VELOX: SPATIAL MEMORY MAKES WORLD MODELS EFFICIENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Video world models turn actions and camera motion into long visual rollouts, making video a natural interface for interactive simulation. Such interaction requires responsive generation, yet coherent rollouts must retain visual memory of what has already been generated. Chunked video diffusion models usually keep this context in growing KV caches, so later chunks become progressively slower. To address this bottleneck, we introduce Velox, a training-free attention router for region-level context selection using frustum slab correspondences from streaming geometry. Velox represents historical blocks as camera frustum slabs with depth intervals and obtains current-block depth through recency-priority multiwarp. For well-covered regions, frustum slab overlap softly modulates each layer's model-derived block importance. Uncovered regions retain the model-derived route without geometric bias. The fixed token-block granularity matches FlashAttention-style tiled execution and reduces the attention term that grows with rollout length. Frustum slab construction runs asynchronously with denoising on the same H200. At chunk 48, Velox accelerates attention by $4.36\times$ and denoising by $2.43\times$. On the full 158-case WBench navigation split, it remains within 0.15 API-free average points of dense LingBot-Fast-Cam. Visual results are available at `https://anonymous.4open.science/w/velox/`, and all code and results will be made public.

1 INTRODUCTION

Video world models are emerging as a practical interface for interactive simulation. A user specifies an action trajectory or camera motion, and the model rolls the world forward by generating future video frames. This direction builds on rapid progress in video diffusion transformers and large video foundation models (Ho et al., 2020; Rombach et al., 2022; Peebles & Xie, 2023; Blattmann et al., 2023; Chen et al., 2024b; Yang et al., 2024b; Kong et al., 2024; Wan et al., 2025; Ma et al., 2025). Recent works increasingly target controllable long video worlds rather than isolated short clips (Alonso et al., 2024; Valevski et al., 2024; Bruce et al., 2024; Parker-Holder et al., 2024; Sand.ai, 2025; Gao et al., 2026; HY-World et al., 2026). For these models, context management is a central methodological challenge. Past frames carry the appearance, geometry, and object state that future frames must respect. Without access to the right history, long rollouts drift, forget revisited regions, or re-synthesize the same scene inconsistently.

The most direct way to preserve this history is to retain an ever-growing visual context. Chunked or causal video diffusion models let each new chunk attend to historical frame tokens through a KV cache. Dense FlashAttention makes this practical at moderate lengths (Dao et al., 2022; Dao, 2023), and causal diffusion or self-forcing training makes long rollouts natural (Chen et al., 2024a; Huang et al., 2025b). The cost, however, grows with the number of generated chunks. Later chunks inherit larger context even when the current camera overlaps only a small part of the past scene. As a result, chunk-wise generation becomes progressively slower exactly when long-horizon consistency matters most.

This motivates context selection rather than naive context accumulation. Frame- and view-memory methods select a smaller set of relevant past observations. Context-as-Memory retrieves historical views by camera field-of-view overlap, and VMem stores surfel-indexed views for consistent revisits (Yu et al., 2025a; Li et al., 2025). AnchorWeave and related works (Xiao et al., 2025; Wu

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.

et al., 2025; Wang et al., 2026) further inject warped renderings from persistent memory through ControlNet (Zhang et al., 2023a) to stabilize camera-controlled generation. These methods show that not all past frames are equally useful, and figure 1 summarizes this progression from dense history to view-level retrieval. Yet frame-level selection remains coarse for transformer attention. It provides the same selected context to all regions in the active chunk, although dynamic scenes often need local temporal continuity and different regions in the same noisy chunk can depend on different parts of the past. Methods that inject warped point renderings into the model input can also require architectural adaptation or fine-tuning, which limits direct reuse of a pretrained video generator.

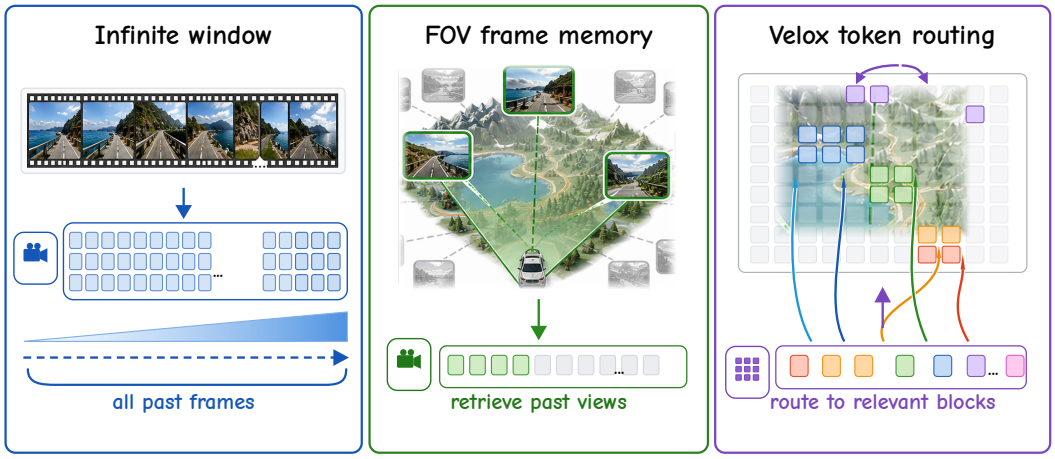


Figure 1: **Three context-management paradigms for long-horizon world generation.** Dense chunked generation keeps an ever-growing visual context. Frame-memory methods retrieve a small set of past views using camera overlap. Velox routes each current latent region to compact, geometry-relevant history blocks using frustum slab correspondences.

