as the rollout horizon extends, indicating stronger resistance to rollout-induced degradation rather than merely a higher initial generation quality. ARiSE mitigates the reliability mismatch arising from recursively reused conditions, whereas ReCAP preserves complementary historical evidence across prolonged observation gaps; together, they suppress compounding autoregressive errors while retaining access to previously established world states. Consequently, EonWeave maintains substantially higher fidelity and semantic consistency throughout extended generation, underscoring the importance of persistent historical modeling for sustaining generation quality as short-term autoregressive conditions lose reliability.

Qualitative results. Fig. 3 presents qualitative comparisons on HM-World. When previously observed content re-enters the field of view after an extended absence, EonWeave more faithfully

recovers established subject identity, object presence, and scene structure. In contrast, competing methods exhibit identity drift, object omission or duplication, and structural distortions, producing conspicuous deviations from earlier observations. EonWeave instead reconstructs revisited content with markedly stronger correspondence to its previously established world state. These comparisons provide direct perceptual evidence of improved revisitation consistency, complementing the quantitative gains in reconstruction fidelity and subject consistency. Further analysis of long-horizon performance and qualitative results is provided in Appendix G.

4.3 Ablation Study

We conduct leave-one-out ablations on Mira-World to quantify the contributions of ReCAP, ARiSE, and RaDAR. As shown in Tab. 3, removing each component in turn while holding the remaining pipeline fixed directly measures the respective contributions of EonWeave's *preserve*, *trust*, and *align* mechanisms.

Table 3: **Ablation study on Mira-World.**

Method	PSNR↑	LPIPS↓	Subj. Cons.↑	$\Delta_{\text{Dft}}^{\text{Qal}}$ ↓
w/o ReCAP	20.678	0.484	0.954	0.417
w/o ARiSE	20.735	0.483	0.957	0.312
w/o RaDAR	22.018	0.497	0.949	0.388
EonWeave (Full)	23.571	0.439	0.968	0.337
EonWeave-DMD	25.222	0.394	0.985	0.321

Removing ReCAP produces the largest collapse in reconstruction fidelity, reducing PSNR by 2.893 dB while increasing LPIPS by 10.3% and $\Delta_{\text{Dft}}^{\text{Qal}}$ by 23.7% (Tab. 3). Since ARiSE and RaDAR remain intact, this degradation shows that downstream source arbitration and retrieval cannot compensate for historical states that are no longer retained. ReCAP therefore acts as the persistence bottleneck of EonWeave: its diversity-gain curation preserves complementary world-state evidence under a fixed memory budget, sustaining both reconstruction quality and rollout stability.

With this persistent evidence still available, removing ARiSE nevertheless incurs a comparable 2.836 dB PSNR loss, together with a 10.1% increase in LPIPS and reduced subject consistency. The degradation therefore arises not from missing memory, but from ineffective source utilization: without asymmetric reliability induction, the model remains biased toward short-term conditions even as their reliability deteriorates. ARiSE converts preserved history into effective conditioning by calibrating reliance between local and persistent evidence, thereby preventing short-term dominance from suppressing informative memory.

Replacing RaDAR with conventional frame-level historical attention yields the smallest PSNR loss among the three ablations (1.553 dB), indicating that ReCAP and ARiSE together already establish a strong persistent-memory substrate by preserving informative history and calibrating reliance on it. However, this variant incurs the most severe perceptual and identity degradation, increasing LPIPS by 13.3%, producing the lowest subject consistency, and raising $\Delta_{\text{Dft}}^{\text{Qal}}$ by 15.1%. This contrast highlights the distinct role of RaDAR: persistent evidence must not only be retained and trusted, but also precisely aligned with the current query. Through noise-adaptive appearance–geometry routing and retrieval-to-reference distillation, RaDAR converts preserved history into structured, query-aligned generative context.

Together, the complementary degradation patterns of ReCAP, ARiSE, and RaDAR substantiate the *trust–preserve–align* design as a tightly coupled solution to persistent evidence retention, reliability calibration, and historical alignment. Further ablation analysis and hyperparameter sensitivity studies are provided in Appendix G.

5 Conclusion

This work establishes *revisitation consistency* as a central capability for long-horizon world generation and introduces EonWeave, a unified *trust–preserve–align* framework for persistent world-state modeling. By jointly calibrating source reliability, preserving complementary historical states under bounded memory, and aligning persistent evidence with current state, EonWeave enables faithful recovery of previously observed content across prolonged temporal gaps. Extensive experiments demonstrate SOTA generation quality and revisitation consistency, with its advantage becoming increasingly pronounced as the autoregressive horizon extends. This horizon-amplified superiority substantiates *trust–preserve–align* as a principled paradigm for persistent evidence integration and establishes EonWeave as a substantive advance toward consistent long-horizon world modeling.

LIMITATIONS AND FUTURE WORK

Bounded Historical Coverage. EonWeave achieves persistent world-state modeling with a fixed-capacity Persistent Scene Atlas, keeping historical storage bounded as the rollout grows. ReCAP mitigates the resulting capacity constraint by preferentially retaining pose–appearance complementary states, while RaDAR avoids forcing unreliable correspondences when suitable evidence is absent. Nevertheless, under arbitrarily long and highly diverse histories, some previously observed states may eventually fall outside the retained atlas. A natural extension is therefore to improve historical coverage under the same memory budget through more information-efficient persistence and consolidation mechanisms.

Routing Scalability. EonWeave keeps both persistent memory and the final query-aligned reference context bounded, preventing attention complexity from scaling with the complete rollout history. RaDAR nevertheless evaluates the available condition and atlas candidates before distillation, whose cost may become more relevant as future systems adopt substantially larger persistent memories. Extending EonWeave with more scalable candidate search and hierarchical routing could further enlarge historical coverage while preserving the compact *trust–preserve–align* interface.

Underdetermined Long-Horizon Prediction. Long-horizon generation is intrinsically underdetermined when future world states have not yet been observed. In Mira-World, only the first 33 frames are provided, while the remaining 297 frames must be generated autoregressively; newly appearing objects, unseen regions, and subsequent scene changes may therefore admit multiple plausible continuations. EonWeave specifically strengthens consistency for previously established world states, but evaluation against a single recorded future inevitably entangles revisitation fidelity with open-ended future prediction. Future work may therefore consider protocols that retain long-horizon revisitation while providing stronger evidence for the states being evaluated, enabling a more precise characterization of persistent world-state consistency.

AI USE STATEMENT

Generative AI tools were used solely for language polishing and literature summarization. AI-assisted editing was limited to improving the clarity, readability, and presentation of author-written text, without generating scientific content or substantive manuscript material. AI-generated literature summaries were used only as reading aids and were verified against the original sources. AI tools were not used for literature retrieval, reference generation, experimental design, data analysis, scientific interpretation, or formulation of the paper's claims and conclusions. All references were selected and verified by the authors, and no AI-generated or fabricated citations were incorporated. The authors reviewed all AI-assisted edits and retain full responsibility for the final manuscript and its scientific content.

ETHICS STATEMENT

This work advances long-horizon world generation by improving *revisitation consistency*: the ability to preserve and recover previously established subjects, objects, and scene states after extended temporal gaps. EonWeave is developed as a research framework for persistent visual modeling, generative world models, simulation, and interactive environments, where stronger temporal persistence is essential for maintaining coherent world states over long horizons. Our study does not involve behavioral interventions or recruited human participants, and the proposed methodology is intended for research and other lawful, non-deceptive applications consistent with applicable licenses, privacy requirements, and responsible AI practices.

By preserving historical evidence and recovering it upon revisitation, EonWeave produces substantially more coherent long-horizon generations, particularly in subject identity, appearance, object presence, and scene structure. This improved persistence should nevertheless be distinguished from factual correctness: like other generative models, EonWeave may synthesize details or transitions that are plausible but not physically or historically authentic. Generated videos should therefore be treated as synthetic content rather than documentary evidence of real people, places, or events. For

applications in which synthetic outputs could be mistaken for authentic observations, clear disclosure and appropriate provenance or content-authentication mechanisms are advisable.

EonWeave builds on a pretrained video-generation backbone and may inherit biases or representation limitations present in the underlying model and training data. Improved long-horizon consistency strengthens the persistence of generated representations but does not itself guarantee fairness or factual accuracy. Accordingly, human-centered applications should avoid treating generated attributes as evidence about real individuals or demographic groups. Likewise, use of real-world visual data should respect applicable dataset licenses, copyright, privacy, consent, and data-governance requirements, particularly when identifiable individuals or sensitive content are involved.

The same ability to maintain identities and scene states over extended sequences could, as with general-purpose video-generation systems, be applied in misleading contexts. EonWeave is not designed for impersonation, deception, surveillance, or fabrication of real-world evidence. Responsible deployment should therefore incorporate safeguards appropriate to the application, including transparent disclosure of synthetic media, provenance tracking, access controls where warranted, and compliance with relevant legal and institutional requirements. Continued research on authenticity verification, bias mitigation, privacy protection, watermarking, and provenance can further complement advances in persistent generation. We view EonWeave primarily as a step toward more coherent and reliable long-horizon world modeling, and encourage its use in settings that preserve transparency, accountability, and respect for affected individuals.

Reproducibility Statement

We provide a comprehensive specification of the methodological, optimization, deployment, and evaluation procedures used throughout this work to facilitate faithful reproduction of EonWeave. Sec. 3 gives the complete formulations of its three core components: the training and inference design of ARiSE, the construction and capacity-controlled curation of the Persistent Scene Atlas in ReCAP, and the noise-adaptive routing and query-aligned reference distillation in RaDAR. Together, these formulations specify the full *trust–preserve–align* pipeline and the interactions among its components.

Sec. 4.1 and Appendix G.1 document the experimental protocol in detail, including the HM-World and Mira-World data configurations, training and test splits, single-shot and autoregressive evaluation settings, compared baselines, evaluation metrics, model backbone, principal hyperparameters, optimization settings, training schedule, and computational resources. All quantitative comparisons and ablation studies follow these stated configurations, providing a consistent experimental basis for reproducing the reported results.

Appendix B further provides explicit algorithmic descriptions of offline training-atlas construction, end-to-end optimization, and autoregressive inference, including the online update of the Persistent Scene Atlas during long-horizon generation. The few-step EonWeave-DMD configuration is separately documented in Appendix F, while Appendix G.5 reports sensitivity analyses for the principal design parameters. Collectively, the main paper and appendix specify the complete path from model formulation and memory construction to optimization, deployment, and evaluation, providing a detailed and internally consistent basis for reproducing the reported EonWeave results.

References

Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32. Curran Associates, Inc., 2019.

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, and Di Zhang. Recammaster: Camera-controlled generative rendering from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14834–14844, 2025.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28. Curran Associates, Inc., 2015.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 24081–24125, 2024.

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*, 2026a.

Zhifei Chen, Luozhou Wang, Guibao Shen, Dongyu Yan, Shuai Yang, Tianshuo Xu, Yihua Du, Wei Wang, Tianyi Gui, Lianghua Huang, et al. Reworld: An interactive world model with long-horizon memory. *arXiv preprint arXiv:2608.23565*, 2026b.

Nan Duan, Haoyang Huang, Weiyang Jin, Haoran Li, Yaowei Li, Yuming Li, Yijun Liu, Xin Lu, Xiaoxiao Ma, Yanwen Ma, et al. Long-horizon audio-visual generation for persistent stories and interactive worlds. *arXiv preprint arXiv:2608.23383*, 2026.

Samuel Garcin, Thomas Walker, Steven McDonagh, Tim Pearce, Hakan Bilen, Tianyu He, Kaixin Wang, and Jiang Bian. Beyond pixel histories: World models with persistent 3d state. *arXiv preprint arXiv:2603.03482*, 2026.

Anirudh Goyal, Alex M. Lamb, Ying Zhang, Saizheng Zhang, Aaron C. Courville, and Yoshua Bengio. Professor forcing: A new algorithm for training recurrent networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29. Curran Associates, Inc., 2016.

Yicong Hong, Yiqun Mei, Chongjian Ge, Yiran Xu, Yang Zhou, Sai Bi, Yannick Hold-Geoffroy, Mike Roberts, Matthew Fisher, Eli Shechtman, Kalyan Sunkavalli, Feng Liu, Zhengqi Li, and Hao Tan. RELIC: Interactive video world model with long-horizon memory. *arXiv preprint arXiv:2512.04040*, 2025.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 167283–167308, 2025.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *Advances in Neural Information Processing Systems*, 37:48955–48970, 2024.

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

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Zhiheng Liu, Xueqing Deng, Shoufa Chen, Angtian Wang, Qiushan Guo, Mingfei Han, Zeyue Xue, Mengzhao Chen, Ping Luo, and Linjie Yang. Worldweaver: Generating long-horizon video worlds via rich perception. *Advances in Neural Information Processing Systems*, 38:46268–46293, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 47500–47510. Curran Associates, Inc., 2023.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14297–14306. IEEE, 2023.

Zhan Peng, Jie Ma, Huiqiang Sun, Chong Gao, Zhijie Xue, Zhiyu Pan, Zhiguo Cao, Jun Liang, and Jing Li. Compression and retrieval: Implicit memory retrieval for video world models. *arXiv preprint arXiv:2606.23105*, 2026.

Ryan Po, Yotam Nitzan, Richard Zhang, Berlin Chen, Tri Dao, Eli Shechtman, Gordon Wetzstein, and Xun Huang. Long-context state-space video world models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8733–8744. IEEE, 2025.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2001–2010, July 2017.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations (ICLR)*, 2022.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4938–4947, June 2020.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations (ICLR)*, 2018.

Tianchang Shen, Sherwin Bahmani, Kai He, Sangeetha Grama Srinivasan, Tianshi Cao, Jiawei Ren, Ruilong Li, Zian Wang, Nicholas Sharp, Zan Gojcic, Sanja Fidler, Jiahui Huang, Huan Ling, Jun Gao, and Xuanchi Ren. Lyra 2.0: Explorable generative 3D worlds. *arXiv preprint arXiv:2604.13036*, 2026.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 267, pp. 56242–56280, 2025.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32211–32252. PMLR, 2023.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8922–8931, June 2021.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 1363–1389. Curran Associates, Inc., 2023. doi: 10.52202/075280-0068.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, April 2004. doi: 10.1109/TIP.2003.819861.

Tianhe Wu, Ruibin Li, Lei Zhang, and Kede Ma. Diversity-preserved distribution matching distillation for fast visual synthesis. *arXiv preprint arXiv:2602.03139*, 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.

Xindi Wu, Sven Elflein, James Lucas, Olga Russakovsky, Laura Leal-Taixé, Despoina Paschalidou, Jonathan Lorraine, and Aljoša Ošep. Addressable memory for video world models. *arXiv preprint arXiv:2608.07408*, 2026c.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. WorldMem: Long-term consistent world simulation with memory. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 49632–49652, 2025.

Tian-Xing Xu, Zi-Xuan Wang, Guangyuan Wang, Li Hu, Zhongyi Zhang, Peng Zhang, Bang Zhang, and Song-Hai Zhang. Ucm: Unified modeling of camera control and memory with time-aware positional encoding warping for world models. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, SIGGRAPH Conference Papers '26, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400725548. doi: 10.1145/3799902.3811088.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in Neural Information Processing Systems*, 37:47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623. IEEE, 2024b.

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers (SIGGRAPH Asia)*, pp. 1–11, 2025. doi: 10.1145/3757377.3763833.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 30546–30566, 2025.

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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, June 2018.

EonWeave: Weaving Persistent Evidence into Consistent Long-Horizon World Generation

Appendix

Anonymous submission

OVERVIEW

The submission consists of two files: `EonWeave.pdf`, containing the main paper and this appendix, and `Supplementary Material.zip`, containing three generation demonstration videos. The PDF provides detailed methodological analyses and experimental results, while the videos enable direct visual assessment of generation quality and temporal coherence.

```
[EonWeave Submission Materials]/
  |-- EonWeave.pdf
  |-- Supplementary Material.zip
```

This appendix follows the terminology and evaluation protocols established in the main paper. It provides the complete training and inference procedures of EonWeave, further analyzes its *trust–preserve–align* design, and presents additional evidence for long-horizon revisitation consistency. Specifically, it examines the three core components: Anchored Reliability Induction via Source Asymmetry (ARiSE), Reference Curation with Adaptive Projection (ReCAP), and Retrieval-aware Distillation into Aligned References (RaDAR).

ARiSE induces reliability-aware conditioning by selectively degrading short-term conditions while preserving clean historical evidence. ReCAP maintains a fixed-capacity Persistent Scene Atlas through pose–appearance diversity-based curation, while RaDAR performs noise-adaptive retrieval and distills the selected evidence into a fixed-size, query-aligned reference grid. During autoregressive inference, generated short-term conditions are re-noised before reuse, and the atlas is updated online using previously generated observations.

The appendix provides a further discussion of related work (Appendix A), detailed methodological and algorithmic analyses (Appendices B–E), an analysis of few-step distribution matching distillation (Appendix F), and extended experimental evaluation and analysis (Appendix G). The latter includes detailed experimental settings, long-horizon rollout analyses, component ablations, qualitative evaluations, and hyperparameter sensitivity studies.

Contents

A. Further Discussion of Related Work 16
B. Training Algorithm 17
C. Further Analysis of ARiSE 19
D. Further Analysis of ReCAP 26
E. Further Analysis of RaDAR 30
F. Analysis of Few-Step Distribution Matching Distillation 34
G. Extended Experimental Evaluation and Analysis 37

A Further Discussion of Related Work

Persistent spatial memory and explicit world state. Long-term spatial memory has emerged as a direct mechanism for preserving scene consistency under revisitation. Wu et al. (2026b) introduce a geometry-grounded long-term spatial memory that stores and retrieves previously observed scene information, providing explicit spatial persistence beyond the active temporal context. EonWeave pursues a different design point: ReCAP maintains a fixed-capacity atlas of pose–appearance-complementary latent observations, while RaDAR transforms retrieved history into query-aligned references, avoiding dependence on an explicitly reconstructed spatial memory representation. ReWorld (Chen et al., 2026b) similarly targets bounded long-horizon memory through mixed attention windows, a fixed-budget KV cache, and pose-indexed landmark retrieval, together with few-step distribution-matching distillation. In contrast, EonWeave couples memory capacity with content-aware curation and adaptive readout: ReCAP selects complementary historical states beyond pose proximity alone, while RaDAR performs noise-adaptive appearance–geometry routing at the patch level. PERSIST (Garcin et al., 2026) takes a more explicit geometric route by evolving a latent 3D state of the environment, camera, and renderer. EonWeave instead preserves historical visual evidence directly in latent space, enabling persistent revisitation within a pretrained video-generation paradigm without maintaining an explicit 3D scene state.

Memory addressability and drift-resistant historical evidence. WorldTrace (Wu et al., 2026c) identifies a complementary failure mode in long-horizon KV memory: temporal RoPE offsets can move beyond the distribution encountered during training, while compression in RoPE-rotated space can further impair memory addressability. It therefore assigns compressed memory slots in-distribution virtual positions and develops complementary memories for temporal coherence and episodic recall. RaDAR addresses a different but related problem: given available historical evidence, it determines *which* observations are relevant to the current query and *how* they should be converted into structured generative context through appearance–geometry routing and reference distillation. WorldWeaver (Liu et al., 2026) instead strengthens long-horizon conditioning by jointly modeling RGB and perceptual signals, exploiting comparatively stable depth cues to construct historical memory and using segmented noise scheduling to mitigate temporal drift. EonWeave is complementary in emphasis: ReCAP explicitly curates pose–appearance-diverse historical states under a fixed memory budget, while ARiSE calibrates reliance between clean persistent evidence and progressively less reliable short-term autoregressive conditions.