Inspired by inverse rendering works (Kerbl et al., 2023; Mildenhall et al., 2021), we observe that regions whose geometry has already been observed should primarily attend to their corresponding past regions. Newly revealed or uncertain regions, in contrast, may require model-derived semantic context and local temporal neighbors. A road block, a facade block, and newly revealed sky in the same noisy chunk should therefore not necessarily attend to the same past views. The question, then, is no longer only which frames to retrieve, but whether long-context video generation can assign each region its own relevant history while preserving local motion continuity and leaving the pretrained generator unchanged.

We consider this question at the token-block level in the block-wise execution model of FlashAttention (Dao, 2023). Our key insight is that camera frustum slabs with depth intervals can provide region-specific routing signals at the same granularity used by tiled attention. Instead of adding a new conditioning branch or fine-tuning the video generator, Velox summarizes generated history as per-block frustum slabs and uses their overlap to guide which historical KV token blocks each current region should see. For each target chunk, Velox obtains query-side depth through recency-priority multiwarp and scores query and historical blocks by world-space OBB frustum slab overlap. In well-covered regions, this score softly reweights each layer's model-derived block importance before selection. Uncovered regions retain the model-derived route without geometric bias. Previous-chunk retention and global-aligned Gilbert neighborhoods preserve local motion continuity without a separate cross-chunk seam table.

Across 48-chunk timing traces and the 158-case WBench navigation split (Ying et al., 2026), Velox reduces the growing long-context attention cost while preserving aggregate quality near dense attention. Its advantage grows with the number of generated chunks. At chunk 48, attention latency decreases from approximately 34.2 ms with dense FlashAttention to 7.8 ms, a $4.36\times$ speedup. Denoising time decreases from approximately 7.19 s to 2.96 s under the same single-H200 asynchronous setting. Built on LingBot-Fast-Cam (Gao et al., 2026), Velox remains within 0.15 points of dense attention on the WBench API-free average and improves the 19-metric overall average from 75.52 to 75.60. These results support the scoped claim that soft frustum slab routing can alleviate the late-chunk attention bottleneck without materially degrading benchmark quality.

Our contributions can be summarized as:

- We introduce Velox, a training-free framework that uses frustum slab correspondences to reduce growing KV attention cost in chunked video world models.
- We devise a soft geometric fusion rule that uses OBB frustum slab overlap to modulate each layer's model-derived block importance while preserving an unbiased fallback for uncovered regions.
- We implement Velox with global-aligned Gilbert KV storage, an optimized sparse attention kernel, and asynchronous frustum slab construction on the same H200 as denoising, reaching a $4.36\times$ attention speedup and a $2.43\times$ denoising speedup at chunk 48.

2 Related Work

2.1 Video World Models and Spatial Memory

Video generation has progressed from latent diffusion to scalable diffusion transformers and rectified flow (Ho et al., 2020; Rombach et al., 2022; Blattmann et al., 2023; Peebles & Xie, 2023; Esser et al., 2024), with foundation models supplying strong visual priors (Yang et al., 2024b; Kong et al., 2024; Wan et al., 2025; HaCohen et al., 2024; Ma et al., 2025; Gao et al., 2025; Team et al., 2025). Long-video methods extend these priors through progressive generation, self-rollout training, and causal chunking (Kim et al., 2024; Jin et al., 2024; Henschel et al., 2024; Xie et al., 2024; Gu et al., 2025; Chen et al., 2024a; Huang et al., 2025b; Sand.ai, 2025; Chen et al., 2025a; Liu et al., 2025a), while interactive models target controllable persistent worlds (Bruce et al., 2024; Parker-Holder et al., 2024; Alonso et al., 2024; Valevski et al., 2024; Feng et al., 2024; Zhang et al., 2025d; Che et al., 2024; Yang et al., 2024a; Yu et al., 2025b; Decart et al., 2024; Gao et al., 2026; HY-World et al., 2026). These directions make history management central: more context supports consistency but makes dense attention increasingly expensive. Spatial-memory methods address this tension by retrieving views, preserving revisits, or expanding reconstructed worlds (Yu et al., 2025a; Li et al., 2025; Xiao et al., 2025; Wu et al., 2025; Huang et al., 2025a; Chen et al., 2025b; Yu et al., 2024b;a; Engstler et al., 2024). Camera-aware generation (Yu et al., 2024c; YU et al., 2025; Bai et al., 2025; He et al., 2024; 2025; Yang et al., 2025c; Ren et al., 2025) and streaming reconstruction (Wang et al., 2024b; Leroy et al., 2024; Cabon et al., 2025; Yang et al., 2025a; Wang et al., 2024a; 2025a;b; Keetha et al., 2025) further make projectable memory practical. Velox uses this memory to route pretrained attention rather than adding a conditioning pathway or retrained memory module.

2.2 Efficient Attention and Context Routing

Long-context transformers reduce attention cost through sparsity, locality, compression, state reuse, heavy-hitter selection, and KV eviction (Kitaev et al., 2020; Zaheer et al., 2020; Beltagy et al., 2020; Wang et al., 2020; Dai et al., 2019; Liu et al., 2023; Lee et al., 2025; Jiang et al., 2024; Li et al., 2024; Zhang et al., 2023b; Kwon et al., 2023), while FlashAttention is the dense IO-aware baseline (Dao et al., 2022; Dao, 2023). Video accelerators additionally exploit spatial, temporal, denoising-step, and feature regularity through adaptive sparsity, caching, and quantized attention (Zhang et al., 2025c; Xi et al., 2025; Yang et al., 2025b; Zhang et al., 2025b; Xia et al., 2025; Ding et al., 2025; Zhang et al., 2024; 2025a; Liu et al., 2024b; 2025b; 2024a; Guo et al., 2024; Nawaz et al., 2026; Feng et al., 2026). These approaches usually derive sparsity from position, cached features, or attention statistics, whereas frame-memory methods select views but retain dense attention within them. Velox connects both directions by using projective correspondence as a soft prior for region-specific attention routes, with model-derived fallback and temporal safeguards.

3 Method

Overview. Long-horizon generation requires persistent visual history, yet dense KV attention makes every current region process all retained blocks. Velox keeps the pretrained generator and conditioning interface unchanged and modifies only the inference-time visibility relation between current query blocks and historical KV blocks. It represents history as camera frustum slabs with depth intervals, uses their overlap as a soft prior over each layer's model-derived block importance, and exposes each current region to a compact route. Figure 2 summarizes this training-free pipeline.

3.1 PRELIMINARIES: CHUNKED WORLD GENERATION

Consider a video world model that generates chunks $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots$ autoregressively. At denoising step s, chunk c has noisy latent tokens $\boldsymbol{z}_c \in \mathbb{R}^{T\times H\times W\times d}$, text/action controls (y_c, a_c), and a camera sequence $\pi_c = \{\pi_{c,t}\}_{t=1}^{T}$, where T is the number of latent frames, $H \times W$ is the latent spatial grid, d is the channel dimension, and each $\pi_{c,t}$ maps 3D scene locations to the corresponding latent/image view. A pretrained diffusion transformer with parameters θ predicts

$$\hat{\boldsymbol{\epsilon}}_c = f_\theta(\boldsymbol{z}_c, s, y_c, a_c, \pi_c, \mathcal{C}_{<c}), \tag{1}$$

where $\mathcal{C}_{<c}$ is the retained historical context, implemented as a visual KV cache in the transformer. If $\mathcal{C}_{<c}$ contains all previous chunks, the attention cost grows with history length even when only a small part of the past scene is visible from the current camera.

We operate at block granularity. Let σ be the global-aligned Gilbert order, and let $\mathcal{B}_c = \{B_i\}_{i=1}^{N}$ partition the current chunk into blocks of b_{tok} tokens. Historical KV blocks use the same size, and M_c denotes their visible count before chunk c. For fixed chunk c and layer ℓ, write Q_i for current block B_i and (K_r, V_r) for visible KV block $r \in \{1, \ldots, M_c\}$.

This block size matches FlashAttention-style execution (Dao et al., 2022; Dao, 2023), which streams fixed query and KV tiles. Abstractly, history attention decomposes into

$$\sum_{i=1}^{N}\sum_{r=1}^{M_c} \text{TileAttn}(Q_i, K_r, V_r), \tag{2}$$

where each interaction computes $Q_i K_r^\top$, updates online softmax, and applies V_r. Dense FlashAttention still schedules every current-history pair. A fixed route removes entire KV tiles before launch while preserving block-contiguous gathers. Token-level or variable-size sparsity would require irregular row-wise indexing.

The active attention window is ordered by global-aligned Gilbert rank. Dense visual attention for block i uses every visible block:

$$\text{Attn}_{\text{dense}}(Q_i) = \text{softmax}\left(Q_i K_{1:M_c}^\top/\sqrt{d}\right) V_{1:M_c}. \tag{3}$$

Velox replaces $1{:}M_c$ by a route $\mathcal{R}_i^\ell \subseteq \{1, \ldots, M_c\}$ at each transformer layer ℓ. The layer index matters because every layer produces its own pooled query-key importance distribution, while the geometric correspondence is shared across layers within a denoising step.

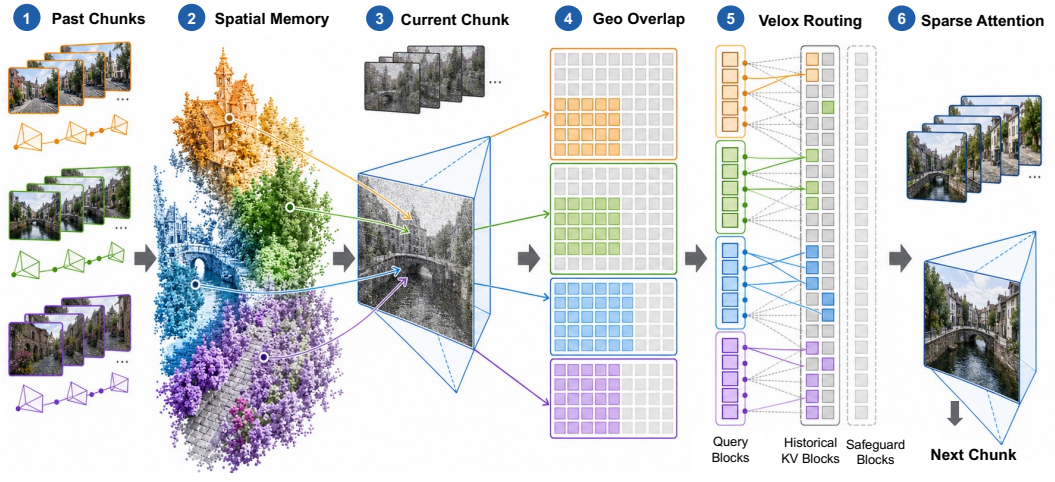


Figure 2: **Velox overview.** Velox leaves the pretrained chunked video generator unchanged and replaces dense historical attention with frustum slab routed attention. Historical chunks are summarized by per-block frustum slabs with depth intervals, target query blocks obtain depth through multiwarp, and block routes softly fuse frustum slab overlap with layer-specific model importance before adding temporal and global Gilbert safeguards. Indexed sparse attention then produces the next chunk.