Long-context architectures and rollout-aware generation. A separate line of work extends temporal persistence through architectural or training-level mechanisms. Po et al. (2025) combine block-wise state-space scanning with dense local attention, propagating long-range information through an efficient recurrent state rather than explicit historical retrieval. EonWeave instead retains an addressable set of selected world states, allowing previously observed evidence to be reactivated according to the current generation query rather than being represented solely through an implicit recurrent context. JoyAI-Echo-1.5 (Duan et al., 2026) combines composable cross-shot memory for long-form video generation, geometry-aware metric camera conditioning for interactive world modeling, and rollout-aware training for stable few-step autoregressive generation. EonWeave focuses specifically on revisitation consistency: ARiSE addresses changing source reliability under recursive generation, ReCAP preserves complementary historical evidence, and RaDAR aligns that evidence with the current query.

Taken together, these works demonstrate complementary strategies for extending memory through explicit geometry, compressed KV state, perceptual conditioning, recurrent sequence modeling, or rollout-aware training. EonWeave differs in treating long-horizon revisitation as a coupled *trust–preserve–align* problem: ARiSE determines how persistent history should compete with deteriorating local context, ReCAP determines what historical evidence remains available under bounded capacity, and RaDAR determines how that evidence is retrieved and transformed into query-aligned generative context.

B TRAINING ALGORITHM

This section provides a complete algorithmic account of how EonWeave constructs, consumes, and updates historical context throughout training and autoregressive deployment. The pipeline consists of three coupled procedures. Algorithm 1 first constructs scene-wise training atlases from causally available ground-truth history. Algorithm 2 then optimizes the generator using the current short-term condition together with the corresponding persistent historical context. Finally, Algorithm 3 instantiates the same separation between recent and persistent evidence during autoregressive rollout, where the Persistent Scene Atlas evolves online as previously generated content becomes history. Together, these procedures specify how EonWeave carries historical information from offline preparation to end-to-end optimization and ultimately to long-horizon generation. The notation used throughout the three algorithms is summarized in Tab. 4.

For each training scene and temporal position, the offline procedure gathers all latent observations preceding the active condition segment and curates from them a fixed-capacity Persistent Scene Atlas. Historical video segments are first encoded into the VAE latent space and associated with their camera states and global latent-time indices. ReCAP then retains a compact subset according to the pose–appearance diversity criterion introduced in the main paper, favoring observations that complement rather than duplicate the states already represented in persistent memory. The retained records are organized according to their original temporal positions, while unused slots, when the available history is shorter than the prescribed capacity, are explicitly excluded from subsequent retrieval. Importantly, the active condition segment itself is not part of its training atlas: persistent memory is reserved for observations that precede the short-term context, thereby maintaining a clean temporal separation between recent evidence and longer-range scene history. The resulting atlases are precomputed from ground-truth observations and remain uncorrupted during model optimization.

Algorithm 2 describes how these persistent records are integrated into a training iteration. The sampled training instance determines a short-term condition, a future target segment, and the atlas assembled from earlier history. Stored global latent-frame indices are expressed relative to the active condition before being consumed by the generator, preserving how far each historical observation lies from the current prediction point rather than treating the atlas as an unordered collection of memories. The target segment follows the standard Flow Matching corruption process (Lipman et al., 2023), whereas ARiSE perturbs the short-term condition through a separate reliability path: with the prescribed probability, non-initial conditions are re-noised at a sampled strength, while the persistent atlas remains clean. The first condition segment is kept uncorrupted, matching the observed context available at the beginning of an autoregressive rollout. EonWeave then processes the corrupted target together with the short-term condition, text and camera conditioning, and the persistent historical records. ReCAP maps atlas observations into the representation space of the corresponding generator blocks, while RaDAR performs query-dependent routing over the available short- and long-term evidence. Optimization remains governed by the original Flow Matching objective and is evaluated only on the target portion of the latent sequence, so no auxiliary source-selection supervision is required. All trainable components are optimized with AdamW (Loshchilov & Hutter, 2019).

Algorithm 3 realizes the corresponding process when ground-truth history is no longer available. The online Persistent Scene Atlas is initialized empty, reflecting the fact that the initial observed segment already serves as the active short-term condition rather than being redundantly stored as long-term memory. At each rollout step, the model generates the next segment from the current condition and the atlas accumulated from earlier observations. Only after this prediction is completed does the condition that has just been consumed become historical and therefore eligible for insertion into the atlas. ReCAP updates the bounded memory using its pose–appearance curation rule, after which the atlas is temporally reordered for subsequent access and its stored latent-frame indices are converted into offsets relative to the next active condition. This update order ensures that the two context pathways remain semantically distinct throughout the rollout: the short-term branch represents the immediately preceding state, whereas the Persistent Scene Atlas contains observations lying further back in the trajectory.

The generated latent itself subsequently follows two deliberately asymmetric paths. An uncorrupted copy is preserved as the strongest available representation of the generated segment and is retained for future atlas curation once that segment leaves the active short-term window. In parallel, ARiSE

applies inference-time re-noising before the same generated segment is recursively reused as the next short-term condition. Training and inference therefore differ in the provenance of persistent history, which is ground-truth during optimization but progressively generated during deployment, while preserving the same underlying information structure: recent evidence may exhibit reduced reliability, persistent evidence is maintained in its cleanest available form, and both are aligned to the current prediction through the same memory interface. In this sense, Algorithms 1–3 provide the procedural realization of EonWeave's *trust–preserve–align* principle: ARiSE regulates the reliability of recursively reused context, ReCAP preserves complementary historical states under a bounded memory budget, and RaDAR aligns the resulting evidence with the queries that require it.

Table 4: Notation used in the EonWeave training and inference algorithms.

Notation	Description
Training Data and Condition Encoding	
$\mathcal{D}, \mathcal{S}_\ell, \ell, N$	Training set $\mathcal{D}$; scene sequence $\mathcal{S}_\ell = \{(V_{\ell,m}, y_{\ell,m}, \Pi_{\ell,m})\}_{m=0}^{N}$; scene index ℓ; number N of target segments following the initial segment. Each sequence contains $N+1$ segments in total.
m, n, k	Segment index m; condition/rollout index n; training-iteration index k. During training, $n \in \{0, \dots, N-1\}$ selects the condition segment, and segment $n+1$ is the target.
$i, j, j^\star$	Atlas-record indices i, j and selected replacement index $j^\star$ in atlas construction. In the inference algorithm, i instead indexes denoising steps, with $i \in \{0, \dots, S-1\}$.
T_{lat}	Number of latent frames in each video segment. Under the fixed-length segmentation used in the algorithms, condition segment n starts at global latent-time index nT_{lat}.
$V_{\ell,m}, C_n, y_{\ell,m}$	Ground-truth video segment $V_{\ell,m}$ at temporal position m; condition segment $C_n \equiv V_{\ell,n}$ during training; text prompt $y_{\ell,m}$ associated with $V_{\ell,m}$.
$\Pi_{\ell,m}, \pi_i$	Camera-pose sequence $\Pi_{\ell,m}$ associated with $V_{\ell,m}$; camera pose π_i associated with historical latent frame Z_i.
$E_{\text{vae}}, D_{\text{vae}}$, E_{text}	Frozen variational autoencoder (VAE) encoder, frozen VAE decoder, and frozen text encoder, respectively.
$E^{\text{cond}}_{\text{cam},\theta}$, $E^{\text{tgt}}_{\text{cam},\theta}$	Trainable block-wise camera encoders for the condition and target camera trajectories, respectively. During training, they map $\Pi_{\ell,n}$ and $\Pi_{\ell,n+1}$ to c_n^{cond} and c_{n+1}^{tgt}. During inference, the scene index ℓ is omitted.
$z_n^{\text{cond}}, z_{n+1}^{\text{tgt}}$	VAE latent representations of the condition and target segments. For a sampled training scene ℓ, $z_n^{\text{cond}} = z_{\ell,n}$ and $z_{n+1}^{\text{tgt}} = z_{\ell,n+1}$.
h_{n+1}^{text}	Text-conditioning embedding extracted from the target-segment prompt $y_{\ell,n+1}$ during training, or y_{n+1} during inference.
$c_n^{\text{cond}}, c_{n+1}^{\text{tgt}}$	Block-wise camera representations of the condition and target camera trajectories, respectively.
Persistent Scene Atlas Construction and Temporal Indexing	
$z_{\ell,m}, \mathcal{Z}_{\ell,<n}$	Offline VAE latent $z_{\ell,m}$ of segment $V_{\ell,m}$; temporally ordered latent history $\mathcal{Z}_{\ell,<n} = (z_{\ell,m})_{m=0}^{n-1}$ preceding condition segment n.
$\text{Align}_{\text{vae}}$	Deterministic alignment operator that associates historical VAE latent frames with their camera poses and global indices on the latent temporal grid.
$\mathcal{H}_{\ell,n}^{\text{GT}}$	Temporally ordered ground-truth history of aligned triples (Z_i, π_i, u_i) preceding condition segment n in scene ℓ, with $\mathcal{H}_{\ell,0}^{\text{GT}} = \emptyset$.
$u_{\ell,n}^{\text{cond}}, u_n^{\text{cond}}$, u_i	Global latent-time origin of condition segment $V_{\ell,n}$ during training; corresponding condition origin during inference; global latent-frame index of historical Atlas Record a_i, respectively. The condition origins satisfy $u_{\ell,n}^{\text{cond}} = u_n^{\text{cond}} = nT_{\text{lat}}$ under the fixed-length segmentation used in the algorithms.
Δ_i	Temporal offset of historical record a_i relative to the current condition origin: $\Delta_i = u_i - u_{\ell,n}^{\text{cond}} < 0$ during training and $\Delta_i = u_i - u_n^{\text{cond}} < 0$ during inference. This offset provides the temporal coordinate used when EonWeave reads the Atlas Record.
$a_i, Z_i, \mathcal{A}_{\ell,n}^{\text{clean}}$	Persistent Atlas Record $a_i = (Z_i, \pi_i, u_i)$; stored historical latent frame Z_i; clean atlas precomputed from ground-truth history preceding condition segment n in scene ℓ, respectively. The initial training atlas satisfies $\mathcal{A}_{\ell,0}^{\text{clean}} = \emptyset$.

Continued on next page

Table 4 (continued)

Notation	Description
K, α, s	Maximum number of Atlas Records $K \geq 2$; pose–appearance distance weight $\alpha \in [0, 1]$; translation normalization scale $s > 0$, respectively.
$\mathcal{A}^{\mathrm{rel}}_{\ell,n}$, $\mathcal{A}^{\mathrm{rel}}_{n}$	Condition-relative views of the offline training atlas and online inference atlas, respectively, obtained by replacing each absolute timestamp u_i with its current relative offset Δ_i. These views are constructed at read time and are not stored as independent atlases.
d, D	Pose–appearance distance $d(a_i, a_j)$, parameterized by α and s, as defined in Eq. 6; nearest-neighbor atlas dispersion $D(\mathcal{A})$, as defined in Eq. 7.
ARiSE and Flow Matching Optimization	
$p_{\mathrm{aug}}, \sigma_{\max}$	ARiSE augmentation probability $p_{\mathrm{aug}} \in [0, 1]$ for non-initial conditions; maximum training noise level $\sigma_{\max} \in [0, 1]$.
$b_n, \sigma_c, \bar{z}^{\mathrm{cond}}_n$	Bernoulli augmentation gate $b_n \sim \mathrm{Bernoulli}(p_{\mathrm{aug}})$; sampled condition-noise level $\sigma_c \sim \mathcal{U}(0, \sigma_{\max})$; condition latent supplied during training, respectively. For $n > 0$, ARiSE constructs $\bar{z}^{\mathrm{cond}}_n = (1 - b_n\sigma_c)z^{\mathrm{cond}}_n + b_n\sigma_c\epsilon_c$. The initial condition remains unchanged: $\bar{z}^{\mathrm{cond}}_0 = z^{\mathrm{cond}}_0$.
$I, \epsilon_c, \epsilon_x$, ϵ_{n+1}	Identity matrix; standard Gaussian noise used for training-time condition perturbation; standard Gaussian noise used for the target Flow Matching path or inference initialization; standard Gaussian noise used to re-noise generated condition segment $n + 1$, respectively. Each noise variable has distribution $\mathcal{N}(0, I)$ with dimensions matching its associated latent.
$t, x_t, v^\star, w(t)$	Flow Matching interpolation time $t \in [0, 1]$; interpolated target state $x_t = (1 - t)z^{\mathrm{tgt}}_{n+1} + t\epsilon_x$; target velocity $v^\star = \epsilon_x - z^{\mathrm{tgt}}_{n+1}$; timestep weighting function from the base Flow Matching objective, respectively.
$\mathcal{L}_{\mathrm{ARiSE}}$	Flow Matching training objective with ARiSE-perturbed short-term conditioning and a clean historical atlas. The timestep weighting $w(t)$ is retained from the base objective.
$v_\theta, \hat{v}, \theta$	EonWeave velocity-field generator; predicted velocity; complete trainable parameter set including the generator, camera encoders, ReCAP projections, and RaDAR parameters, respectively.
T_{train}, η	Number of training iterations and AdamW learning rate, respectively.
Autoregressive Inference	
$V_n, y_n, \Pi_n, V_{0:N}$	Video segment at rollout position n, where V_0 is the observed initial segment and V_n for $n \geq 1$ is generated; text prompt y_n for generated segment n; camera-pose sequence Π_n for segment n; complete video sequence $V_{0:N} = V_0 \parallel V_1 \parallel \cdots \parallel V_N$, respectively. Here, $\parallel$ denotes temporal concatenation, and $C_n \equiv V_n$ during inference.
$\mathcal{A}_n, \mathcal{W}$	Online Persistent Scene Atlas available before generating V_{n+1}; capacity-controlled ReCAP write operator, respectively. The atlas contains only history preceding C_n and satisfies $\mathcal{A}_0 = \emptyset$ and $\lvert\mathcal{A}_n\rvert \leq K$. After V_n has served as the active condition, the update is $\mathcal{A}_{n+1} = \mathcal{W}(\mathcal{A}_n, V_n; K, \alpha, s)$.
$S, \{t_i\}_{i=0}^{S}, x_{t_i}$	Number of Flow Matching integration steps during inference; decreasing timestep schedule $1 = t_0 > \cdots > t_S = 0$; intermediate target latent at timestep t_i, respectively.
$\hat{z}_n, \tilde{z}^{\mathrm{cond}}_n$, $\sigma^\star$	Generated latent of segment V_n before ARiSE re-noising, for $n \geq 1$; condition latent supplied when generating V_{n+1}; fixed inference-time re-noising strength, respectively. For $n \geq 1$, the reused condition is $\tilde{z}^{\mathrm{cond}}_n = (1 - \sigma^\star)\hat{z}_n + \sigma^\star\epsilon_n$. The initial observed condition remains unchanged: $\tilde{z}^{\mathrm{cond}}_0 = z^{\mathrm{cond}}_0$. The strength satisfies $0 \leq \sigma^\star \leq \sigma_{\max}$.

C Further Analysis of ARiSE

This section provides additional analysis of *Anchored Reliability Induction via Source Asymmetry (ARiSE)*. We further examine why the combination of *condition re-noising*, a *persistent reliability anchor*, and *inference-time reliability alignment* is well matched to long-horizon autoregressive generation. Rather than introducing an additional mechanism, the following analysis develops the reliability interpretation already implicit in the training and inference procedures of ARiSE.

Algorithm 1 Offline Construction of Scene-wise Training Atlases

Input: Training scenes $\mathcal{D}$
Output: Fixed clean training atlases $\{\mathcal{A}^{\text{clean}}_{\ell,n}\}$
Hyperparameters: Atlas capacity $K \geq 2$, distance weight $\alpha \in [0,1]$, and translation scale $s > 0$

for all scene ℓ with sequence $\{(V_{\ell,m}, \Pi_{\ell,m})\}_{m=0}^{N} \in \mathcal{D}$ **do**
  Encode all segments as $z_{\ell,m} \leftarrow E_{\text{vae}}(V_{\ell,m})$, $m = 0, \ldots, N$
  Initialize $\mathcal{A}^{\text{clean}}_{\ell,0} \leftarrow \emptyset$
  for $n = 1, \ldots, N-1$ **do**
    Construct the historical latent sequence $\mathcal{Z}_{\ell,<n} \leftarrow (z_{\ell,m})_{m=0}^{n-1}$
    Align historical latents with camera poses and global latent-time coordinates: $\mathcal{H}^{\text{GT}}_{\ell,n} \leftarrow \text{Align}_{\text{vae}}(\mathcal{Z}_{\ell,<n}, \Pi_{\ell,0:n-1})$
    Initialize $\mathcal{A}^{\text{clean}}_{\ell,n} \leftarrow \emptyset$
    for all $(Z_i, \pi_i, u_i) \in \mathcal{H}^{\text{GT}}_{\ell,n}$ in chronological order **do**
      Construct Atlas Record $a_i \leftarrow (Z_i, \pi_i, u_i)$
      if $|\mathcal{A}^{\text{clean}}_{\ell,n}| < K$ **then**
        Insert a_i into $\mathcal{A}^{\text{clean}}_{\ell,n}$
      else
        $j^\star \leftarrow \arg\max_{j:a_j \in \mathcal{A}^{\text{clean}}_{\ell,n}} D\Big((\mathcal{A}^{\text{clean}}_{\ell,n} \setminus \{a_j\}) \cup \{a_i\}\Big)$
        if $D\Big((\mathcal{A}^{\text{clean}}_{\ell,n} \setminus \{a_{j^\star}\}) \cup \{a_i\}\Big) > D(\mathcal{A}^{\text{clean}}_{\ell,n})$ **then**
          Replace $a_{j^\star}$ with a_i
        end if
      end if
    end for
  end for
end for
return $\{\mathcal{A}^{\text{clean}}_{\ell,n}\}$

C.1 From Error Simulation to Reliability Control

At generation step n, EonWeave conditions on two qualitatively different information sources. The short-term condition C_n provides temporally adjacent evidence and is therefore highly informative for local dynamics, whereas the Persistent Scene Atlas $\mathcal{A}_n$ preserves observations from earlier world states. During conventional training, both sources originate from ground-truth video. The adjacent condition z_n^{cond} is sufficiently informative that the generator can often minimize the objective without substantially exploiting persistent history. During autoregressive inference, however, C_n is itself generated and progressively inherits errors from preceding segments, while earlier Atlas Records can remain comparatively stable. The resulting mismatch is therefore not merely a shift in the condition distribution, but a change in the *relative reliability* of the two conditioning sources. This train–inference discrepancy is closely related to exposure bias in autoregressive sequence modeling, where optimization uses clean histories but deployment conditions on model-generated contexts (Bengio et al., 2015; Goyal et al., 2016).