3.2 FRUSTUM SLAB CORRESPONDENCE

Velox does not route by brittle reconstructed sample identity. Instead, it stores each historical KV block as one or more *frustum slab primitives*, each comprising a source-camera token rectangle and robust depth interval:

$$\psi_{j,r,m} = \big(r,\ t,\ \mathbf{R}_{j,r,m},\ [z^{-}_{j,r,m}, z^{+}_{j,r,m}],\ \nu_{j,r,m}\big)\,, \tag{4}$$

Here j, r, m, and t index the source chunk, global Gilbert block, primitive, and latent frame. The source-camera token rectangle is $\mathbf{R}_{j,r,m}$, its percentile depth interval is $[z^{-}_{j,r,m}, z^{+}_{j,r,m}]$, and $\nu_{j,r,m}$ stores validity statistics. Blocks crossing frame or chunk boundaries have multiple primitives, aggregated by source chunk before scoring.

Historical frustum slab construction. After chunk j is decoded, the reconstruction backend produces depth, confidence, masks, and camera poses for the generated frames. Velox transiently unprojects reliable depth samples into 3D, projects them into the chunk's latent cameras, buckets them by global Gilbert block, and computes a depth quantile ladder for each block fragment. Let $\mathrm{Quantile}_p(\cdot)$ denote the empirical p-th percentile over valid source-camera depths, and let p_{lo} and p_{hi} be fixed low and high percentile levels. The frustum slab interval is:

$$z^{-}_{j,r,m} = \mathrm{Quantile}_{p_{\mathrm{lo}}}(\mathcal{Z}_{j,r,m})\,, \qquad z^{+}_{j,r,m} = \mathrm{Quantile}_{p_{\mathrm{hi}}}(\mathcal{Z}_{j,r,m})\,, \tag{5}$$

where $\mathcal{Z}_{j,r,m}$ is the set of valid source-camera depths collected for that primitive. Validity gates require enough samples, enough covered tokens, bounded relative depth spread, and a minimum frustum slab thickness. This yields a compact frustum slab table for stored history without carrying dense reconstructed geometry in the routing score.

Target-side multiwarp depth. To score a current query block, Velox also needs a depth interval in the target camera. The current implementation obtains this Q-depth with recency-priority multiwarp. It caches the first and last pixel-frame depths of recent reconstructed chunks, then forward-warps all cached frames into the target chunk's latent cameras. For each target token u, let f index cached source frames, let $n_f(u)$ be the number of warped depth samples that land on u, and let k_{warp} be the sample floor. The newest cached frame with at least k_{warp} samples wins:

$$f^{\star}(u) = \max\{f \mid n_f(u) \geq k_{\mathrm{warp}}\}, \qquad z_Q(u) = \mathrm{Quantile}_{50}\big(\mathcal{Z}_{f^{\star}(u),u}\big)\,. \tag{6}$$

Here $\mathcal{Z}_{f,u}$ is the depth sample set warped from cached frame f to target token u, and $\mathrm{Quantile}_{50}$ is the median depth. If no cached frame reaches the sample floor, $f^{\star}(u)$ is undefined, the token has no Q-depth, and the row can fall back to model-derived routing. Intuitively, this rule favors the most recent reliable geometry, while still allowing revisits to be filled by older observations when the current camera returns to a previously seen surface. Appendix B.3 formalizes source filtering, target-camera reprojection, winner selection, and query-block aggregation.

World-space frustum slab overlap. Given a query block i with multiwarp depth and the historical primitives attached to KV block r, Velox approximates each camera frustum slab by a camera-facing oriented bounding box. Let $\Psi_r = \{\psi_{j,r,m}\}$ be the primitive set attached to historical block r, and let $\mathcal{O}^{Q}_{i}$ and $\mathcal{O}^{K}_{j,r,m}$ denote the query and historical OBBs in world coordinates. The geometric score is:

$$g_{i,r} = \max_{\psi_{j,r,m} \in \Psi_r} \frac{\mathrm{Vol}\Big(\mathcal{O}^{Q}_{i} \cap \mathcal{O}^{K}_{j,r,m}\Big)}{\mathrm{Vol}\Big(\mathcal{O}^{Q}_{i}\Big) + \mathrm{Vol}\big(\mathcal{O}^{K}_{j,r,m}\big) - \mathrm{Vol}\Big(\mathcal{O}^{Q}_{i} \cap \mathcal{O}^{K}_{j,r,m}\Big) + \epsilon}. \tag{7}$$

Here $\epsilon > 0$ prevents division by zero. The implementation estimates the intersection with a deterministic OBB grid after a separating-axis test, then applies grouped fragment aggregation for multi-fragment KV columns. This score is view-aware because both the target and historical blocks carry camera orientation, token footprint, and depth extent. It also avoids tying correspondence to exact reconstructed sample identities, which are brittle under generated-video artifacts.

Coverage. The router separates geometric score from geometric trust. A token is covered when its multiwarp depth lies inside at least one valid historical frustum slab that contains the token ray within a spatial and depth margin. The block coverage is:

$$\rho_i = \frac{1}{|B_i|} \sum_{u \in B_i} \mathbf{1}\big[\exists r,\ \exists \psi_{j,r,m} \in \Psi_r :\ u \in \Pi_c(\psi_{j,r,m}) \wedge z_Q(u) \in [z^{-}_{j,r,m}, z^{+}_{j,r,m}]\big]\,, \tag{8}$$

where $\Pi_c(\psi_{j,r,m})$ denotes the projection footprint of the frustum slab primitive in the current target view, and η_{cov} is the coverage threshold. A block is covered when $\rho_i \geq \eta_{\text{cov}}$. Thus $g_{i,r}$ identifies which historical KV blocks overlap the current frustum, while ρ_i decides whether geometry is reliable enough to drive the route.