ARiSE addresses this mismatch without attempting to reproduce the exact morphology of future rollout errors. Instead, it explicitly controls the reliability of the short-term source through the corruption family

$$\mathcal{R}_\sigma(z) = (1-\sigma)z + \sigma\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{17}$$

where I denotes the identity covariance. For a non-initial condition, an active augmentation event uses $\sigma = \sigma_c \sim \mathcal{U}(0, \sigma_{\max})$; otherwise the condition remains unchanged, and the initial observed condition is always kept uncorrupted. The Persistent Scene Atlas remains uncorrupted throughout training. ARiSE therefore exposes the generator to a continuum ranging from highly reliable local evidence to progressively degraded short-term evidence, while preserving a comparatively stable historical source. Its role is not to imitate a particular rollout artifact, but to create a controlled reliability deficit in the local pathway so that persistent evidence becomes useful under the original generation objective.

Algorithm 2 EonWeave Training Procedure

Input: Training scenes $\mathcal{D}$ and fixed atlases $\{\mathcal{A}^{\text{clean}}_{\ell,n}\}$ from Algorithm 1
Output: Optimized parameters θ
Hyperparameters: Training iterations T_{train}, learning rate η, augmentation probability $p_{\text{aug}} \in [0,1]$, and maximum noise level $\sigma_{\max} \in [0,1]$

for $k = 1, \dots, T_{\text{train}}$ **do**
  Sample $(\ell, \mathcal{S}_\ell) \leftarrow \text{Sample}(\mathcal{D})$ and $n \sim \text{Unif}(\{0, \dots, N-1\})$
  Set $C_n \leftarrow V_{\ell,n}$ and retrieve $\mathcal{A}^{\text{clean}}_{\ell,n}$
  Encode the condition and target segments: $z^{\text{cond}}_n \leftarrow E_{\text{vae}}(C_n) = z_{\ell,n}$, $z^{\text{tgt}}_{n+1} \leftarrow E_{\text{vae}}(V_{\ell,n+1}) = z_{\ell,n+1}$
  Extract the text condition $h^{\text{text}}_{n+1} \leftarrow E_{\text{text}}(y_{\ell,n+1})$
  Encode the camera conditions: $(c^{\text{cond}}_n, c^{\text{tgt}}_{n+1}) \leftarrow (E^{\text{cond}}_{\text{cam},\theta}(\Pi_{\ell,n}), E^{\text{tgt}}_{\text{cam},\theta}(\Pi_{\ell,n+1}))$
  Determine the global latent-time origin of the condition segment: $u^{\text{cond}}_{\ell,n} \leftarrow n \cdot T_{\text{lat}}$
  Convert absolute atlas timestamps to relative temporal offsets: $\mathcal{A}^{\text{rel}}_{\ell,n} \leftarrow \{(Z_i, \pi_i, \Delta_i) \mid (Z_i, \pi_i, u_i) \in \mathcal{A}^{\text{clean}}_{\ell,n},\ \Delta_i = u_i - u^{\text{cond}}_{\ell,n}\}$
  if $n = 0$ **then**
    $\bar{z}^{\text{cond}}_n \leftarrow z^{\text{cond}}_n$
  else
    Sample $b_n \sim \text{Bernoulli}(p_{\text{aug}})$, $\sigma_c \sim \mathcal{U}(0, \sigma_{\max})$, and $\epsilon_c \sim \mathcal{N}(0, I)$
    $\bar{z}^{\text{cond}}_n \leftarrow (1 - b_n\sigma_c) z^{\text{cond}}_n + b_n \sigma_c \epsilon_c$
  end if
  Sample $t \sim \mathcal{U}(0,1)$ and $\epsilon_x \sim \mathcal{N}(0, I)$
  Construct the Flow Matching state $x_t \leftarrow (1-t) z^{\text{tgt}}_{n+1} + t\epsilon_x$ and $v^\star \leftarrow \epsilon_x - z^{\text{tgt}}_{n+1}$
  Predict $\hat{v} \leftarrow v_\theta(x_t, t, \bar{z}^{\text{cond}}_n, h^{\text{text}}_{n+1}, c^{\text{cond}}_n, c^{\text{tgt}}_{n+1}, \mathcal{A}^{\text{rel}}_{\ell,n})$
  Compute $\mathcal{L}_{\text{ARiSE}} \leftarrow w(t) \|v^\star - \hat{v}\|_2^2$ and update θ using AdamW at learning rate η
end for
return θ

C.2 Quantifying the Induced Reliability Spectrum

The corruption family in Eq. 17 can be interpreted through the linear dependence between the perturbed condition and its clean counterpart. Consider the normalized analytical setting

$$\mathbb{E}[z] = \mathbb{E}[\epsilon] = 0, \qquad \text{Cov}(z) = \text{Cov}(\epsilon) = I, \qquad z \perp \epsilon. \tag{18}$$

Conditioned on augmentation, let

$$\bar{z}_\sigma = (1-\sigma) z + \sigma \epsilon. \tag{19}$$

Its cross-covariance with the original condition and its marginal covariance are

$$\text{Cov}(\bar{z}_\sigma, z) = (1-\sigma) I, \qquad \text{Cov}(\bar{z}_\sigma) = \left[(1-\sigma)^2 + \sigma^2\right] I. \tag{20}$$

This yields the normalized linear-correlation proxy

$$R(\sigma) = \frac{1-\sigma}{\sqrt{(1-\sigma)^2 + \sigma^2}}, \tag{21}$$

whose derivative is

$$\frac{\mathrm{d}R}{\mathrm{d}\sigma} = -\frac{\sigma}{\left[(1-\sigma)^2 + \sigma^2\right]^{3/2}} \le 0. \tag{22}$$

Within this normalized model, increasing σ therefore monotonically reduces the linear dependence retained from the original local observation. Sampling σ_c continuously from $[0, \sigma_{\max}]$ consequently defines an ordered spectrum of short-term reliability rather than a collection of unrelated perturbation strengths.

The same behavior can be characterized through corruption energy. Conditioned on augmentation,

$$\bar{z}_\sigma - z = \sigma(\epsilon - z), \tag{23}$$

Algorithm 3 EonWeave Inference Procedure

Input: Initial segment V_0, text conditions $\{y_n\}_{n=1}^{N}$, camera trajectories $\{\Pi_n\}_{n=0}^{N}$, and trained parameters θ
Output: Generated video sequence $V_{0:N}$
Hyperparameters: Flow steps S, atlas capacity K, distance weight $\alpha \in [0,1]$, translation scale $s > 0$, and re-noising strength $0 \le \sigma^\star \le \sigma_{\max}$

Initialize $\mathcal{A}_0 \leftarrow \emptyset$, $C_0 \leftarrow V_0$, $z_0^{\mathrm{cond}} \leftarrow E_{\mathrm{vae}}(C_0)$, and $\widetilde{z}_0^{\mathrm{cond}} \leftarrow z_0^{\mathrm{cond}}$
Set the initial latent-time origin $u_0^{\mathrm{cond}} \leftarrow 0$
for $n = 0, \dots, N-1$ **do**
  Extract the text condition $h_{n+1}^{\mathrm{text}} \leftarrow E_{\mathrm{text}}(y_{n+1})$
  Encode the camera conditions: $(c_n^{\mathrm{cond}}, c_{n+1}^{\mathrm{tgt}}) \leftarrow (E_{\mathrm{cam},\theta}^{\mathrm{cond}}(\Pi_n), E_{\mathrm{cam},\theta}^{\mathrm{tgt}}(\Pi_{n+1}))$
  Convert absolute atlas timestamps to offsets relative to the current condition: $\mathcal{A}_n^{\mathrm{rel}} \leftarrow \{(Z_i, \pi_i, \Delta_i) \mid (Z_i, \pi_i, u_i) \in \mathcal{A}_n,\ \Delta_i = u_i - u_n^{\mathrm{cond}}\}$
  Set $1 = t_0 > \cdots > t_S = 0$, sample $\epsilon_x \sim \mathcal{N}(0, I)$, and initialize $x_{t_0} \leftarrow \epsilon_x$
  for $i = 0, \dots, S-1$ **do**
    $\hat{v} \leftarrow v_\theta(x_{t_i}, t_i, \widetilde{z}_n^{\mathrm{cond}}, h_{n+1}^{\mathrm{text}}, c_n^{\mathrm{cond}}, c_{n+1}^{\mathrm{tgt}}, \mathcal{A}_n^{\mathrm{rel}})$
    $x_{t_{i+1}} \leftarrow x_{t_i} + (t_{i+1} - t_i)\hat{v}$
  end for
  Decode the generated segment: $\hat{z}_{n+1} \leftarrow x_{t_S}$, $V_{n+1} \leftarrow D_{\mathrm{vae}}(\hat{z}_{n+1})$
  Update the Persistent Scene Atlas: $\mathcal{A}_{n+1} \leftarrow \mathcal{W}(\mathcal{A}_n, V_n; K, \alpha, s)$, with $|\mathcal{A}_{n+1}| \le K$
  Advance the autoregressive condition: $C_{n+1} \leftarrow V_{n+1}$, $z_{n+1}^{\mathrm{cond}} \leftarrow \hat{z}_{n+1}$
  Update the latent-time origin: $u_{n+1}^{\mathrm{cond}} \leftarrow (n+1) \cdot T_{\mathrm{lat}}$
  if $n + 1 < N$ **then**
    Sample $\epsilon_{n+1} \sim \mathcal{N}(0, I)$
    $\widetilde{z}_{n+1}^{\mathrm{cond}} \leftarrow (1 - \sigma^\star) z_{n+1}^{\mathrm{cond}} + \sigma^\star \epsilon_{n+1}$
  end if
end for
return $V_{0:N} = V_0 \| V_1 \| \cdots \| V_N$

and for a d-dimensional standardized latent,

$$\mathbb{E}\left[\|\bar{z}_\sigma - z\|_2^2\right] = 2d\sigma^2. \tag{24}$$

For a non-initial condition, the actual gated perturbation is

$$\bar{z} = (1 - b\sigma)z + b\sigma\epsilon, \qquad b \sim \mathrm{Bernoulli}(p_{\mathrm{aug}}), \qquad \sigma \sim \mathcal{U}(0, \sigma_{\max}). \tag{25}$$

Using $b^2 = b$ and $\mathbb{E}[\sigma^2] = \sigma_{\max}^2/3$ gives

$$\begin{aligned} \mathbb{E}\left[\|\bar{z} - z\|_2^2\right] &= 2d\,\mathbb{E}[b]\mathbb{E}[\sigma^2] \\ &= \frac{2d}{3} p_{\mathrm{aug}} \sigma_{\max}^2. \end{aligned} \tag{26}$$

This decomposition separates the roles of the two ARiSE hyperparameters: p_{aug} controls how frequently optimization leaves the clean-condition regime, whereas $\sigma_{\max}$ controls how far the reliability spectrum extends once perturbation is activated. These calculations rely on a simplified isotropic model rather than the actual statistics of VAE latents, but they make explicit a property already present in the training rule: the local source is varied along an ordered reliability axis while the persistent source remains unchanged.

C.3 Why Asymmetry Induces Memory Utilization

The asymmetric construction changes the marginal utility of the atlas pathway. Under clean-condition training, the temporally adjacent z_n^{cond} already explains much of the target segment, so persistent history provides relatively limited additional information. As σ_c increases, part of this local information becomes uncertain, while complementary persistent evidence remains accessible through the clean atlas. Historical context can therefore reduce the original Flow Matching objective

more effectively precisely when the short-term source becomes less reliable, creating an optimization incentive for memory utilization without an explicit "use-memory" loss.

A simple two-source estimation model illustrates this reliability shift. Let a persistent world-state variable h be observed through a short-term source c and a memory source m,

$$c = h + \eta_c, \qquad m = h + \eta_m, \qquad \eta_c \sim \mathcal{N}(0, s_c^2 I), \qquad \eta_m \sim \mathcal{N}(0, s_m^2 I), \qquad \eta_c \perp \eta_m. \tag{27}$$

Under an uninformative prior on h, equivalently considering minimum-variance unbiased fusion of the two observations, the precision-weighted estimate is

$$\hat{h} = \frac{s_c^{-2} c + s_m^{-2} m}{s_c^{-2} + s_m^{-2}}. \tag{28}$$

If w_c and w_m denote the corresponding normalized coefficients, then

$$\frac{w_m}{w_c} = \frac{s_c^2}{s_m^2}. \tag{29}$$

Increasing uncertainty in the local observation therefore raises the relative contribution of the more stable memory source. The actual generator is nonlinear and does not satisfy this Gaussian model, so σ_c should not be identified directly with s_c. Nevertheless, ARiSE induces the same qualitative reliability shift: the short-term source is selectively weakened while the atlas remains intact.

The asymmetry is important because it creates training examples in which the short-term condition is degraded while relevant historical evidence remains available. Corrupting both sources would also weaken the historical evidence that the model is intended to use. By keeping the training atlas clean, ARiSE encourages the model to rely on persistent memory when recent context becomes unreliable. Continuous sampling of σ_c further avoids reducing this interaction to a binary choice between "local" and "memory"; the resulting behavior is better understood as soft source arbitration over a range of reliability levels.

The same principle also explains the necessity of constructing the training atlas from ground-truth history. As indicated by Eq. 29, greater uncertainty in historical memory reduces its contribution relative to the short-term condition, thereby weakening the learning signal for persistent-memory utilization. By preserving a clean ground-truth atlas rather than using error-contaminated generated history during training, ARiSE maintains a strong reliability contrast, enabling the model to develop a strong capability to utilize historical evidence. Consistent with the importance of this mechanism, the component ablation results in Tab. 7 show that removing ARiSE while retaining the remaining pipeline reduces PSNR from 23.571 to 20.735 dB and SSIM from 0.703 to 0.518 on Mira-World.

C.4 Why Re-noise Generated Conditions at Inference?

Training-time reliability induction alone does not remove the deployment mismatch. A generated segment does not generally contain additive Gaussian noise; its errors may instead appear as appearance drift, color bias, texture distortion, or inaccurate geometry. Such structured artifacts can resemble legitimate content and may therefore be recursively propagated when the generated segment is reused directly as the next condition. To expose the effect of inference-time re-noising, write

$$\hat{z}_n = z_n + \delta_n, \tag{30}$$

where z_n denotes an ideal latent and δ_n summarizes the discrepancy accumulated during autoregressive rollout. For a fixed inference strength $\sigma^\star \in [0, \sigma_{\max}]$, ARiSE applies

$$\begin{aligned} \tilde{z}_n^{\text{cond}} &= (1-\sigma^\star)\hat{z}_n + \sigma^\star \epsilon_n \\ &= (1-\sigma^\star) z_n + \underbrace{(1-\sigma^\star)\delta_n + \sigma^\star \epsilon_n}_{\text{effective condition uncertainty}}, \qquad \epsilon_n \sim \mathcal{N}(0, I). \end{aligned} \tag{31}$$

The deterministic contribution of the inherited discrepancy is therefore attenuated by $(1-\sigma^\star)$ before the condition is reused. Re-noising does not need to identify or explicitly remove δ_n; it instead prevents structured rollout artifacts from being propagated with the same authority as an unperturbed local condition.

At the same time, $\widetilde{z}_n^{\mathrm{cond}}$ belongs to the same corruption family used during training. Since $\sigma^\star \leq \sigma_{\max}$, inference maps generated conditions back into a reliability regime explicitly encountered during optimization. We therefore interpret this operation as *reliability canonicalization*: rather than requiring the model to recognize the precise morphology of arbitrary autoregressive artifacts, it transforms the generated condition into a familiar uncertain state in which persistent evidence has already been trained to become useful. Importantly, only the copy used for immediate short-term reuse is re-noised. The original generated latent remains available and becomes eligible for ReCAP curation after the corresponding segment completes its role as the active condition:

$$\hat{z}_n \longrightarrow \begin{cases} \widetilde{z}_n^{\mathrm{cond}}, & \text{short-term reuse,} \\ \text{future ReCAP eligibility,} & \text{persistent preservation.} \end{cases} \tag{32}$$

The two branches therefore preserve the intended source asymmetry at inference as well: recursive local context is reliability-controlled, whereas persistent preservation uses the generated representation without re-noising.

C.5 A Local View of Rollout-Error Contraction

The decomposition in Eq. 31 also provides a local view of inherited error propagation. Consider an actual generated latent $\hat{z}_n = z_n + \delta_n$ and its ideal counterpart z_n. If both are re-noised using the same Gaussian realization ϵ_n, define

$$\widetilde{z}_{n,\mathrm{ideal}}^{\mathrm{cond}} = (1-\sigma^\star) z_n + \sigma^\star \epsilon_n. \tag{33}$$

Their discrepancy after ARiSE is exactly

$$\widetilde{z}_n^{\mathrm{cond}} - \widetilde{z}_{n,\mathrm{ideal}}^{\mathrm{cond}} = (1-\sigma^\star)\delta_n. \tag{34}$$

Under this coupled-noise comparison, the deterministic discrepancy inherited from the previous rollout is contracted by the explicit factor $1-\sigma^\star$ before it enters the next generation step.

To examine how this discrepancy may propagate across segments, hold the atlas, text condition, camera condition, and sampling randomness fixed, and let F_θ denote the effective local mapping induced by the generation procedure from the short-term condition representation to the next generated latent. Suppose that in a neighborhood of the rollout trajectory,

$$\|F_\theta(c_1, \mathcal{A}) - F_\theta(c_2, \mathcal{A})\|_2 \leq L_c \|c_1 - c_2\|_2 , \tag{35}$$

where $L_c \geq 0$ denotes a local condition-sensitivity constant. Define

$$\delta_{n+1}^{\mathrm{prop}} = F_\theta\big(\widetilde{z}_n^{\mathrm{cond}}, \mathcal{A}\big) - F_\theta\big(\widetilde{z}_{n,\mathrm{ideal}}^{\mathrm{cond}}, \mathcal{A}\big) . \tag{36}$$

Combining Eqs. 34 and 35 yields

$$\big\|\delta_{n+1}^{\mathrm{prop}}\big\|_2 \leq L_c(1-\sigma^\star) \|\delta_n\|_2 . \tag{37}$$

Let $e_n = \|\delta_n\|_2$ denote the total rollout discrepancy magnitude, and write

$$\delta_{n+1} = \delta_{n+1}^{\mathrm{prop}} + r_{n+1}, \qquad \xi_{n+1} = \|r_{n+1}\|_2, \tag{38}$$

where r_{n+1} collects newly introduced generation error and other effects not captured by the isolated short-term pathway. The triangle inequality gives

$$e_{n+1} \leq q e_n + \xi_{n+1}, \qquad q = L_c(1-\sigma^\star), \tag{39}$$

and unrolling the recurrence yields

$$e_N \leq q^N e_0 + \sum_{j=1}^{N} q^{N-j} \xi_j. \tag{40}$$

Within this local model, decreasing q attenuates the contribution of discrepancies originating further back in the rollout, while $q < 1$ corresponds to a contractive regime for the component propagated through the short-term condition. This is not a global stability guarantee for the nonlinear generator; it characterizes only the isolated local pathway under the assumptions above. It also exposes the limitation of aggressive re-noising: reducing $1-\sigma^\star$ suppresses inherited discrepancy, but simultaneously attenuates useful recent information that persistent memory may not contain.

C.6 The Persistent Atlas as a Reliability Anchor

The benefit of condition re-noising is fundamentally coupled to persistent memory. Re-noising alone removes information, while memory alone does not prevent the generator from over-relying on a degraded short-term condition. ARiSE combines the two by deliberately reducing the authority of the local source while leaving a complementary historical pathway available. ARiSE does not rely on persistent history being error-free. Although part of the historical evidence at inference comes from previously generated results and may therefore contain some deviation, as long as the atlas retains information complementary to the current short-term condition, re-noising the latter to reduce its excessive dominance can strengthen the contribution of historical evidence, enabling the model to more effectively exploit recent information and persistent history in a complementary manner. This division is particularly relevant to revisitation: immediate motion and newly observed content should remain governed primarily by the recent condition, whereas identities, appearance, spatial configuration, and scene states established much earlier may need to be recovered from persistent history after long temporal gaps.

The same asymmetry also facilitates optimization of the ReCAP pathway. Without ARiSE, a highly informative adjacent condition can explain much of the target before persistent-memory projections receive substantial learning pressure. Selectively weakening z_n^{cond} increases the marginal utility of historical evidence under the original Flow Matching loss, encouraging the generator to exploit the atlas without introducing auxiliary source-selection supervision or changing the single-segment training objective. ARiSE therefore does not replace local context with memory; it establishes the reliability conditions under which the two sources can assume complementary roles.

C.7 Choosing the Inference Reliability Level

The inference strength $\sigma^\star$ controls a natural trade-off. With $\sigma^\star = 0$, the generated condition is reused unchanged, preserving the maximum amount of recent information but also leaving inherited rollout artifacts unaffected by ARiSE. Increasing $\sigma^\star$ suppresses this propagated discrepancy and moves the local condition further into the reliability regime in which persistent evidence has been trained to become useful. If $\sigma^\star$ becomes too large, however, useful short-term information such as transient motion or newly observed content absent from the atlas is also weakened. We summarize this behavior qualitatively as

$$\mathcal{E}(\sigma^\star) \approx \underbrace{\mathcal{E}_{\mathrm{prop}}(\sigma^\star)}_{\text{propagated condition error}} + \underbrace{\mathcal{E}_{\mathrm{loss}}(\sigma^\star)}_{\text{suppressed local information}}, \tag{41}$$

where $\mathcal{E}$ denotes a qualitative total rollout-error proxy, $\mathcal{E}_{\mathrm{prop}}$ captures discrepancy propagated through the short-term condition, and $\mathcal{E}_{\mathrm{loss}}$ represents useful local information weakened by re-noising. The first term is expected to decrease with stronger re-noising, whereas the second tends to increase.

The Persistent Scene Atlas mitigates the second cost only for information that has already been observed and retained. Persistent identities or scene states can remain accessible even after their local representation is weakened, whereas genuinely new motion or appearance information cannot in general be reconstructed from memory. This interaction motivates a moderate $\sigma^\star$ within the reliability range encountered during training rather than aggressive corruption. More broadly, it captures the central principle of ARiSE: instead of requiring the model to infer the reliability of arbitrary autoregressive artifacts from scratch, expose it to a controlled reliability spectrum during training and map generated conditions back into that spectrum during inference.

In summary, ARiSE couples three mutually reinforcing properties: controlled degradation creates an information deficit that increases the utility of persistent memory; source asymmetry preserves a stable historical anchor from which persistent world-state information can be recovered; and inference-time re-noising maps recursively generated conditions back into the reliability regime encountered during training. The analyses above further show that the training perturbation induces an ordered reliability spectrum and that inference-time re-noising locally attenuates discrepancies inherited through the short-term pathway. Together, these mechanisms provide a source-arbitration principle for long-horizon generation without auxiliary reliability labels, explicit source-selection supervision, or multi-segment training unrolling.

D FURTHER ANALYSIS OF RECAP

This section provides further analysis of *Reference Curation with Adaptive Projection (ReCAP)*. We examine why the combination of *pose–appearance diversity*, *positive-gain replacement*, and *block-adaptive projection* provides an effective interface for persistent scene memory under a fixed capacity. Rather than introducing an additional memory mechanism, the following analysis develops the coverage interpretation already implicit in ReCAP: curation determines which historically established states remain represented under the memory budget, while adaptive projection determines how those retained states are exposed to different transformer blocks of the generator.

D.1 FROM HISTORY ACCUMULATION TO COVERAGE PRESERVATION

A long autoregressive rollout continually produces new observations, whereas persistent memory must remain bounded. The central problem is therefore not simply to compress an increasingly long sequence, but to allocate a fixed number of records across historically established world states. Temporal density is a poor proxy for this objective: adjacent observations are often highly redundant in appearance and viewpoint, whereas an isolated observation may capture a distinctive configuration or viewing state that becomes relevant only when the corresponding content is revisited much later. This bounded-coverage view is closely related to coreset selection, which seeks a compact subset that preserves the utility of a substantially larger collection (Sener & Savarese, 2018).

ReCAP addresses this distinction through the nearest-neighbor dispersion objective in Eq. 7. For an atlas $\mathcal{A}$ containing at least two records, define the local isolation of a_i as

$$r_i(\mathcal{A}) = \min_{\substack{a_j \in \mathcal{A} \\ j \neq i}} d(a_i, a_j), \qquad D(\mathcal{A}) = \sum_{a_i \in \mathcal{A}} r_i(\mathcal{A}). \tag{42}$$

A small $r_i(\mathcal{A})$ indicates that a nearby alternative is already retained, whereas a large value indicates that a_i occupies a relatively sparse region of the adopted pose–appearance space. Under a fixed memory budget, this nearest-neighbor structure directly distinguishes redundancy from coverage: records concentrated in the same pose–appearance region contribute small isolation terms, whereas records distributed across complementary regions yield larger nearest-neighbor separations. Accordingly, summing these terms makes $D(\mathcal{A})$ an atlas-level surrogate for coverage breadth, favoring configurations that allocate limited memory across complementary observed states rather than repeatedly sampling redundant regions along the temporal trajectory.

This interpretation is particularly relevant to revisitation because the future utility of an observation is generally unknown when it is first encountered. A locally distinctive frame may never be needed again, whereas an apparently incidental view can later become important evidence for a re-entering subject, configuration, or viewpoint. ReCAP therefore does not attempt to predict future relevance directly. Instead, it preserves observations that remain visually or geometrically complementary to the current atlas, using pose–appearance diversity as a task-agnostic surrogate for maintaining recoverable historical support.

D.2 WHY POSITIVE-GAIN CURATION PRESERVES COMPLEMENTARY STATES

Once the atlas reaches capacity, ReCAP admits a new observation only when replacing an existing record strictly increases the dispersion objective. For a candidate replacement of a_i by a_{new}, define

$$\mathcal{A}^{(i \leftarrow \text{new})} = \mathcal{A} \setminus \{a_i\} \cup \{a_{\text{new}}\}. \tag{43}$$

For the selected index $i^\star$ in Eq. 8, an accepted replacement satisfies

$$D\Big(\mathcal{A}^{(i^\star \leftarrow \text{new})}\Big) = D(\mathcal{A}) + \gamma_{i^\star} > D(\mathcal{A}). \tag{44}$$

Recency alone is therefore insufficient for admission after the budget is saturated. The incoming observation must improve the atlas-level dispersion enough to compensate for the record displaced by its insertion. In this sense, the write rule allocates memory according to complementary pose–appearance support rather than continually replacing older observations with recent ones. Fixed-memory continual-learning methods similarly preserve representative or diverse exemplars from a

non-stationary stream (Rebuffi et al., 2017; Aljundi et al., 2019), although ReCAP defines diversity jointly in pose–appearance space and optimizes the retained set for later retrieval.

A simple discrete-state abstraction exposes this bias. Suppose each observation belongs to a persistent state ξ_i, and consider a full atlas with $K \geq 2$ under the idealized distance

$$d_{\mathrm{toy}}(a_i, a_j) = \mathbf{1}[\xi_i \neq \xi_j]\,, \tag{45}$$

where $\mathbf{1}[\cdot]$ is the indicator function. Let n_q denote the number of records representing state q. A record belonging to a duplicated state has a same-state neighbor at distance zero, whereas a state represented exactly once has nearest-neighbor distance one. Hence,

$$D_{\mathrm{toy}}(\mathcal{A}) = \sum_q \mathbf{1}[n_q = 1]\,. \tag{46}$$

If a state occupies exactly two slots and an unseen state arrives, replacing either duplicate makes the remaining old record unique while introducing a new unique state. The resulting gain therefore reflects not only the novelty of the incoming observation, but also how the replacement redistributes a limited set of memory slots across distinct states.

The actual pose–appearance space is continuous and substantially richer than this discrete abstraction. Equation 46 therefore does not imply that ReCAP recovers a globally optimal historical subset. It instead makes explicit the inductive bias of the positive-gain rule: repeated observations compete for overlapping support, whereas observations occupying underrepresented regions are favored when their admission improves the atlas configuration as a whole.

D.3 Monotonicity and Configuration-Level Replacement

The positive-gain rule yields a simple structural property of the saturated atlas. Define

$$d_{\mathrm{app}}(a_i, a_j) = 1 - \cos(Z_i, Z_j), \qquad d_{\mathrm{pose}}(a_i, a_j) = \frac{1}{2}\left[\frac{\theta_{ij}}{\pi} + \min\left(\frac{\|\mathbf{t}_i - \mathbf{t}_j\|_2}{s}, 1\right)\right], \tag{47}$$

so that the distance in Eq. 6 becomes

$$d(a_i, a_j) = (1-\alpha) d_{\mathrm{app}}(a_i, a_j) + \alpha d_{\mathrm{pose}}(a_i, a_j). \tag{48}$$

For $0 \leq \alpha \leq 1$, cosine similarity gives

$$0 \leq d_{\mathrm{app}}(a_i, a_j) \leq 2, \qquad 0 \leq d_{\mathrm{pose}}(a_i, a_j) \leq 1, \tag{49}$$

and therefore

$$0 \leq d(a_i, a_j) \leq 2(1-\alpha) + \alpha = 2 - \alpha. \tag{50}$$

For a full atlas containing K records, this immediately implies

$$0 \leq D(\mathcal{A}) \leq K(2-\alpha). \tag{51}$$

Let t index candidate updates after the atlas has reached capacity, and let D_t denote the dispersion under the current pose–appearance distance. Since a replacement is accepted only when its gain is positive,

$$D_t(\mathcal{A}_{t+1}) = \begin{cases} D_t(\mathcal{A}_t) + \gamma_{i^\star}, & \gamma_{i^\star} > 0, \\ D_t(\mathcal{A}_t), & \gamma_{i^\star} \leq 0. \end{cases} \tag{52}$$

Hence,

$$D_t(\mathcal{A}_{t+1}) \geq D_t(\mathcal{A}_t). \tag{53}$$

Thus, every accepted replacement increases the atlas dispersion under the current distance criterion, favoring complementary historical evidence without requiring global optimization over all historical subsets.

The replacement gain also reveals why ReCAP is not equivalent to scoring incoming novelty independently. For

$$\mathcal{A}' = \mathcal{A} \setminus \{a_i\} \cup \{a_{\mathrm{new}}\}, \tag{54}$$

define the isolation of the incoming record after replacement as

$$r_{\text{new}}(\mathcal{A}') = \min_{a_j \in \mathcal{A}' \setminus \{a_{\text{new}}\}} d(a_{\text{new}}, a_j). \tag{55}$$

Using the definition of $r_j(\cdot)$ in Eq. 42, the replacement gain decomposes exactly as

$$\begin{aligned}\gamma_i = r_{\text{new}}(\mathcal{A}') - r_i(\mathcal{A}) \\ + \sum_{\substack{a_j \in \mathcal{A} \\ j \neq i}} \left[r_j(\mathcal{A}') - r_j(\mathcal{A}) \right].\end{aligned} \tag{56}$$

The first line compares the isolation contributed by the incoming record with that of the removed record, while the second captures how the replacement changes the nearest-neighbor structure of every surviving record. A candidate may therefore be individually distinctive yet still be rejected if admitting it destroys more valuable existing coverage; conversely, removing a redundant record may improve several surviving isolation terms simultaneously. ReCAP thus evaluates the *configuration-level consequence* of a write rather than reducing persistent-memory curation to a per-record novelty score.

D.4 Why Pose and Appearance Must Be Coupled

The two components in Eq. 48 capture different forms of historical complementarity. Appearance alone is insufficient because geometrically distinct viewpoints may have similar visual representations. Geometry alone is insufficient because different identities, object configurations, or scene states may occur at nearly identical camera poses. With $\alpha = 0$, the neighborhood structure is determined entirely by d_{app}, so geometrically distinct views can be treated as redundant whenever their visual representations are similar. With $\alpha = 1$, the neighborhood structure depends entirely on d_{pose}, potentially ignoring visually meaningful state variation at similar viewpoints. Intermediate values allow both forms of difference to affect the nearest-neighbor structure that drives ReCAP curation.

The clipped translation component,

$$d_{\text{trans}}(a_i, a_j) = \min\left(\frac{\|\mathbf{t}_i - \mathbf{t}_j\|_2}{s}, 1 \right), \tag{57}$$

where $s > 0$ is the translation normalization scale, further controls the contribution of geometric separation. Once $\|\mathbf{t}_i - \mathbf{t}_j\|_2 > s$,

$$\frac{\partial d_{\text{trans}}}{\partial \|\mathbf{t}_i - \mathbf{t}_j\|_2} = 0. \tag{58}$$

Additional translation beyond this scale therefore cannot continuously increase the pose distance. The geometric term remains sensitive to nearby viewpoint changes while saturation prevents large spatial separation from dominating the joint diversity measure.

More broadly, appearance and geometry answer complementary questions about persistent evidence: appearance characterizes *what* world state was observed, whereas camera pose constrains *from where* that evidence was established. Revisitation can depend on either dimension and often on both. Their coupling therefore defines redundancy in a space better matched to persistent scene evidence than either appearance or viewpoint alone.

D.5 Block-Adaptive Projection as a Readout Interface

Curation determines which observations survive, but preservation alone does not make the same stored representation equally suitable throughout the generator. Atlas Records are stored as VAE latents, whereas the internal representations consumed at different transformer blocks vary with network depth. ReCAP therefore separates persistent storage from block-specific consumption through the block-dependent projection already defined in Eq. 9:

$$\mathbf{k}_i^{(\ell)} = \underbrace{W_{\text{atlas}}^{(\ell)} \mathcal{S}(Z_i)}_{\text{visual component}} + \underbrace{W_{\text{cam}}^{(\ell)} e(\pi_i)}_{\text{viewpoint component}}, \qquad \mathbf{v}_i^{(\ell)} = W_{\text{atlas}}^{(\ell)} \mathcal{S}(Z_i), \tag{59}$$

where $\mathcal{S}(\cdot)$ tokenizes the stored latent and $e(\pi_i)$ encodes its camera pose. The key therefore contains both projected visual content and a viewpoint cue for correspondence, whereas the value retains the projected visual component. Camera information can consequently influence where a historical observation should be matched without being directly inserted into the retrieved visual payload.

The block index ℓ supplies the second dimension of adaptation. Let $\mathcal{H}^{(\ell)}$ denote the internal representation space at transformer block ℓ. Although the persistent record $a_i = (Z_i, \pi_i, u_i)$ remains unchanged in storage, its readout

$$\Phi_\ell(a_i) = \left(\mathbf{k}_i^{(\ell)}, \mathbf{v}_i^{(\ell)}\right) \tag{60}$$

depends on the transformer block at which the record is consumed. ReCAP therefore does not repeatedly rewrite the stored historical observation to follow the evolving representation hierarchy of the generator. Instead, a stable stored record is translated into a block-specific interface at read time, allowing the same persistent evidence to support multiple representation levels without requiring multiple stored copies.

This distinction also locates ReCAP precisely within the *trust–preserve–align* pipeline. ARiSE changes the relative utility of persistent evidence by controlling the reliability of the short-term source. ReCAP then determines which historical states remain available and exposes them in representations compatible with the current generator block. RaDAR subsequently performs query-dependent routing and distillation over these candidates. ReCAP therefore connects persistent preservation to later alignment: diversity-gain curation determines what survives, while block-adaptive projection determines how the surviving evidence becomes consumable.

D.6 Choosing the Atlas Capacity

The atlas capacity K controls a natural trade-off between historical coverage and persistent-memory overhead. If K is too small, distinct historical states compete for the same limited slots, increasing the possibility that evidence required by a later revisitation is no longer represented. Increasing K allows more complementary states to remain available, but it also increases storage and enlarges the persistent candidate pool examined by subsequent routing. This behavior can be summarized by the qualitative proxy

$$\mathcal{E}_{\mathrm{mem}}(K) \approx \underbrace{\mathcal{E}_{\mathrm{miss}}(K)}_{\text{unrepresented historical evidence}} + \beta \underbrace{\mathcal{C}_{\mathrm{route}}(K)}_{\text{persistent candidate overhead}}, \tag{61}$$

where $\mathcal{E}_{\mathrm{mem}}$ denotes a conceptual memory-related cost, $\mathcal{E}_{\mathrm{miss}}$ represents loss caused by historical evidence that is no longer represented, $\mathcal{C}_{\mathrm{route}}$ represents the cost of processing the retained candidate pool, and $\beta \geq 0$ is a trade-off coefficient. The first term is expected to decrease as additional complementary observations can be retained, whereas the second generally increases with the size of the persistent candidate set. This is a qualitative decomposition rather than an additional optimization objective used by EonWeave.

Importantly, preservation and query-time utilization operate at different levels. ReCAP determines which observations remain *available* under the global memory budget, but retention does not imply that every stored record should influence every current query. RaDAR performs a subsequent query-dependent reduction. If P_q denotes the number of current query patches used by RaDAR, the resulting hierarchy can be summarized as

$$\underbrace{\text{growing historical observations}}_{\text{rollout history}} \xrightarrow{\mathrm{ReCAP}} \underbrace{\mathcal{A}_n}_{|\mathcal{A}_n| \leq K} \xrightarrow{\mathrm{RaDAR}} \underbrace{\left\{(\mathbf{k}_{\mathrm{ref}}^{(p)}, \mathbf{v}_{\mathrm{ref}}^{(p)})\right\}_{p=1}^{P_q}}_{P_q \text{ query-aligned references}}. \tag{62}$$

ReCAP therefore bounds and diversifies the evidence that remains available, whereas RaDAR determines how that retained evidence becomes active for the current generation state. Increasing the atlas capacity broadens the candidate universe available to retrieval; it does not force all retained history into the final generative context.

In summary, ReCAP couples three mutually reinforcing properties. Pose–appearance curation distributes a bounded set of records across complementary regions of the observed historical state space; once capacity is reached, positive-gain replacement ensures that every accepted replacement strictly increases the same dispersion criterion used for curation; and block-adaptive projection

translates stable persistent records into the representation spaces in which they are consumed. Together, these mechanisms turn fixed-capacity persistent memory from a passive historical archive into a coverage-oriented write–read interface for long-horizon revisitation.

E Further Analysis of RaDAR

This section provides further analysis of *Retrieval-aware Distillation into Aligned References (RaDAR)*. We examine why the combination of *noise-adaptive routing*, *mutual correspondence*, *selective abstention*, and *query-aligned distillation* is well suited to heterogeneous historical evidence. Rather than introducing an additional retrieval mechanism, the following analysis develops the interpretation already implicit in RaDAR: retrieval identifies potentially useful evidence, whereas distillation determines how that evidence is presented to each current query without allowing retrieval breadth or ambiguity to directly determine the final attention context.