3.3 Velox Attention from Frustum Slab Correspondence

The geometric score should guide a route without imposing the same hard block set on every layer. Reconstruction can leave Q-depth holes, stale warped depths can survive disocclusions, and different transformer layers can favor different visual evidence. A hard geometric top-K route would override these layer-specific preferences. Velox instead uses frustum slab overlap to modulate the model's own block importance.

Layer-specific model importance. Let $\bar{q}_i^{\ell}$ be the mean query vector in block i at layer ℓ, and let $\bar{k}_r^{\ell}$ be the mean key vector in visible KV block r. We compute:

$$p_{i,r}^{\ell} = \text{softmax}_r \left(\bar{q}_i^{\ell} \left(\bar{k}_r^{\ell} \right)^{\top} / \sqrt{d} \right), \tag{9}$$

where the softmax is normalized over all visible KV blocks. This distribution preserves the pretrained transformer's layer-specific notion of useful context.

Soft geometric fusion. For a well-covered query block with $\rho_i \geq \eta_{\text{cov}}$, we first suppress overlap values at or below the noise threshold τ and rescale the remaining values:

$$\widehat{g}_{i,r} = \mathbf{1}[g_{i,r} > \tau] \min\left(\frac{g_{i,r}}{g_{\text{ref}}}, 1 \right), \tag{10}$$

where g_{ref} maps the observed frustum slab IoU range to a bounded boost. We then multiply the model importance by this geometric prior and renormalize:

$$\widetilde{p}_{i,r}^{\ell} = \frac{p_{i,r}^{\ell} \left(1 + \lambda \widehat{g}_{i,r}\right)}{\sum_{r'=1}^{M_c} p_{i,r'}^{\ell} \left(1 + \lambda \widehat{g}_{i,r'}\right)}. \tag{11}$$

Here $\lambda > 0$ controls the fusion strength. A block with $g_{i,r} \geq g_{\text{ref}}$ receives the maximum multiplier $1 + \lambda$, while a block without vetted geometric overlap retains its original model importance. Because the multiplication acts on $p_{i,r}^{\ell}$, geometry can promote a model-supported correspondence without assigning every layer an identical route.

Covered and uncovered routes. Let $\mathcal{P}_{c-1}$ denote the previous chunk's KV blocks. The route materializer reserves these blocks before selecting older history, so local temporal evidence consumes the same bounded budget as the long-range route. For well-covered blocks, Velox fills the remaining $K_{\text{cov}} - |\mathcal{P}_{c-1}|$ positions with the largest fused scores:

$$\mathcal{R}_i^{\ell} = \mathcal{P}_{c-1} \cup \text{Top}_{K_{\text{cov}} - |\mathcal{P}_{c-1}|} \left(\{ \widetilde{p}_{i,r}^{\ell} : r \notin \mathcal{P}_{c-1} \} \right), \qquad \rho_i \geq \eta_{\text{cov}}. \tag{12}$$

For blocks with insufficient geometric coverage, Velox leaves the model distribution unchanged. It sorts $\{p_{i,r}^{\ell}\}$ and takes the smallest prefix whose cumulative mass reaches η_{imp}, subject to a route cap K_{unc}:

$$\mathcal{R}_i^{\ell} = \mathcal{P}_{c-1} \cup \text{Prefix}_{\eta_{\text{imp}}, K_{\text{unc}} - |\mathcal{P}_{c-1}|} \left(\{ p_{i,r}^{\ell} : r \notin \mathcal{P}_{c-1} \} \right), \qquad \rho_i < \eta_{\text{cov}}. \tag{13}$$

This branch lets the pretrained attention signal handle sky, newly exposed regions, and holes in the frustum slab table. The calibrated configuration uses $\tau = 0.1$, $g_{\text{ref}} = 0.2$, $\lambda = 2$, $\eta_{\text{cov}} = 0.9$, $K_{\text{cov}} = 80$, and $K_{\text{unc}} = 150$.

Global-aligned Gilbert neighborhoods. Velox preserves short-range video continuity through a global/window 26-neighbor adjacency over Gilbert blocks. Earlier versions used a repeated chunk-local Gilbert order and a separate cross-chunk seam table. The current implementation instead builds a single global-aligned Gilbert block domain incrementally. It constructs the single-chunk Gilbert curve once, flips chunk-local coordinates according to chunk parity and the base-frame exit quadrant, and thereby matches the full-sequence Gilbert prefix for every generated chunk. Cross-chunk boundary neighbors become ordinary temporal neighbors in the visible global window. No separate seam table is needed.

Let $\mathcal{L}_i^{\mathrm{glob}}$ denote the visible 26-neighborhood of query block i in this global Gilbert block domain, and let $\mathcal{F}_0$ denote optional reference-frame blocks. The materialized route is:

$$\overline{\mathcal{R}}_i^{\ell} = \mathcal{R}_i^{\ell} \cup \mathcal{L}_i^{\mathrm{glob}} \cup \mathcal{F}_0. \tag{14}$$

The spatial neighbors and optional reference blocks are added as deterministic safeguards outside the route budget. The sparse visual-attention output is

$$\mathrm{Attn}_{Velox}(Q_i) = \mathrm{softmax}\Big(Q_i K_{\overline{\mathcal{R}}_i^{\ell}}^{\top}/\sqrt{d}\Big)\, V_{\overline{\mathcal{R}}_i^{\ell}}, \tag{15}$$

at every transformer layer. With this formulation, Velox preserves each layer's model-derived ranking, injects geometry only where correspondence is reliable, and still exposes the sparse kernel to a bounded block list.