E.1 From Retrieval to Evidence Distillation

Sparse historical retrieval is often treated as the final memory interface: high-scoring historical tokens are selected and directly exposed to downstream attention. This conflates two distinct operations. Retrieval determines which observations may contain relevant information, whereas generation requires a representation of how the selected evidence should support a particular current query. This distinction becomes especially important when candidates originate from different times, viewpoints, and sources, since their native token organization is not naturally aligned with the spatial structure of the current generation state.

RaDAR separates these two roles. Let the current hidden representation contain P_q query patches, and let $\mathcal{I}_p$ denote the candidate support selected for query patch p. For each selected candidate $m \in \mathcal{I}_p$, let $\bar{\mathbf{k}}_m$ and $\bar{\mathbf{v}}_m$ denote its pooled generative key and value, and let $\rho_{p,m}$ denote the mutual correspondence score defined below. Together with the dustbin score $\rho_{p,\varnothing}$ and the learnable null reference $(\mathbf{k}_\varnothing, \mathbf{v}_\varnothing)$, Eq. 15 maps the variable candidate support into a single pair

$$\left(\mathbf{k}_{\mathrm{ref}}^{(p)}, \mathbf{v}_{\mathrm{ref}}^{(p)}\right). \tag{63}$$

The retrieved candidates therefore serve as evidence from which a reference is constructed rather than being exposed as the final reference themselves. Producing one such pair for every query patch transfers heterogeneous evidence into a representation indexed by the current query layout, explicitly decoupling retrieval breadth from the structure of the context consumed by the generator.

E.2 Why Noise-Adaptive Routing Stabilizes Correspondence

The reliability of correspondence cues changes throughout denoising. At strongly corrupted states, the current hidden representation contains less discriminative appearance information, while camera-ray and patch-location geometry remains directly available. As denoising proceeds, appearance becomes increasingly informative and can distinguish states that geometry alone may not resolve. Diffusion-model features have been shown to support semantic, geometric, and temporal correspondence, while their usefulness depends on the selected denoising timestep and network layer (Luo et al., 2023; Tang et al., 2023). RaDAR captures this variation through the score in Eq. 11. Writing

$$s_{\mathrm{app}}(p, m; t) = \cos\left(\mathbf{r}_q^{(p)}, \mathbf{r}_k^{(m)}\right), \qquad s_{\mathrm{geo}}(p, m) = \cos\left(\mathbf{g}_q^{(p)}, \mathbf{g}_k^{(m)}\right), \tag{64}$$

the routing score can be expressed as

$$s_{p,m}(t) = \big(1 - \lambda(t)\big) s_{\mathrm{app}}(p, m; t) + \lambda(t) s_{\mathrm{geo}}(p, m), \qquad \lambda(t) = \mathrm{sigmoid}(a\bar{t} + b), \tag{65}$$

where $\bar{t} \in [0, 1]$ is the normalized corruption level, and a, b are learnable scalars initialized to encode a monotonic prior that favors geometry under stronger corruption and appearance as denoising proceeds, while allowing the routing schedule to adapt during optimization.

A two-candidate margin model makes the role of this interpolation explicit. For query patch p, let m^+ denote a desired correspondence and m^- a distractor, and define

$$\Delta_{\mathrm{app}}(t) = s_{\mathrm{app}}(p, m^+; t) - s_{\mathrm{app}}(p, m^-; t), \qquad \Delta_{\mathrm{geo}} = s_{\mathrm{geo}}(p, m^+) - s_{\mathrm{geo}}(p, m^-). \tag{66}$$

The resulting score margin is

$$\Delta s(t) = \big(1 - \lambda(t)\big)\Delta_{\mathrm{app}}(t) + \lambda(t)\Delta_{\mathrm{geo}}. \tag{67}$$

When appearance is misleading but geometry favors the desired correspondence,

$$\Delta_{\mathrm{app}}(t) < 0, \qquad \Delta_{\mathrm{geo}} > 0, \tag{68}$$

the combined margin remains positive whenever

$$\lambda(t) > \frac{-\Delta_{\mathrm{app}}(t)}{\Delta_{\mathrm{geo}} - \Delta_{\mathrm{app}}(t)}. \tag{69}$$

This relation is not an optimal gating rule for the learned router, but it exposes the intended role of noise adaptation: geometric evidence can stabilize correspondence when appearance is unreliable, while appearance can provide increasingly fine discrimination as the visual representation becomes more informative. The learned $\lambda(t)$ allows this balance to vary continuously rather than imposing a fixed appearance–geometry mixture throughout denoising.

E.3 Mutual Correspondence as Two-Sided Competition

A large query-to-candidate score alone does not imply a distinctive correspondence. Repetitive textures, background regions, or visually generic historical patches may be compatible with many current queries simultaneously. Transformer-based matchers such as LoFTR likewise use cross-image context to disambiguate correspondences in low-texture and repetitive regions (Sun et al., 2021). RaDAR therefore combines query-side preference and candidate-side specificity through

$$\rho_{p,m} = \pi_{m|p}\pi_{p|m}, \tag{70}$$

where $\pi_{m|p}$ measures how strongly query p selects candidate m, while $\pi_{p|m}$ measures how strongly candidate m favors query p relative to the other current queries.

Consider a simplified candidate m that has approximately equal scores for $R \geq 1$ query patches and negligible affinity to the remaining queries. Candidate-side normalization then gives approximately

$$\pi_{p|m} \approx \frac{1}{R}, \qquad \rho_{p,m} \approx \frac{1}{R}\pi_{m|p}. \tag{71}$$

A candidate that appears broadly compatible with many current locations is therefore suppressed relative to one that is specific to a particular query, even when its query-side similarity is large. Mutual routing does not enforce a hard one-to-one assignment; instead, it introduces a soft specificity correction before sparse candidate selection.

The same effect can be written exactly in log space while retaining the query-side dustbin. For temperature $\tau > 0$ and dustbin logit z, define

$$Z_p = \exp(z/\tau) + \sum_m \exp\Big(\frac{s_{p,m}}{\tau}\Big), \qquad Z_m = \sum_i \exp\Big(\frac{s_{i,m}}{\tau}\Big). \tag{72}$$

Using the mutual normalization in Eq. 12,

$$\log \rho_{p,m} = \frac{2s_{p,m}}{\tau} - \log Z_p - \log Z_m. \tag{73}$$

The raw correspondence score therefore competes in two directions. The two normalization terms account for competition among historical candidates for each query and among current queries for each candidate. In addition,

$$0 \leq \rho_{p,m} \leq \pi_{m|p}, \qquad 0 \leq \rho_{p,m} \leq \pi_{p|m}, \tag{74}$$

because both factors lie in $[0, 1]$. A correspondence cannot therefore obtain a large mutual score when either its query-side preference or its candidate-side specificity is weak. This two-sided competition explains why RaDAR can prefer a more query-specific historical candidate even when one-directional similarities are comparable.

E.4 The Dustbin as a Selective Abstention State

Some current queries may have no appropriate condition or atlas correspondence. This occurs naturally in newly observed regions, for states not preserved by the bounded atlas, or when all available historical candidates are ambiguous. Forcing every query to select a real candidate would convert the absence of convincing evidence into an artificial historical match. An analogous dustbin is used by SuperGlue to absorb unmatched features in differentiable partial assignment (Sarlin et al., 2020); RaDAR adapts this principle to query-conditioned historical retrieval. RaDAR instead includes a learnable dustbin logit z in query-side normalization:

$$\pi_{\varnothing|p} = \frac{\exp(z/\tau)}{\exp(z/\tau) + \sum_{m=1}^{M} \exp(s_{p,m}/\tau)}, \tag{75}$$

where M denotes the number of real candidates considered for query p. If every real candidate lies at least a margin $\Delta > 0$ below the dustbin,

$$s_{p,m} \le z - \Delta \qquad \forall m, \tag{76}$$

then

$$\pi_{\varnothing|p} \ge \frac{1}{1 + M\exp(-\Delta/\tau)}. \tag{77}$$

Thus, as all real candidates become less competitive relative to the dustbin, a larger fraction of query-side mass can be assigned to the null state rather than to the least implausible historical match.

This abstention mass is retained during reference distillation. From Eq. 14,

$$\alpha_{p,\varnothing} + \sum_{m\in\mathcal{I}_p} \alpha_{p,m} = 1, \qquad \alpha_{p,\varnothing} \ge 0, \qquad \alpha_{p,m} \ge 0. \tag{78}$$

The resulting reference therefore interpolates between selected evidence and the learnable null reference. The dustbin should not be interpreted as a calibrated uncertainty estimator; it provides an operational abstention state that prevents historical evidence from being forced into every query when the current candidate set offers no convincing correspondence.

E.5 Query-Aligned Distillation as Normalized Aggregation

After routing, RaDAR distills the selected support into exactly one reference pair for each query patch. Because the normalized weights satisfy Eq. 78, the resulting references are convex combinations of the selected candidate representations and the null state:

$$\mathbf{k}_{\mathrm{ref}}^{(p)} \in \operatorname{conv}\left(\left\{\bar{\mathbf{k}}_m\right\}_{m\in\mathcal{I}_p} \cup \{\mathbf{k}_\varnothing\}\right), \tag{79}$$

with an analogous relation for $\mathbf{v}_{\mathrm{ref}}^{(p)}$. This normalized aggregation prevents the magnitude of the reference from growing merely because more candidates have been retained. For example, if the selected and null values satisfy

$$\|\bar{\mathbf{v}}_m\|_2 \le B, \qquad \|\mathbf{v}_\varnothing\|_2 \le B \tag{80}$$

for some $B \ge 0$, then

$$\begin{aligned}\left\|\mathbf{v}_{\mathrm{ref}}^{(p)}\right\|_2 &\le \sum_{m\in\mathcal{I}_p} \alpha_{p,m}\|\bar{\mathbf{v}}_m\|_2 + \alpha_{p,\varnothing}\|\mathbf{v}_\varnothing\|_2 \\ &\le B.\end{aligned} \tag{81}$$

A related local property holds when the routing support and weights are fixed. Let the selected values change by $\delta\mathbf{v}_m$ while the null reference remains unchanged. The induced perturbation of the distilled reference is

$$\delta\mathbf{v}_{\mathrm{ref}}^{(p)} = \sum_{m\in\mathcal{I}_p} \alpha_{p,m}\delta\mathbf{v}_m, \tag{82}$$

which implies

$$\begin{aligned}\left\|\delta\mathbf{v}_{\mathrm{ref}}^{(p)}\right\|_2 &\le \sum_{m\in\mathcal{I}_p} \alpha_{p,m}\|\delta\mathbf{v}_m\|_2 \\ &\le (1-\alpha_{p,\varnothing}) \max_{m\in\mathcal{I}_p} \|\delta\mathbf{v}_m\|_2 .\end{aligned} \tag{83}$$

Under this fixed-routing comparison, reference distillation therefore does not amplify the largest perturbation among the selected real values, and nonzero null mass further reduces their aggregate contribution. This is a property of the aggregation operation itself, not a global stability guarantee for RaDAR: perturbations to queries or candidates can also alter the routing scores, selected support, and weights.

The same normalized representation provides a soft treatment of ambiguous evidence. When several candidates receive comparable mutual scores, RaDAR need not expose each of them independently or commit to a single hard match. Their evidence can remain distributed through the normalized coefficients until it is represented by the single query-aligned reference and fused with the local context. Distillation therefore changes retrieval from direct token insertion into a bounded evidence-aggregation stage.

E.6 Choosing Retrieval Breadth Without Expanding Attention

RaDAR uses separate condition and atlas retrieval budgets. Define

$$\mathcal{I}_p^{\mathrm{cond}} = \mathrm{TopIdx}_{r_c}(\{\rho_{p,m}\}_{m\in\mathcal{J}_{\mathrm{cond}}}), \qquad \mathcal{I}_p^{\mathrm{atlas}} = \mathrm{TopIdx}_{r_a}(\{\rho_{p,m}\}_{m\in\mathcal{J}_{\mathrm{atlas}}}), \tag{84}$$

so that

$$\mathcal{I}_p = \mathcal{I}_p^{\mathrm{cond}} \cup \mathcal{I}_p^{\mathrm{atlas}}, \qquad \left|\mathcal{I}_p^{\mathrm{cond}}\right| \le r_c, \qquad \left|\mathcal{I}_p^{\mathrm{atlas}}\right| \le r_a. \tag{85}$$

Here r_c and r_a denote the condition and atlas retrieval budgets, respectively. When atlas candidates are available, this separation reserves selection capacity for persistent evidence instead of forcing all candidates to compete for a single shared top-k support. It does not force persistent memory to contribute to the final reference: selected candidates are still weighted by their mutual correspondence scores, and the dustbin remains available when the selected historical evidence is weak.

Increasing r_c or r_a broadens the support available to distillation, but it does not increase the number of references ultimately presented by RaDAR. Because one pair is produced for every query patch,

$$\left|\left\{\left(\mathbf{k}_{\mathrm{ref}}^{(p)}, \mathbf{v}_{\mathrm{ref}}^{(p)}\right)\right\}_{p=1}^{P_q}\right| = P_q. \tag{86}$$

If the intra-segment local context contains L_{local} key–value positions, the final attention context therefore has length

$$L_{\mathrm{ctx}} = L_{\mathrm{local}} + P_q. \tag{87}$$

By contrast, a design that directly appended every selected candidate independently for every query could require up to

$$L_{\mathrm{local}} + P_q(r_c + r_a) \tag{88}$$

positions in the corresponding worst case. RaDAR therefore decouples retrieval breadth from the width of the historical representation presented to the final attention operator.

This fixed final cardinality should not be confused with constant total computation. Routing descriptors and correspondence scores must still be computed over the available candidate set, so retrieval cost remains dependent on the number of condition and atlas candidates examined before distillation. What RaDAR bounds is the size of the distilled reference context, not the cost of discovering that context. ReCAP and RaDAR therefore solve complementary compression problems: ReCAP bounds the persistent candidate universe across time, whereas RaDAR reduces the available condition and atlas evidence to one current reference per query patch. Together with ARiSE, the resulting organization follows EonWeave's *trust–preserve–align* principle: ARiSE controls the relative reliability of short- and long-term evidence, ReCAP preserves complementary historical states under a bounded budget, and RaDAR aligns the available evidence with the current queries that require it.

In summary, RaDAR couples four mutually reinforcing properties. Noise-adaptive routing changes the balance between appearance and geometry over the denoising trajectory; mutual normalization suppresses candidates that are attractive but insufficiently query-specific; the dustbin provides an explicit null state when no convincing correspondence is available; and normalized reference distillation converts variable candidate support into fixed-size, query-aligned context. The analyses above further show that mutual routing implements two-sided competition and that, for fixed routing, the distillation operation behaves as a bounded convex aggregation of selected evidence. Together, these mechanisms transform historical retrieval from direct token selection into a structured evidence-to-reference interface for long-horizon generation.

Table 5: **Full EonWeave versus its DMD-distilled variant under HM-World autoregressive generation.** Both variants use the same autoregressive evaluation pipeline. NFE denotes the number of network function evaluations required to generate one segment; the full model uses explicit CFG and therefore requires two network evaluations per denoising step. The $\pm$ values denote standard deviations. The Δ row reports EonWeave-DMD minus EonWeave (Full), computed from the underlying full-precision measurements; consequently, small discrepancies may arise when subtracting the rounded values displayed in the table. Best results are shown in bold.

Method	**Inference Efficiency**			**Visual Fidelity**			**Revisitation Consistency**		**Degradation**
	Steps	CFG	NFE/Seg. ↓	PSNR ↑	SSIM ↑	LPIPS ↓	Subj. Cons. ↑	Backg. Cons. ↑	$\Delta^{\mathrm{Qal}}_{\mathrm{Dft}}$ ↓
EonWeave (Full)	50	✓	100	22.604 ± 3.780	0.6817 ± 0.1201	0.2893 ± 0.0889	0.9567	0.9603	0.2176
EonWeave-DMD	**4**	–	**4**	$\mathbf{23.685 \pm 3.707}$	$\mathbf{0.7098 \pm 0.1146}$	$\mathbf{0.2770 \pm 0.0872}$	**0.9603**	**0.9652**	**0.2136**
Δ (DMD − Full)	–	–	$25\times$ reduction	$+1.081$	$+0.0282$	-0.0122	$+0.0036$	$+0.0048$	-0.0040

F Analysis of Few-Step Distribution Matching Distillation

In addition to evaluating the full EonWeave model, we examine how few-step distribution matching distillation alters both segment-level generation quality and long-horizon autoregressive behavior. The original EonWeave generator performs 50 denoising steps with explicit classifier-free guidance (CFG), requiring conditional and unconditional network evaluations at every step. Its distilled counterpart, denoted as **EonWeave-DMD**, contracts this inference process to only 4 denoising steps and eliminates explicit CFG during deployment. Few-step generation has also been studied through deterministic implicit sampling, progressive sampler distillation, and consistency learning (Song et al., 2021; Salimans & Ho, 2022; Song et al., 2023). Both variants are evaluated on exactly the same HM-World autoregressive test set, comprising $1{,}000$ rollout sequences and $3{,}000$ generated chunks, while sharing the same autoregressive pipeline, context-propagation procedure, and persistent-memory pathway. The comparison therefore isolates the transition from the original guided multi-step generator to the complete few-step distilled inference regime, including its associated sampling procedure.

The distillation procedure is constructed around the actual few-step deployment trajectory rather than isolated regression to teacher predictions. Starting from the guided EonWeave teacher, the student is optimized under the same four-step denoising schedule used at inference. Backward simulation exposes the student to the sequence of states induced by this shortened trajectory, thereby reducing the discrepancy between optimization-time inputs and those recursively encountered during deployment. Within this trajectory, the first denoising step is supervised through teacher velocity regression, whereas subsequent steps are optimized by distribution matching following the role-separated design of DP-DMD (Wu et al., 2026a). The guided teacher prediction is incorporated directly into the student target, allowing the guidance effect to be absorbed by a single student branch and removing the need for explicit CFG at inference. EonWeave-DMD should therefore not be viewed as a four-step imitation of the intermediate states traversed by the original sampler, but as a separately optimized few-step generator whose shortened inference path is trained directly for its deployment regime.

Quality–Computation Characteristics. Tab. 5 summarizes the overall comparison. Despite reducing the denoising trajectory from 50 steps to 4, EonWeave-DMD improves every reported reconstruction and consistency metric. PSNR increases from 22.604 dB to 23.685 dB, yielding a gain of 1.081 dB, while SSIM rises from 0.6817 to 0.7098. The improvement is not confined to distortion-oriented criteria: LPIPS decreases from 0.2893 to 0.2770, accompanied by higher Subject Consistency and Background Consistency. The agreement across pixel-level, perceptual, and temporal criteria makes a purely distortion-driven explanation less compelling. These measurements nevertheless characterize fidelity and consistency rather than sample diversity, and therefore do not establish preservation of the complete teacher distribution.

The computational reduction becomes particularly pronounced once CFG is accounted for. The full model requires two network evaluations at each of its 50 denoising steps, whereas EonWeave-DMD performs a single evaluation at each of four steps. The number of network function evaluations

Table 6: **Chunk-wise comparison under the HM-World autoregressive protocol.** Chunk 1 is generated directly from the initial conditioning frames, whereas Chunks 2 and 3 additionally depend on previously generated content.

Chunk	Method	PSNR ↑	SSIM ↑	LPIPS ↓	Subj. Cons. ↑
1	EonWeave (Full)	25.862	0.788	0.2028	0.9560
1	**EonWeave-DMD**	**26.987**	**0.811**	**0.1942**	**0.9580**
2	EonWeave (Full)	21.999	0.673	0.2930	0.9552
2	**EonWeave-DMD**	**23.125**	**0.707**	**0.2770**	**0.9585**
3	EonWeave (Full)	19.949	0.585	0.3720	0.9591
3	**EonWeave-DMD**	**20.943**	**0.612**	**0.3600**	**0.9644**

(NFE) per generated segment is therefore reduced from 100 to 4:

$$\frac{\mathrm{NFE}_{\mathrm{Full}}}{\mathrm{NFE}_{\mathrm{DMD}}} = \frac{50 \times 2}{4} = 25. \tag{89}$$

We report this as a $25\times$ reduction in NFE rather than an equivalent wall-clock speedup, since realized runtime additionally depends on implementation and system overhead. Notably, under the evaluated setting, this substantial contraction of iterative computation is accompanied by improved rather than degraded generation metrics.

Chunk-Wise Behavior and Long-Horizon Stability. The chunk-wise results in Tab. 6 further localize the source of this advantage. EonWeave-DMD already outperforms the full model on the first generated chunk, before any previously generated chunk can recursively enter the conditioning history. First-chunk PSNR increases from 25.862 dB to 26.987 dB, corresponding to an improvement of approximately 1.13 dB, with concurrent gains in SSIM and LPIPS.

Crucially, this margin remains comparatively stable throughout the rollout, with PSNR gains of approximately $+1.13$, $+1.13$, and $+0.99$ dB across the three generated chunks. The improvement therefore cannot be explained solely by reduced late-stage autoregressive error accumulation. Rather, the distilled generator establishes a stronger segment-level operating point before cross-segment recursion begins, and this advantage remains largely intact after its outputs are repeatedly reincorporated into subsequent conditioning history. Consistently, $\Delta^{\mathrm{Qal}}_{\mathrm{Dft}}$ decreases slightly from 0.2176 to 0.2136, indicating that the aggressive reduction in sampling length does not incur additional long-horizon degradation under the evaluated protocol.

Distribution Matching as Inference-Path Reparameterization. The observation that a four-step student can outperform its 50-step teacher becomes less surprising once distribution matching is distinguished from trajectory imitation. Conventional trajectory distillation encourages the student to reproduce states or transitions associated with a particular teacher sampler. DMD instead optimizes the generated distribution through a score-difference objective (Yin et al., 2024b); reproducing the teacher's complete sequence of intermediate states is therefore not itself the optimization target. Related observations in DMD2 further show that few-step distribution-matched generators can, under appropriate settings, surpass the multi-step models from which their supervision is derived (Yin et al., 2024a).

This distinction motivates a more useful interpretation of EonWeave-DMD: distillation *reparameterizes the inference path*. The original model realizes generation through 50 sequentially coupled denoising transitions, whereas the distilled model learns a substantially shorter four-step path toward the distribution targeted by distillation. Because the objective is distributional rather than trajectory preserving, the student can redistribute the required transformation across this shorter sequence instead of inheriting every intermediate computation of the teacher sampler. Sampling contraction is therefore not merely a reduction in numerical steps; it defines a different finite dynamical path through which the final sample is produced.

Backward simulation is central to making this contracted path self-consistent. Standard Flow Matching training supervises predictions at individually sampled noise levels, whereas inference recursively feeds the output of one denoising transition into the next. DMD2 characterizes this discrep-

ancy as a training–inference input mismatch and mitigates it by optimizing on states generated by the model's own inference process (Yin et al., 2024a). EonWeave-DMD follows the same principle: the four-step trajectory is explicitly simulated during training, so subsequent objectives are evaluated on states induced by the shortened sampling process itself. The student is therefore optimized as a coherent four-step generator rather than assembled from independently trained local transitions.

This distinction becomes particularly relevant in autoregressive world generation, where recurrence occurs at two nested scales. Within each generated segment, deviations may propagate across successive denoising transitions; across segments, the completed generation becomes part of the conditioning history and can influence all subsequent predictions. A longer inner trajectory introduces more recurrent state transitions before a segment is propagated to the outer autoregressive process. EonWeave-DMD substantially contracts this inner recurrence, while backward simulation aligns optimization with the states created by that contraction. This provides a coherent interpretation of the pattern in Tab. 6: the performance advantage is established before cross-segment error propagation begins and remains largely preserved once generated segments enter future context.

The role-separated objective further structures this shortened trajectory. The first denoising step receives direct teacher velocity supervision, whereas later transitions are optimized through distribution matching following DP-DMD (Wu et al., 2026a). The initial transition therefore provides an explicit teacher-anchored entry into the few-step trajectory, while subsequent transitions retain the flexibility of distribution-level optimization. This separation is also relevant to the reverse-KL tendency associated with DMD, which can emphasize high-density regions and become mode seeking when applied uniformly. Combining an explicitly anchored first transition with distribution matching at later stages provides a principled separation between trajectory anchoring and subsequent distributional refinement, rather than imposing a single objective uniformly across the complete sampling path.

Guidance folding fits naturally within the same inference-path interpretation. The original model realizes CFG by repeatedly evaluating conditional and unconditional branches and combining them at every denoising step. During distillation, the guided teacher prediction is instead incorporated directly into the student target, following prior work on classifier-free guidance distillation (Meng et al., 2023). EonWeave-DMD consequently learns the guided transformation within a single inference branch rather than reconstructing it through repeated conditional–unconditional evaluations. Combined with four-step sampling, this reduces the deployed process from 100 network evaluations to 4 and removes repeated guidance recombination from the inference trajectory.

Taken together, backward simulation, role-separated supervision, distribution matching, and guidance folding are better understood as coupled elements of a single inference regime rather than independent explanations for the observed gain. They define a short, deployment-aligned, guidance-absorbed trajectory whose optimization explicitly reflects the states encountered during few-step sampling. Under this view, the student need not constitute a uniformly degraded approximation of its teacher. It can instead realize an inference path better adapted to the particular four-step operating regime, even though the teacher remains the source of supervision.

Interaction with Persistent Historical Modeling. Importantly, few-step distillation modifies the *intra-segment* generation path without changing the structural role of EonWeave's persistent-history interface. The distilled generator continues to receive the same two classes of contextual evidence as the full model: a recent autoregressive condition whose reliability is regulated by ARiSE, and persistent historical states maintained by ReCAP and transformed into query-aligned references by RaDAR. What changes is the trajectory through which the current segment is synthesized from this context, rather than the organization of the context itself.

Distillation and persistent historical modeling therefore operate at different but complementary temporal scales. Distribution matching contracts the iterative dynamics used to realize an individual segment, whereas the *trust–preserve–align* pipeline governs how information is trusted, retained, and reused across segments. In particular, shortening the denoising path does not remove the historical states curated by ReCAP, alter the source asymmetry induced by ARiSE, or replace the query-dependent historical readout performed by RaDAR. The few-step student retains the same persistent-evidence pathway while learning a different finite trajectory for transforming that evidence, together with the current condition, into the next segment.

This separation is important for understanding why aggressive sampling contraction remains compatible with long-horizon world modeling. Distillation addresses the computational and optimization structure within a generated segment, but does not by itself solve the problem that previously observed world states may disappear from the active condition for extended periods. Conversely, persistent historical modeling determines what information remains available across the rollout and how it is reintroduced when relevant, but does not prescribe the number of denoising transitions used to synthesize each segment. EonWeave-DMD combines these two axes: it substantially shortens the local generative path while preserving the cross-segment historical interface through which remote world states remain accessible during autoregressive generation.

The resulting organization also clarifies how distillation fits into the broader design of EonWeave. ARiSE, ReCAP, and RaDAR determine the structure of the evidence supplied to generation—respectively regulating source reliability, preserving complementary historical states, and aligning the retained evidence with the current query. DMD instead changes how the generator traverses from noise to the current segment once this evidence has been provided. The two mechanisms are therefore complementary rather than substitutive: the former maintains long-range informational continuity across the rollout, while the latter compresses the local inference trajectory used to realize each step of that rollout. This distinction allows the *trust–preserve–align* organization to remain intact even after the sampling process is reduced from 50 guided denoising steps to a four-step single-branch student.

Implications and Claim Boundary. The empirical evidence is consistent with this inference-path reparameterization interpretation. EonWeave-DMD achieves higher PSNR and SSIM, lower LPIPS, and stronger subject and background consistency while requiring $25\times$ fewer network evaluations. Its approximately 1 dB PSNR advantage is already established in the first generated chunk and remains throughout the autoregressive rollout, while the measured long-horizon quality drop is marginally smaller. The observed gain is therefore more consistent with an improvement established by the few-step generator itself and subsequently preserved through autoregressive propagation than with an advantage created only by reduced late-stage error accumulation.

The current comparison nevertheless evaluates the distilled system as a whole. Distribution matching, backward simulation, role-separated supervision, guidance folding, and sampling length are modified jointly, so their individual causal contributions cannot be disentangled from the teacher–student comparison alone. Moreover, although the consistent improvements across distortion, perceptual, and temporal metrics argue against a simple degradation in fidelity, none of these measurements directly evaluates sample diversity. The results should therefore not be interpreted as evidence that DMD eliminates mode collapse, preserves the complete teacher distribution, or generally enables few-step students to outperform their teachers.

The conclusion supported by the present HM-World autoregressive evaluation is narrower. Distribution matching distillation transforms EonWeave's costly guided sampler into a substantially shorter inference trajectory that is explicitly optimized for its own deployment process. Under this setting, the resulting four-step generator reduces computation by $25\times$ in NFE while improving the evaluated generation metrics and retaining its advantage under recursive rollout. Determining how much of this behavior arises from trajectory alignment, objective design, guidance folding, or sampling contraction individually would require a controlled decomposition of these distillation components.

G Extended Experimental Evaluation and Analysis

G.1 Experimental Setup

Datasets. We evaluate EonWeave on two complementary benchmarks that emphasize different aspects of persistent world modeling and long-horizon generation.

HM-World (Chen et al., 2026a) is a synthetic revisitation benchmark at 480×832, with decoupled camera and subject trajectories that cause scene content to repeatedly leave the field of view and later reappear. This controlled setting directly evaluates whether previously observed world states can be faithfully recovered across temporal gaps. We consider two protocols. Under *single-shot generation*, following HyDRA, the model is conditioned on 77 frames and generates the subsequent 77 frames in a single pass, evaluating revisitation consistency within a fixed generation horizon.

Under *autoregressive generation*, the model is conditioned on 33 frames and recursively generates three additional 33-frame segments, further testing whether persistent world-state information can be maintained under accumulated rollout errors. Following the protocol of HyDRA (Chen et al., 2026a), we use $54{,}915$ HM-World sequences for training and reserve $1{,}000$ sequences for testing. All methods are evaluated under the same dataset split and experimental protocol to ensure a controlled and fair comparison.

Mira-World (Ju et al., 2024) provides a substantially longer and more diverse evaluation setting, comprising $12{,}390$ sequences of 330 frames at 640×360. We use $12{,}140$ sequences for training and reserve 250 held-out sequences for evaluation. Each sequence is divided into 33-frame segments: starting from the first observed segment, the model autoregressively generates up to nine subsequent segments to complete the full 330-frame sequence. Compared with HM-World, this protocol imposes a markedly more challenging regime, combining diverse real-world content with prolonged recursive generation, where prediction errors accumulate and previously observed world states must remain recoverable across substantially longer temporal gaps.

Baselines. We compare EonWeave against representative state-of-the-art methods, including ReCamMaster (Bai et al., 2025), DFoT (Song et al., 2025), Context-as-Memory (Yu et al., 2025), WorldMem (Xiao et al., 2025), UCM (Xu et al., 2026), HyDRA (Chen et al., 2026a), and Self-Forcing (Huang et al., 2025). To ensure a controlled comparison, all methods use the same benchmark training and held-out test splits. Baselines are initialized from their officially released checkpoints and trained following their method-specific official optimization recipes until convergence, allowing each approach to be evaluated at its converged operating point rather than under an arbitrarily shared training budget. A fixed random seed is used consistently throughout training and evaluation to control stochastic variation. After training, all methods are evaluated under identical conditioning, generation, autoregressive rollout, and metric protocols. We additionally construct **AR-HyDRA**, an autoregressive extension of HyDRA, and train and evaluate it under the same data and rollout settings, enabling a direct and controlled comparison under recursive long-horizon generation.

Metrics. Our evaluation considers three complementary dimensions of generation performance. First, *visual fidelity* is measured against paired ground-truth frames using frame-averaged PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018), capturing pixel-level, structural, and perceptual quality. Second, *revisitation consistency* is evaluated using VBench Subject Consistency and Background Consistency (Huang et al., 2024), following the standardized evaluation adopted by HyDRA (Chen et al., 2026a) for direct benchmark comparability. Although these sequence-level metrics do not explicitly condition on individual exit–re-entry events, they quantify the two persistent properties that revisitation fundamentally requires: preservation of subject identity and appearance, measured through DINO feature similarity, and preservation of scene-level background state, measured through CLIP feature similarity. On HM-World, where controlled trajectories repeatedly move previously observed content out of view and subsequently reintroduce it, failures to recover established subjects or scene context consequently manifest as degraded consistency along these dimensions. We therefore use them as standardized quantitative proxies for revisitation consistency and complement them with re-entry-focused qualitative comparisons that directly examine the recovery of subject identity, object presence, and scene structure after prolonged absence.

Finally, we quantify *long-horizon degradation* using the normalized PSNR drift index $\Delta_{\mathrm{Dft}}^{\mathrm{Qal}}$. Let N denote the number of evaluation sequences, and let $P_{i,1}$ and P_{i,L_i} be the mean frame-wise PSNR values, in dB, of the first and final generated chunks of sequence i, respectively. We define

$$\Delta_{\mathrm{Dft}}^{\mathrm{Qal}} = \frac{1}{N} \sum_{i=1}^{N} \frac{P_{i,1} - P_{i,L_i}}{P_{i,1}}. \tag{90}$$

Rather than interpreting this quantity as a physical percentage change in reconstruction error, we use it as a dimensionless normalization of rollout-induced PSNR drift. This normalization is performed independently for each sequence before averaging, reducing sensitivity to differences in initial reconstruction quality and enabling more comparable assessment of degradation across heterogeneous sequences. In contrast, an absolute PSNR drop measures raw quality loss in dB but does not account for variation in the starting fidelity of individual sequences. We therefore use $\Delta_{\mathrm{Dft}}^{\mathrm{Qal}}$ together with

Table 7: **Expanded ablation study on Mira-World.**

Method	PSNR↑	SSIM↑	LPIPS↓	Subj. Cons.↑	Backg. Cons.↑	$\Delta_{\text{Dft}}^{\text{Qal}}$ ↓
w/o ReCAP	20.678	0.655	0.484	0.954	0.966	0.417
w/o ARiSE	20.735	0.518	0.483	0.957	0.971	0.312
w/o RaDAR	22.018	0.654	0.497	0.949	0.966	0.388
EonWeave (Full)	23.571	0.703	0.439	0.968	0.976	0.337
EonWeave-DMD	25.222	0.751	0.394	0.985	0.985	0.321

absolute PSNR, endpoint quality, and segment-wise PSNR trajectories, providing complementary views of both generation quality and its stability over long autoregressive rollouts. The first generated chunk excludes the conditioning segment, and lower values indicate stronger resistance to rollout-induced degradation.

Implementation Details. EonWeave is built upon Wan2.1-T2V-1.3B. The Persistent Scene Atlas maintains $K{=}18$ records, with the pose–appearance weighting coefficient set to $\alpha{=}0.4$. For ARiSE, we use an augmentation probability of $p_{\text{aug}}{=}0.5$ and a maximum training noise level of $\sigma_{\max}{=}0.5$; during inference, the re-noising strength is set to $\sigma^{\star}{=}0.15$. RaDAR operates on 2×2 spatial patches, with retrieval budgets $r_c{=}r_a{=}4$. The intra-segment local context is constructed using a 5-frame window.

The base EonWeave model is trained for 15,000 steps on 32 A800 GPUs using AdamW with a learning rate of 10^{-4}. We further apply distribution matching distillation (DMD) to obtain a few-step inference variant. Specifically, the distilled model performs generation with four denoising steps and does not require classifier-free guidance during inference, while retaining the same autoregressive conditioning and persistent-memory pipeline as the full model. This distilled variant is used to provide efficient inference without changing the long-horizon generation protocol described above.

G.2 Extended Analysis of Component Contributions

To examine whether ReCAP, ARiSE, and RaDAR affect the same aspect of long-horizon generation or address distinct failure modes, we extend the ablation study in the main paper with SSIM and Background Consistency. Each variant removes one component while leaving the remainder of the pipeline unchanged. As summarized in Tab. 7, the additional metrics distinguish reference-based structural fidelity from feature-level temporal consistency, enabling a more fine-grained analysis of the three components.

The results for the metrics already reported in the main paper follow the same trends. Removing ReCAP or ARiSE causes the two largest drops in PSNR, whereas removing RaDAR produces the highest LPIPS and the lowest Subject Consistency. These patterns are consistent with the intended roles of the three components: ReCAP preserves historical evidence, ARiSE regulates the use of unreliable autoregressive conditions, and RaDAR adapts retrieved evidence to the current query. More importantly, the additional metrics reveal distinct degradation profiles rather than a uniform loss of quality.

The most pronounced difference appears in structural fidelity. Without ARiSE, SSIM falls from 0.703 to 0.518, a reduction of 0.185. By comparison, removing ReCAP or RaDAR reduces SSIM by only 0.048 or 0.049, respectively. The SSIM penalty without ARiSE is therefore nearly four times as large, even though its PSNR reduction is comparable to that caused by removing ReCAP. This behavior is consistent with the intended role of ARiSE: reducing excessive reliance on recursively generated local conditions whose reliability deteriorates as prediction errors accumulate. The resulting metric profile suggests that reliability alignment is particularly important for maintaining spatial and structural correspondence during autoregressive generation.

Background Consistency exhibits a different pattern. Removing either ReCAP or RaDAR lowers the score from 0.976 to 0.966, whereas the model without ARiSE retains a higher value of 0.971 despite its substantially lower SSIM. This divergence indicates that temporal background consistency alone does not guarantee accurate reconstruction: a generated scene can remain stable at the feature level while deviating from the reference in local appearance or geometry. The similar background penal-

Table 8: **Segment-wise PSNR evolution on Mira-World.** EonWeave maintains the highest generation fidelity at every segment throughout the nine-step autoregressive rollout. For brevity, EonWeave and HyDRA denote EonWeave-DMD and AR-HyDRA, respectively.