Gilbert KV layout. Velox stores current and historical visual KV blocks in the same global-aligned Gilbert order used by the route. This keeps neighboring image regions close in memory and reduces per-call permutation overhead. The layout is an implementation optimization. The soft route in Eqs. (11) and (15) is independent of the particular space-filling curve, but global-aligned Gilbert makes local neighborhood preservation and block-contiguous KV storage share one coordinate frame.

3.4 Complexity and Design Rationale

The benefit of Velox comes from separating two quantities that dense attention couples together: the amount of visual memory retained by the world model and the number of KV blocks exposed to each query block. With the fixed block size b_{tok}, dense attention between N current blocks and M_c visible KV blocks has cost

$$\mathcal{O}\big(NM_c b_{\mathrm{tok}}^2 d\big)\,. \tag{16}$$

As more chunks are retained, M_c grows even when the current camera needs only a small part of the past. Velox keeps the memory available but caps the exposed route. Let $K_{\max} = \max(K_{\mathrm{cov}}, K_{\mathrm{unc}})$ and let $\Delta_{\mathrm{safe}} = |\mathcal{L}_i^{\mathrm{glob}} \cup \mathcal{F}_0|$ denote the small add-on safeguard budget:

$$\mathcal{O}\left(\sum_{i=1}^{N} |\overline{\mathcal{R}}_i^{\ell}| b_{\mathrm{tok}}^2 d\right) \approx \mathcal{O}\big(N(K_{\max} + \Delta_{\mathrm{safe}}) b_{\mathrm{tok}}^2 d\big) \quad \text{per sparse attention layer.} \tag{17}$$

The approximation highlights the intended regime: the route and safeguard budgets are fixed, while M_c grows with history length.

The same block unit also matches FlashAttention-style execution. Dense FlashAttention evaluates each query-key tile efficiently, but it still schedules all NM_c tile pairs. Velox changes the tile schedule to $\sum_i |\overline{\mathcal{R}}_i^{\ell}|$. A skipped block removes both the $QK^{\top}$ tile product and the KV tile load. Thus, routing reduces KV traffic when attention is memory-bound, and it reduces tile matrix products when attention is compute-bound. The calibrated route applies at every transformer layer, so the growing historical term is bounded throughout the denoising network.

4 Experiments

4.1 Implementation Details

We instantiate Velox on the LingBot-Fast-Cam inference stack (Gao et al., 2026) with a Wan-style video diffusion transformer. The reported configuration uses recency-priority multiwarp Q-depth, world-space OBB frustum slab IoU, grouped fragment aggregation, and global-aligned Gilbert KV storage. We apply fixed soft geometric fusion to well-covered rows at every transformer layer, with $\tau = 0.1$, $g_{\mathrm{ref}} = 0.2$, $\lambda = 2$, and $\eta_{\mathrm{cov}} = 0.9$. Covered rows retain $K_{\mathrm{cov}} = 80$ blocks. Uncovered rows use the model-derived cumulative-mass route with $K_{\mathrm{unc}} = 150$ and no route-size floor.

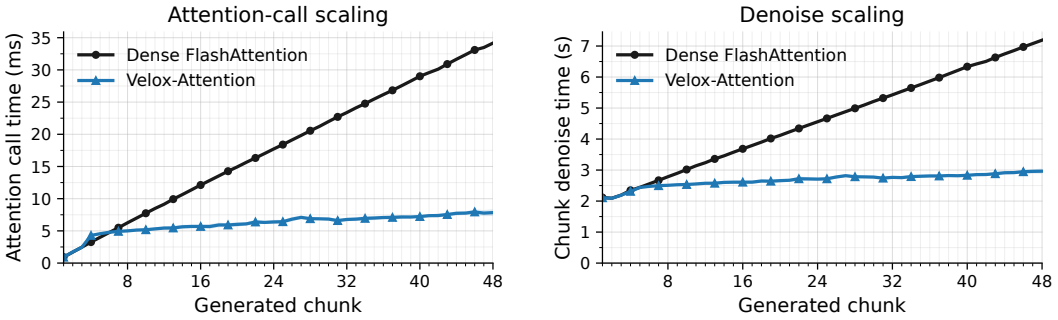
The measured sparse path is a native Triton indexed block-sparse attention implementation rather than a dense call with a Python mask. Current queries are reordered once per chunk, KV caches are stored in global-aligned Gilbert order, and the kernel consumes routed block lists with a runtime route-length bound. The previous chunk is always retained within the route budget, while current-chunk local

neighbors and reference-frame blocks act as deterministic safeguards. Frustum slab state is prepared asynchronously with a two-chunk freshness lag, so chunk c normally consumes geometry through chunk $c-2$. In the reported timing traces, DiT denoising and asynchronous geometry construction run concurrently on one H200.

We evaluate quality on all 158 cases of the WBench navigation split (Ying et al., 2026). The three LingBot-based rows use identical generated cases and the Gemini judge for VLM-scored metrics. Published WBench rows use the official Doubao judge, so judge-dependent values should not be compared across these two groups. We therefore include an API-free average that excludes VLM-scored setting metrics and is comparable across judge backends.

4.2 Attention Scaling over 48 Chunks

Figure 3 reports 48-chunk timing traces. Dense FlashAttention grows with the retained context, while Velox remains nearly flat because each query block receives a bounded route. At chunk 48, per-call attention latency decreases from approximately 34.2 ms to 7.8 ms, corresponding to a $4.36\times$ speedup. Under the same single-H200 asynchronous setting, chunk denoising time decreases from approximately 7.19 s to 2.96 s, corresponding to a $2.43\times$ speedup. The widening separation across the rollout shows that the gain stems from bounding the growing historical attention term rather than from a fixed short-context overhead.



(a) Per-call attention time.

(b) Chunk denoising time.

Figure 3: **Runtime scaling over a 48-chunk rollout.** The left plot measures per-call visual-attention latency. The right plot measures per-chunk DiT denoising time while asynchronous geometry construction shares the same H200. Velox increasingly separates from dense FlashAttention as the retained context grows.

4.3 Comparison on WBench

Table 1 compares dense LingBot-Fast-Cam, model-only block-sparse routing, and Velox under the same 158-case protocol. Velox remains within 0.15 points of dense attention on the API-free average and improves the 19-metric overall average from 75.52 to 75.60. Against the block-sparse baseline, soft geometric fusion raises the API-free average from 73.11 to 73.51, the overall average from 75.01 to 75.60, and spatial consistency from 71.8 to 74.3. Thus, geometry adds value beyond the matched sparse routing budget while preserving aggregate quality near dense attention.

4.4 Geometry and Model-Importance Agreement

Soft fusion assumes that frustum slab overlap identifies useful history without imposing a hard replacement set. We audit this relation using offline affinity dumps. For layer ℓ, threshold τ, and well-covered query block i, let $\mathcal{H}_i$ be the K_{hist} candidate history blocks after removing the self block, $\mathcal{G}_i(\tau) = \{h \in \mathcal{H}_i : g_{i,h} > \tau\}$ the $n_i(\tau)$ positive-overlap blocks, and $\mathcal{P}_i^{50}(\ell)$ the model-preferred top-50 after averaging affinity over batch and heads.

With $a_i(\ell,\tau) = |\mathcal{G}_i(\tau) \cap \mathcal{P}_i^{50}(\ell)|$, geometric recall is $R_{\text{geo}}(\ell,\tau) = \mathbb{E}_i[a_i(\ell,\tau)/n_i(\tau)]$, against the random expectation $50/143 = 0.350$. The fixed-budget set $\mathcal{U}_i^{50}(\ell,\tau)$ contains all positive-overlap blocks and the highest-importance outside blocks needed to reach 50. Its agreement is

Table 1: **WBench navigation results on 158 cases.** Scores are in $[0, 100]$. Published baselines use Doubao, and the final three rows use Gemini for VLM-scored metrics. API-free excludes Setting, while Overall averages all 19 metrics.

Method	Average						Consistency							
	API-free Avg.	Overall Avg.	Quality Score	Setting Adh.	Interaction Adh.	Physical Score	Back. Const.	Spatial Const.	Gated Spatial	Persp. Const.	Segment Cont.	Geom. Const.	Photo. Const.	Subject Const.
Text-driven video generation models														
Seedance 1.5	73.03	79.68	82.1	82.9	68.0	68.4	89.6	72.7	72.4	70.5	96.2	82.4	76.8	90.1
Wan 2.7	72.35	80.76	81.5	91.4	66.0	71.8	89.4	71.0	71.0	78.2	92.4	83.7	76.4	90.7
Kling 3.0	74.03	81.53	81.4	91.0	70.3	69.3	92.3	75.2	75.1	76.8	93.0	88.9	79.9	88.5
YUME 1.5	71.87	76.51	77.7	72.4	72.0	65.2	90.3	71.5	71.4	48.0	99.4	88.0	83.3	88.8
HY-Video 1.5	74.12	81.16	77.6	85.6	71.8	67.4	92.1	79.2	75.1	86.6	99.4	94.6	80.3	91.6
LTX 2.3	69.42	76.21	77.1	85.2	67.6	64.9	88.3	70.2	70.2	69.8	75.8	76.9	79.2	87.2
LongCat-Video	71.83	78.65	75.4	72.3	63.1	68.9	95.1	83.3	66.2	81.5	99.4	95.4	82.2	93.4
Kairos 3.0	69.93	75.34	74.0	70.3	65.1	60.4	91.1	76.8	62.0	76.3	94.3	89.0	80.8	90.8
Camera-controlled world models														
SANA-WM	74.65	77.86	79.3	76.1	82.1	61.9	90.5	77.1	76.4	49.0	97.5	88.7	81.2	85.4
HY-World 1.5	77.78	80.43	78.1	72.2	87.5	66.3	92.7	90.6	84.9	62.5	100.0	92.0	83.1	89.1
Fantasy-World	72.65	77.58	72.4	71.3	72.1	66.8	94.2	80.6	64.2	79.8	100.0	95.3	84.8	92.5
InSpatio-World	73.95	77.99	71.5	71.4	72.8	65.2	95.0	93.8	66.5	72.5	100.0	97.3	87.4	94.4
Astra	65.67	67.31	67.1	59.6	67.7	51.4	85.3	64.7	63.3	30.0	86.6	85.6	87.5	83.5
Ours built on LingBot-Fast-Cam (Gemini judge)														
LingBot-Fast-Cam (Gao et al., 2026)	**73.66**	75.52	77.7	79.9	**79.4**	**39.0**	87.9	73.0	73.0	**81.7**	90.5	**80.5**	77.6	**87.2**
Block-Sparse	73.11	75.01	77.9	79.1	78.0	38.6	87.6	71.8	71.8	75.7	**94.3**	80.4	77.1	85.8
Velox	73.51	**75.60**	**78.1**	**80.3**	78.8	38.8	**88.0**	**74.3**	**74.3**	77.0	92.4	**80.5**	**78.0**	86.4

$R_{50}(\ell, \tau) = \mathbb{E}_i[|\mathcal{P}_i^{50}(\ell) \cap \mathcal{U}_i^{50}(\ell, \tau)|/50]$. We omit rows with no positive overlap or non-finite scores and average over rows with at least 0.9 coverage from the last chunk of four calibration scenes and four denoising steps.