Method	**PSNR (dB) ↑ across Generated Segments**								
	S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**
WorldMem	24.603	21.246	18.832	16.798	14.904	13.228	11.939	10.948	10.180
UCM	24.259	21.094	19.113	17.600	16.125	14.896	13.809	12.850	12.009
ReCamMaster	26.437	22.731	20.939	19.452	18.448	17.707	16.985	16.377	15.882
HyDRA	28.121	22.641	19.950	17.666	16.483	15.907	15.665	15.303	15.132
Self-Forcing	26.267	20.839	19.406	18.651	18.169	17.755	17.321	16.950	16.650
EonWeave	**32.596**	**28.182**	**26.189**	**24.952**	**24.086**	**23.437**	**22.928**	**22.484**	**22.074**

ties observed without ReCAP and RaDAR are consistent with two complementary requirements. Relevant scene evidence must remain available in the atlas and must subsequently be presented in a form compatible with the current query.

A cross-metric comparison further separates preservation from alignment. The ReCAP and RaDAR ablations obtain nearly identical SSIM values (0.655 and 0.654), yet the RaDAR ablation has a higher LPIPS (0.497 versus 0.484) and a lower Subject Consistency (0.949 versus 0.954). Conversely, the ReCAP ablation incurs the larger PSNR loss. This pattern suggests that evidence preservation has a stronger effect on pixel-level reconstruction, whereas structured historical readout is especially important for perceptual correspondence and subject preservation. Preserving historical content and aligning it with the current query therefore provide complementary rather than interchangeable benefits.

A different pattern emerges for relative degradation. The model without ARiSE achieves the lowest $\Delta_{\mathrm{Dft}}^{\mathrm{Qal}}$ of 0.312, even though its PSNR and SSIM remain substantially below those of the full model. This metric captures the relative quality change between the first and final generated chunks rather than the absolute quality of the trajectory. The chunk-wise statistics make this distinction explicit: without ARiSE, PSNR starts at 26.821 rather than 30.819 and ends at 18.450 rather than 20.339. The lower relative drop therefore describes a flatter decline along an already weaker trajectory, not a better final state. This observation underscores the need to interpret relative degradation together with absolute fidelity when evaluating long-horizon robustness.

Overall, the full model outperforms every component ablation in PSNR, SSIM, LPIPS, Subject Consistency, and Background Consistency. The distinct failure profiles support the joint use of reliability-aware conditioning, bounded evidence preservation, and query-aligned readout.

Under four-step distillation, EonWeave-DMD further improves PSNR by 1.651 dB and SSIM by 0.048 while reducing LPIPS by 0.045 relative to the full model. It also improves Subject and Background Consistency by 0.017 and 0.009, respectively. These results indicate that the complete EonWeave pipeline remains compatible with four-step distillation while retaining strong long-horizon fidelity and consistency.

G.3 Long-Horizon Performance Evolution

Autoregressive world generation recursively reuses predicted segments as future conditions, allowing local errors to propagate and progressively reshape subsequent generation. While $\Delta_{\mathrm{Dft}}^{\mathrm{Qal}}$ summarizes the overall degradation between the first and final generated chunks, it does not reveal how fidelity evolves throughout the rollout. We therefore examine segment-wise PSNR across the nine generated segments of Mira-World in Tab. 8 and Fig. 4, characterizing both sustained generation fidelity and the progression of degradation with autoregressive depth. For conciseness, **EonWeave** denotes EonWeave-DMD and **HyDRA** denotes its autoregressive extension, AR-HyDRA, throughout this analysis.

EonWeave exhibits a pronounced and persistent fidelity advantage throughout the entire autoregressive trajectory. At S1, it reaches 32.596 dB, exceeding the highest competing PSNR by 4.475 dB.

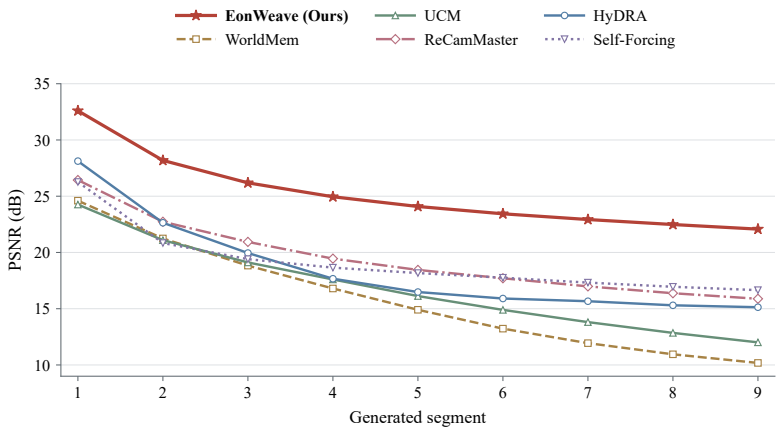


Figure 4: **PSNR evolution over nine generated segments on Mira-World.** EonWeave sustains substantially higher fidelity throughout the rollout and exhibits markedly slower degradation as the generation horizon extends.

More importantly, this separation does not diminish as generated segments are recursively reused as future conditions: from S2 through S9, EonWeave maintains a margin exceeding 5 dB at every generation step. By S9, it still achieves 22.074 dB, remaining 5.424 dB above the highest competing result. The advantage is therefore not confined to the initial prediction quality, but persists under repeated autoregressive feedback, indicating substantially stronger preservation of usable visual information as the rollout horizon increases.

The shape of this trajectory provides a complementary view of long-horizon stability. Let

$$d_s = P_s - P_{s+1}, \qquad s \in \{1, \dots, 8\}, \tag{91}$$

where P_s denotes the PSNR of generated segment s. For EonWeave, the successive decreases are

$$\{d_s\}_{s=1}^{8} = \{4.414,\ 1.993,\ 1.237,\ 0.866,\ 0.649,\ 0.509,\ 0.444,\ 0.410\}\ \text{dB}. \tag{92}$$

The incremental loss decreases monotonically across all eight transitions. Although the total S1–S9 decrease is 10.522 dB, 41.9% of this loss occurs in the first transition alone. The average decrease subsequently falls from 2.128 dB per transition over S1–S5 to only 0.503 dB over S5–S9, corresponding to a 76.4% reduction in the rate of additional degradation. Indeed, the entire four-transition loss from S5 to S9 is only 2.012 dB, less than half of the initial S1–S2 decrease. Thus, as autoregressive depth increases, EonWeave enters a markedly flatter degradation regime while retaining a substantially higher absolute fidelity level.

This distinction between *retained fidelity* and *magnitude of decline* is essential for interpreting long-horizon behavior. Several competing methods also exhibit progressively smaller segment-wise losses, but their trajectories flatten only after fidelity has already deteriorated to considerably lower levels. Self-Forcing provides a particularly illustrative example: its total S1–S9 decrease is 9.617 dB, numerically smaller than EonWeave's 10.522 dB, yet its final PSNR remains 5.424 dB lower. A smaller absolute decline therefore does not by itself imply stronger long-horizon generation. The more informative behavior exhibited by EonWeave is the combination of a substantially elevated quality trajectory and progressively diminishing incremental degradation. Absolute PSNR, $\Delta_{\text{Dft}}^{\text{Qal}}$, and the segment-wise trajectory consequently capture complementary aspects of long-horizon performance: the fidelity retained, the normalized endpoint drift, and the temporal structure through which that drift accumulates.

This behavior is consistent with the persistent historical conditioning introduced by EonWeave. As autoregressive generation proceeds, recursively generated short-term conditions become increasingly susceptible to accumulated prediction errors, while previously observed subjects and scene states may have already disappeared from the active context. The *trust–preserve–align* design provides an additional path through which such remote evidence remains usable: ARiSE calibrates reliance between progressively less reliable local conditions and persistent history, ReCAP preserves

complementary observations within a bounded Persistent Scene Atlas, and RaDAR converts the retained evidence into query-aligned references for the current generation step. Rather than relying exclusively on an increasingly self-generated context, EonWeave can therefore continue grounding generation in previously established world states. The sustained performance separation in Fig. 4 is consistent with this mechanism, while the controlled ablations separately establish the contributions of its individual components.

Overall, the segment-wise analysis reveals a stronger property than endpoint improvement alone: EonWeave maintains a substantial fidelity margin at every evaluated autoregressive depth while its additional quality loss progressively contracts. This trajectory complements the aggregate long-horizon results and provides direct evidence that the benefit of persistent historical modeling is sustained, rather than transient, across extended recursive generation. The corresponding perceptual behavior, including the preservation of returning subjects and previously established scene structure, is further examined in Sec. G.4.

G.4 Qualitative Analysis of Long-Horizon Revisitation

Figs. 5–8 provide additional long-horizon autoregressive comparisons on HM-World and MiraWorld, complementing the revisitation example in Fig. 3. For brevity, we use **HyDRA** as a unified shorthand for the original single-shot HyDRA and its autoregressive extension, AR-HyDRA. The first image column shows the observed context, while the remaining columns sample the subsequent rollout at approximately uniform temporal intervals. These comparisons are designed to assess a property that is not fully captured by the realism of an individual generated frame: whether a world state established much earlier remains recoverable after it leaves the active field of view and the model undergoes repeated autoregressive updates. A generated frame may remain locally plausible while becoming historically inconsistent, for example by introducing a different-looking instance of a returning subject or by rewriting scene structure that has already been observed. We therefore examine the results along three closely related axes: preservation of revisited subjects and objects, persistence of previously established scene structure, and resistance to progressive degradation over extended rollout.

The distinction is particularly clear in the HM-World example of Fig. 3. The observed context establishes both a distinctive foreground subject and a recognizable urban environment. After the subject leaves the field of view and subsequently re-enters, several competing methods gradually lose correspondence with this initial state. ReCamMaster develops severe geometric deformation in the surrounding scene and eventually fails to recover the subject; Self-Forcing similarly loses the returning subject in later frames; WorldMem undergoes pronounced large-scale structural degradation; and HyDRA, although retaining a recognizable subject during part of the revisitation interval, no longer preserves it at the end of the displayed trajectory. UCM retains a giraffe-like foreground, but its appearance, proportions, and surrounding scene deviate visibly from the state established in the initial context. EonWeave, by contrast, preserves a recognizable subject across Frames 103, 117, and 130, including substantially more stable body proportions and appearance patterns, while simultaneously maintaining the coarse organization of the red-brick street environment. The enlarged subject crops make this distinction particularly visible: the advantage lies not merely in generating a plausible foreground category, but in recovering an instance that remains visually correspondent to the subject observed before the prolonged absence.

This observation highlights a characteristic failure mode of long-horizon generation that is easily overlooked when frames are judged independently. Once an entity disappears from the current view, short-term autoregressive conditioning no longer contains direct evidence of its established appearance. A later re-entry can therefore remain visually realistic while being incorrect with respect to the generated world's own history. Across the additional HM-World sequences in Fig. 5, Fig. 6, and Fig. 7(a), competing approaches exhibit this failure through identity drift, missing or replaced foreground content, duplicated objects, and changes to persistent background organization. Such inconsistencies become increasingly apparent when later frames are compared against the initial observation rather than evaluated in isolation. EonWeave maintains substantially stronger continuity across these revisitation events, preserving previously established subjects and their relationship to the surrounding scene after they have remained outside the active context for extended intervals. These results therefore reinforce the distinction underlying *revisitation consistency*: long-horizon

world generation requires not only smooth transitions between adjacent observations, but also correspondence to world states established across large temporal gaps.

A second class of failure emerges progressively as autoregressive rollout continues. Because every predicted segment is recursively reused as future context, deviations introduced at one stage can persist and interact with newly introduced errors at later steps. In the qualitative comparisons, this process manifests as gradual foreground degradation, changes in scene appearance or illumination, drift in background geometry, and, in more severe cases, large-scale structural or stripe-like artifacts. The effect is particularly visible when early and late rollout samples are compared: several baselines begin from locally plausible states yet progressively lose recognizable scene organization as generation proceeds. EonWeave exhibits substantially less severe deterioration, retaining coherent foregrounds and recognizable scene layouts further into the trajectory. This observation complements the segment-wise analysis in Fig. 4: the quantitative trajectory characterizes how reconstruction quality evolves with rollout length, whereas the qualitative examples expose the perceptual form of this degradation, including the disappearance of established entities, changes in their identity or appearance, and progressive rewriting of the surrounding world.

The longer and more diverse Mira-World sequences provide a complementary test in which persistent information must survive through substantially more autoregressive transitions and less constrained scene evolution. In Fig. 7(b) and Fig. 8, foreground configuration, viewpoint, illumination, and visible scene content continue to change throughout the rollout. As the sequences approach their later portions, competing methods tend to exhibit increasingly pronounced changes in recognizable foregrounds or the surrounding spatial organization. EonWeave instead maintains more coherent subjects and recognizable scene layouts throughout the extended trajectory, including cases in which the current view differs substantially from the observation in which the relevant world state was originally established. This property is particularly important for persistent memory: successful long-horizon retrieval cannot reduce to reproducing a stored observation under a nearly identical viewpoint. Historical evidence must remain useful when the current query has changed both spatially and visually, motivating a readout mechanism that adapts retained history to the current generation state.

Importantly, stronger historical correspondence does not arise from freezing the generated scene. Across both benchmarks, subjects continue to move, camera viewpoints change, and newly visible regions evolve while previously established identities and scene structure remain comparatively stable. Persistent consistency therefore coexists with short-term dynamics rather than replacing them. This distinction separates revisitation consistency from temporal copying: the objective is not to reproduce an earlier frame, but to preserve those properties of the world that should persist while allowing its current state to continue evolving. The qualitative results consequently indicate that EonWeave can use remote history as a constraint on the present generation without forcing the current state to collapse back to a past observation.

The observed failure modes also clarify why the three components of EonWeave address complementary aspects of long-horizon historical modeling. When recursively generated conditions become progressively less reliable, ARiSE reduces the tendency to treat them as uniformly trustworthy local evidence. When useful content has moved beyond the active context window, ReCAP preserves complementary historical states under a bounded memory budget instead of allowing recent observations to overwrite persistent evidence. Finally, because a retained observation may differ substantially from the current query in viewpoint and spatial organization, RaDAR converts the available history into query-aligned references rather than exposing heterogeneous historical records directly to attention. The qualitative comparisons alone do not isolate the causal contribution of any individual component; this role is served by the controlled ablations in the main paper. Taken together, however, the qualitative evidence and ablation results provide a coherent perceptual account of EonWeave's *trust–preserve–align* pipeline: previously established evidence is less likely to be forgotten, remains accessible after prolonged absence, and can be incorporated in a form compatible with the current generation state.

G.5 Hyperparameter Sensitivity Analysis

We further examine the sensitivity of EonWeave to two key hyperparameters: the Persistent Scene Atlas capacity K and the ARiSE re-noising strength $\sigma^\star$. These parameters respectively govern the

Table 9: **Hyperparameter sensitivity analysis on Mira-World.** We investigate two key factors in EonWeave: (1) the persistent memory capacity controlled by the Persistent Scene Atlas size K, which determines the amount of scene evidence retained for long-horizon generation; and (2) the inference-time reliability alignment strength controlled by the ARiSE re-noising magnitude $\sigma^\star$, which adjusts the perturbation applied to autoregressive latent conditions. For each experiment, only the corresponding hyperparameter is varied while keeping the remaining configurations fixed. All metrics are reported with four decimal places to better distinguish subtle performance variations among different settings. Best and second-best results within each group are highlighted accordingly.

Setting	PSNR ↑	SSIM ↑	LPIPS ↓	Subj. Cons. ↑	Backg. Cons. ↑	$\Delta_{\text{Dft}}^{\text{Qal}}$ ↓
Persistent Scene Atlas memory capacity K ($\sigma^\star = 0.15$)						
9	25.1127	0.7474	0.3950	0.9851	0.9850	0.3239
18	25.2223	0.7506	0.3940	0.9849	0.9852	0.3207
27	25.1418	0.7474	0.3954	0.9848	0.9851	0.3230
36	25.1928	0.7499	0.3961	0.9851	0.9851	0.3193
45	25.1937	0.7486	0.3952	0.9851	0.9853	0.3227
54	25.2027	0.7490	0.3944	0.9850	0.9851	0.3193
ARiSE re-noising strength $\sigma^\star$ ($K = 54$)						
0.05	24.9932	0.7375	0.3918	0.9853	0.9853	0.3287
0.10	25.1373	0.7448	0.3930	0.9851	0.9854	0.3242
0.15	25.2027	0.7490	0.3944	0.9850	0.9851	0.3193
0.20	25.1851	0.7499	0.3964	0.9852	0.9851	0.3218
0.25	25.1840	0.7486	0.3988	0.9853	0.9850	0.3220
0.30	25.1620	0.7461	0.3998	0.9854	0.9846	0.3233

amount of persistent scene evidence retained by ReCAP and the degree of reliability calibration applied to short-term autoregressive conditions by ARiSE. All experiments in this subsection are conducted with the EonWeave-DMD variant; for brevity, we refer to it simply as **EonWeave** throughout the following analysis. We vary one hyperparameter at a time while keeping the remaining settings fixed (the $\sigma^\star$ sweep is conducted at $K = 54$ and the K sweep at $\sigma^\star = 0.15$), thereby isolating its individual effect on long-horizon generation performance. The resulting sensitivity profiles on Mira-World are reported in Tab. 9.

Effect of the Persistent Scene Atlas Capacity. The capacity of the Persistent Scene Atlas determines the amount of persistent scene evidence that can be retained throughout the generation process. Increasing the atlas capacity allows EonWeave to preserve richer world states, which can potentially improve reconstruction fidelity and revisitation consistency. However, increasing the capacity beyond a sufficient level does not necessarily provide proportional gains, as additional records may contribute only limited complementary information under the evaluated setting.

To examine this trade-off, we vary the atlas capacity K while fixing the ARiSE re-noising strength to $\sigma^\star = 0.15$. As shown in Tab. 9, increasing K from 9 to 18 modestly improves reconstruction quality, with PSNR increasing from 25.113 dB to 25.222 dB and SSIM improving from 0.7474 to 0.7506. This improvement indicates that limited persistent memory capacity constrains the ability of EonWeave to preserve important scene states during long-horizon generation. However, further increasing K does not lead to consistent improvements across all metrics. For example, $K = 54$ achieves competitive reconstruction quality with 25.203 dB PSNR and obtains the best $\Delta_{\text{Dft}}^{\text{Qal}}$ score together with $K = 36$, but the improvement over $K = 18$ remains marginal. This observation suggests that once sufficient persistent information has been stored, additional atlas capacity provides diminishing returns.

Meanwhile, the subject and background consistency metrics remain highly stable across different capacity settings, indicating that EonWeave preserves reliable temporal coherence under varying persistent memory capacities. Considering the overall reconstruction fidelity and consistency perfor-

mance, $K = 18$ provides the most favorable operating point among the evaluated settings, achieving a balanced trade-off between persistent scene preservation and representation compactness.

Effect of the ARiSE Re-noising Strength. The ARiSE re-noising strength $\sigma^\star$ controls the degree of perturbation applied to the generated latent condition during inference-time reliability alignment. A moderate re-noising strength enables the model to better handle imperfect autoregressive conditions caused by accumulated prediction errors, whereas excessive perturbation may introduce unnecessary uncertainty and weaken the preservation of reliable latent information.

To evaluate the influence of $\sigma^\star$, we fix the Persistent Scene Atlas capacity to $K = 54$ and vary the re-noising strength. The results in Tab. 9 demonstrate a clear trade-off between robustness to condition errors and latent fidelity. When the re-noising strength is relatively small, accumulated errors in the generated condition are insufficiently attenuated, resulting in degraded reconstruction quality. As the re-noising strength increases, the reconstruction performance gradually improves and reaches the best PSNR at the default setting $\sigma^\star = 0.15$.

However, continuously increasing the re-noising strength does not provide further benefits. Beyond the optimal operating point, reconstruction quality gradually saturates or decreases, indicating that excessive latent corruption may weaken the useful information preserved in the autoregressive condition. Meanwhile, the subject and background consistency metrics remain stable across different settings, demonstrating that ARiSE maintains robust world evolution under varying reliability alignment strengths. Therefore, $\sigma^\star = 0.15$ provides an effective balance between suppressing accumulated condition errors and preserving accurate latent states.

Overall, the sensitivity analysis demonstrates that both the Persistent Scene Atlas capacity and the ARiSE re-noising strength require careful calibration. Rather than maximizing persistent memory capacity or re-noising strength independently, EonWeave achieves favorable performance by maintaining an appropriate balance between persistent scene evidence preservation and inference-time reliability alignment.

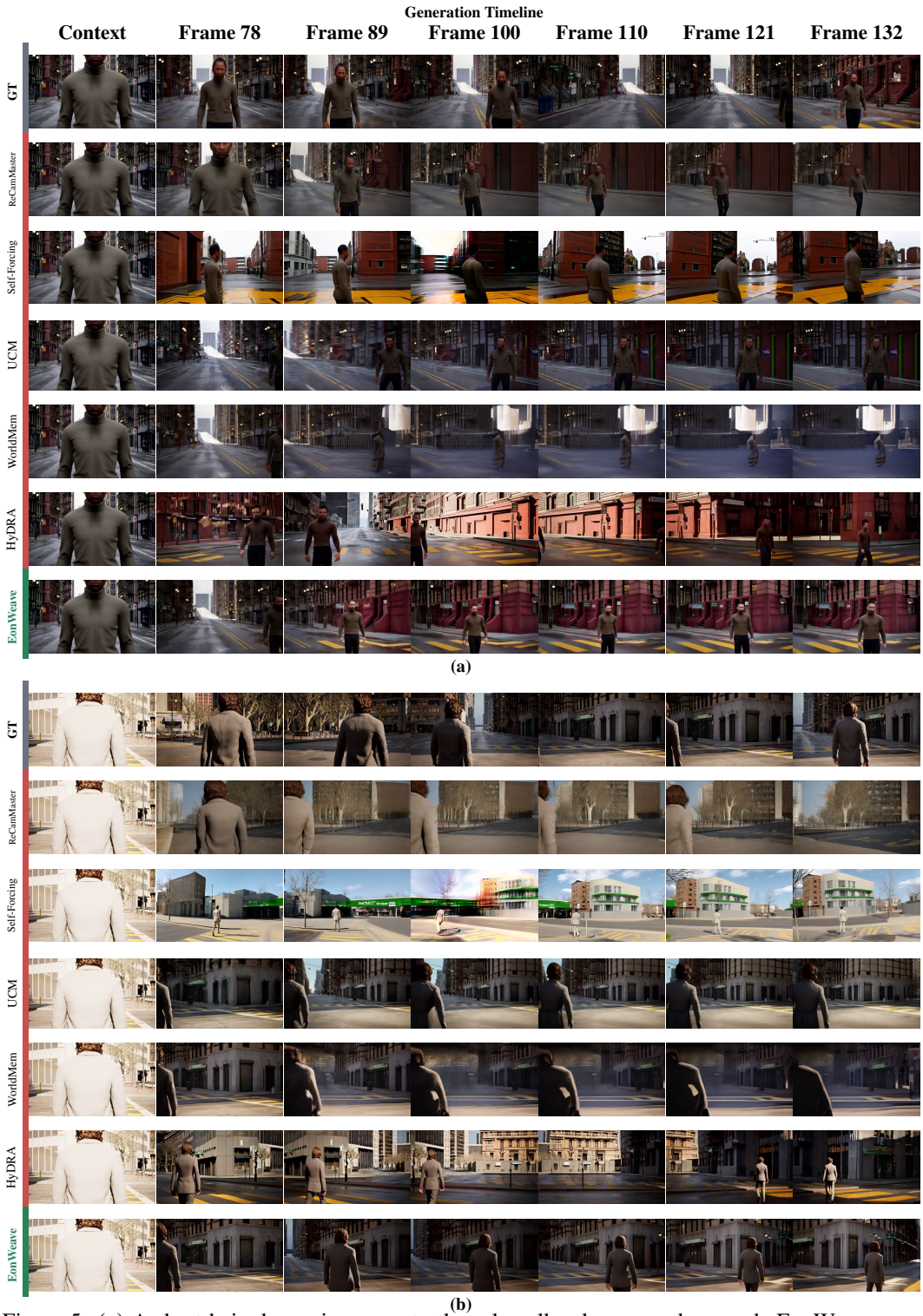


Figure 5: **(a)** A short-haired man in a gray turtleneck walks along an urban road. EonWeave preserves his facial identity, clothing, and body proportions while keeping the road geometry and surrounding urban environment coherent across viewpoint changes. **(b)** A brown-haired man in a light-colored jacket is viewed from behind at an urban intersection. EonWeave maintains his hairstyle, jacket silhouette, and identity together with the surrounding building geometry throughout the rollout.

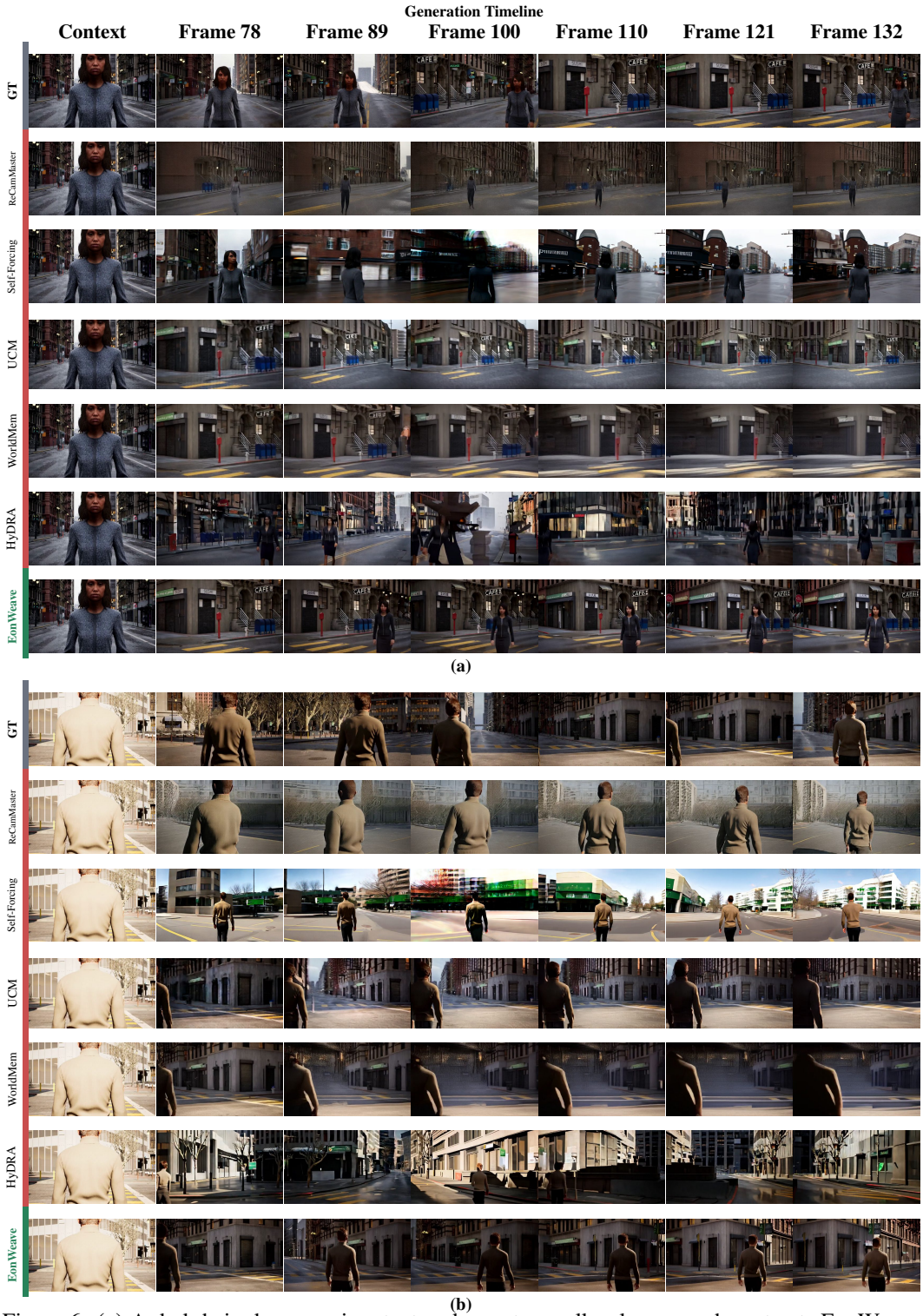
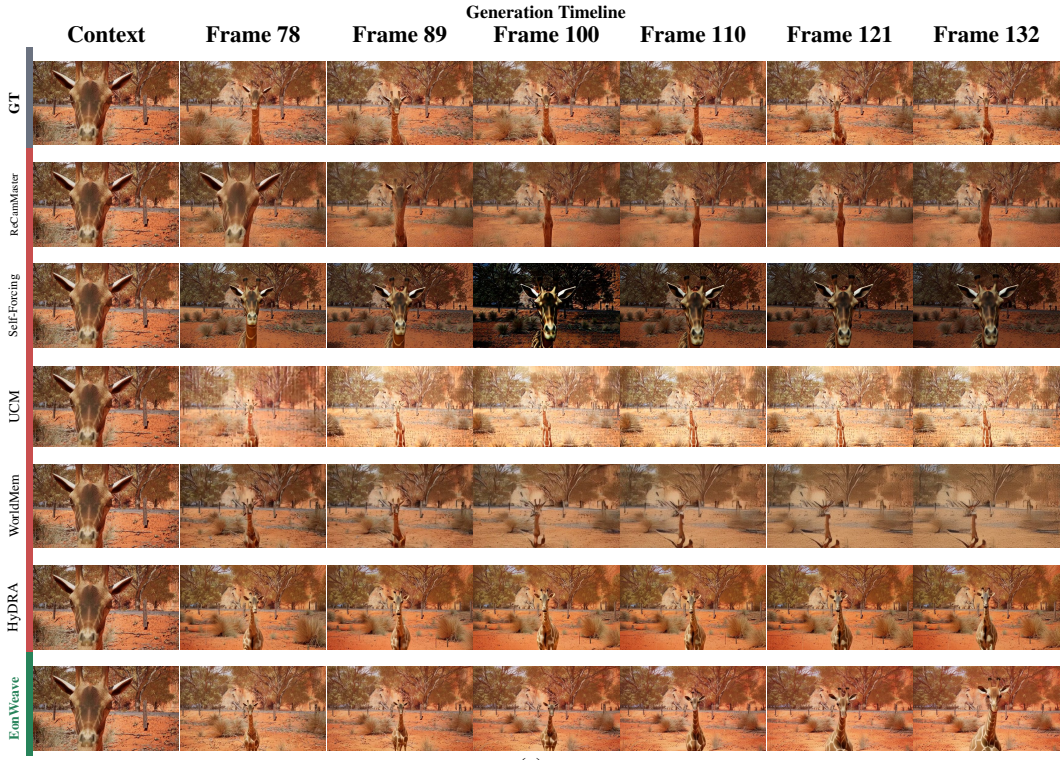
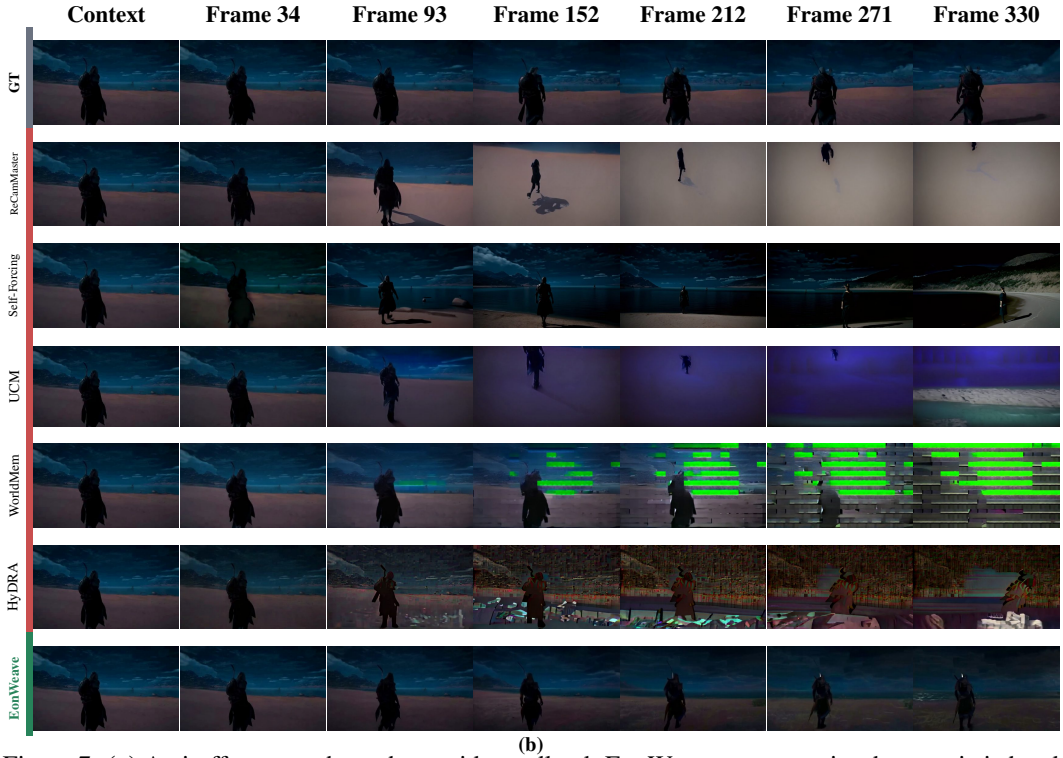


Figure 6: **(a)** A dark-haired woman in a textured gray top walks along an urban street. EonWeave preserves her facial features, hairstyle, clothing, and body proportions while keeping the street layout coherent across the rollout. **(b)** A brown-haired man in a beige sweater is followed from behind through a city intersection. EonWeave maintains his brown hair, beige sweater, and silhouette together with the surrounding intersection geometry across changing views.



(a)



(b)

Figure 7: **(a)** A giraffe moves through an arid woodland. EonWeave preserves its characteristic head shape, coat pattern, and neck proportions together with the visual continuity of the arid habitat. **(b)** A dark-clad archer carrying a bow crosses a nighttime shoreline. EonWeave maintains the figure's silhouette, costume, and carried equipment while preserving the dark shoreline appearance under changing poses and viewpoints.

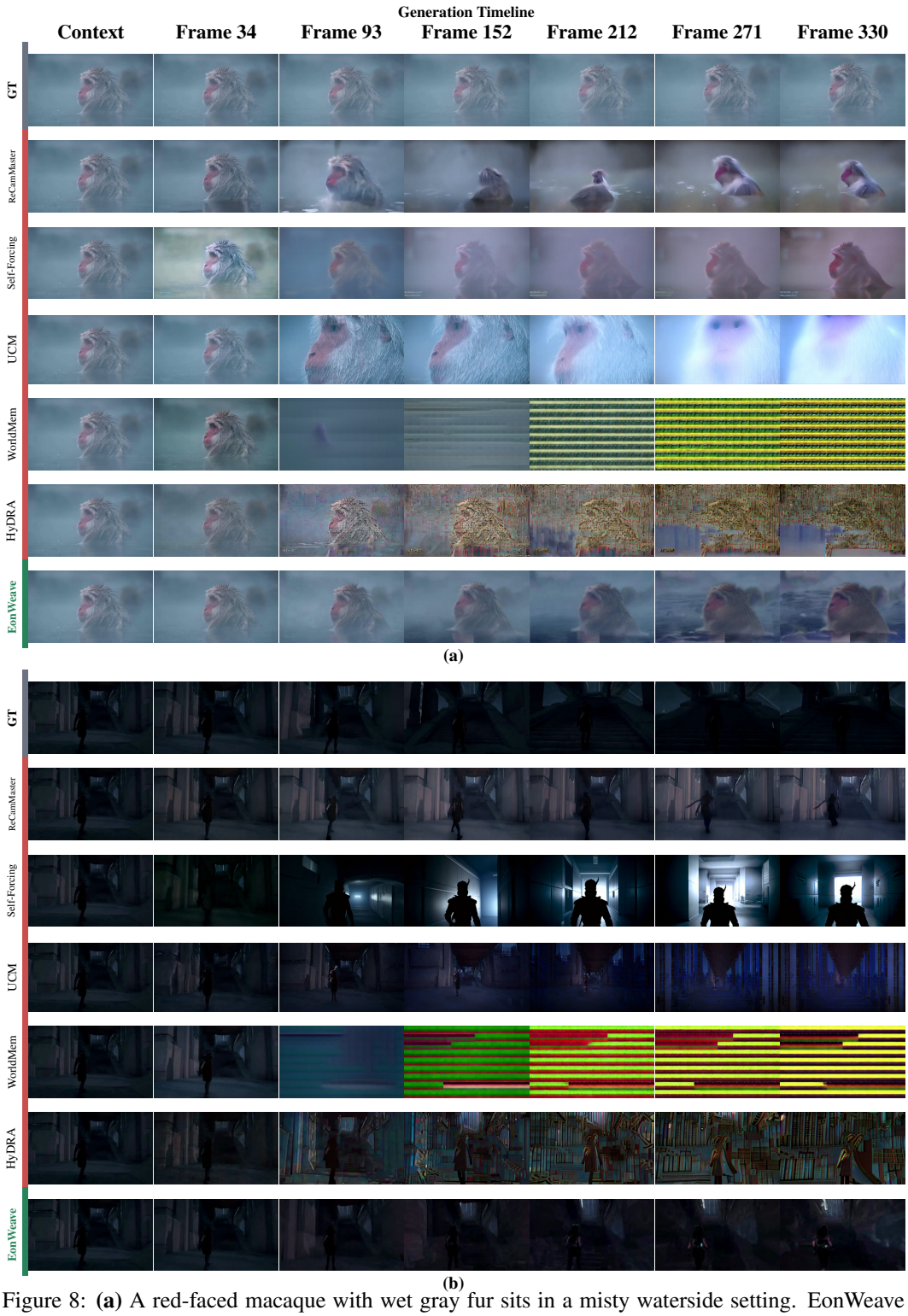


Figure 8: **(a)** A red-faced macaque with wet gray fur sits in a misty waterside setting. EonWeave preserves its facial structure, fur texture, and head profile together with the misty waterside background throughout the rollout. **(b)** A black-clad man with a sword strapped across his back ascends a staircase in a dim interior. EonWeave preserves his appearance and back-mounted sword while correctly rendering the changing body pose, perspective, and surrounding geometry as he climbs the stairs.