At $\tau = 0.1$, Table 2 shows agreement above the random expectation, with layer 14 reaching $R_{\text{geo}} = 0.7302$ and $R_{50} = 0.8909$. The layer variation motivates retaining each model-derived distribution instead of imposing a layer-invariant geometric set. In a separate calibration over 48 pooled-importance grids, $\lambda = 2$ and $g_{\text{ref}} = 0.2$ raise fused top-80 positive-overlap recall from 0.602 to 0.852 while changing 9.0% of selected blocks. This operating point adds a measured geometric bias without replacing the model ranking.

Table 2: **Geometry and model-importance agreement at $\tau = 0.1$.** R_{geo}, lift over the random expectation, and fixed-budget overlap R_{50} are defined in the text.

Layer	R_{geo}	Lift	R_{50}
10	0.6812	+0.3315	0.8723
14	**0.7302**	**+0.3806**	**0.8909**
19	0.6929	+0.3433	0.8770
23	0.6729	+0.3233	0.8716
24	0.6848	+0.3351	0.8739
25	0.6652	+0.3155	0.8666

5 Conclusion

We present Velox, a training-free router that projects spatial memory into the current latent camera view and uses frustum slab overlap to modulate layer-specific block importance while retaining model-derived fallback and local temporal safeguards. At chunk 48, Velox accelerates attention by $4.36\times$ and denoising by $2.43\times$ with asynchronous geometry construction on one H200. Across all 158 WBench navigation cases, it remains within 0.15 API-free average points of dense LingBot-Fast-Cam and slightly improves the 19-metric overall average. These results show that region-level spatial correspondence can turn persistent world memory into a practical routing signal for long-horizon generation.

6 Limitations

The evaluation covers one pretrained generator, one routing operating point, and one benchmark split. Several WBench metrics rely on VLM judges and can vary across judge backends. Broader validation should include additional generators, matched-budget fusion ablations, runtime confidence intervals, and rollouts beyond 48 chunks.

AI USE STATEMENT

We used generative AI tools only for language polishing. We did not use generative AI tools for research ideation, experimental design, code development, or data analysis. The authors reviewed all polished text and take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Section 3 specifies the complete routing approach, and Section 4 reports the evaluation protocol and implementation details. The appendix provides the routing configuration, geometry construction, execution path, and validation scope needed to reproduce the reported results. Anonymous visual results are available through the project page linked in the abstract, and all code and results will be made public.

REFERENCES

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos J. Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. *Advances in Neural Information Processing Systems*, 37:58757–58791, 2024.

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. Must3r: Multi-view network for stereo 3d reconstruction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1050–1060, 2025.

Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. Gamegen-x: Interactive open-world game video generation. *arXiv preprint arXiv:2411.00769*, 2024.

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

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Juncheng Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengchen Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025a.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7310–7320, 2024b.

Luxi Chen, Zihan Zhou, Min Zhao, Yikai Wang, Ge Zhang, Wenhao Huang, Hao Sun, Ji-Rong Wen, and Chongxuan Li. Flexworld: Progressively expanding 3d scenes for flexible-view synthesis. *arXiv preprint arXiv:2503.13265*, 2025b.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

Etched Decart, Quinn McIntyre, Spruce Campbell, Xinlei Chen, and Robert Wachen. Oasis: A universe in a transformer. *URL: https://oasis-model. github. io*, 2024.

Hangliang Ding, Dacheng Li, Runlong Su, Peiyuan Zhang, Zhijie Deng, Ion Stoica, and Hao Zhang. Efficient-vdit: Efficient video diffusion transformers with attention tile, 2025. URL https://arxiv.org/abs/2502.06155.

Paul Engstler, Andrea Vedaldi, Iro Laina, and Christian Rupprecht. Invisible stitch: Generating smooth 3d scenes with depth inpainting. In *Arxiv*, 2024.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Ruili Feng, Han Zhang, Zhantao Yang, Jie Xiao, Zhilei Shu, Zhiheng Liu, Andy Zheng, Yukun Huang, Yu Liu, and Hongyang Zhang. The matrix: Infinite-horizon world generation with real-time moving control. *arXiv preprint arXiv:2412.03568*, 2024.

Weilun Feng, Guoxin Fan, Haotong Qin, Chuanguang Yang, Mingqiang Wu, Yuqi Li, Xiangqi Li, Zhulin An, Libo Huang, Dingrui Wang, et al. Worldcache: Accelerating world models for free via heterogeneous token caching. *arXiv preprint arXiv:2603.06331*, 2026.

Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025.

Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, Yihang Chen, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.

Yuchao Gu, Weijia Mao, and Mike Zheng Shou. Long-context autoregressive video modeling with next-frame prediction. *arXiv preprint arXiv:2503.19325*, 2025.

Junxian Guo, Haotian Tang, Shang Yang, Zhekai Zhang, Zhijian Liu, and Song Han. Block Sparse Attention. https://github.com/mit-han-lab/Block-Sparse-Attention, 2024.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. *arXiv preprint arXiv:2503.10592*, 2025.

Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. *arXiv preprint arXiv:2403.14773*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020.

Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson W. H. Lau, Wangmeng Zuo, et al. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *arXiv preprint arXiv:2506.04225*, 2025a.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025b.

Team HY-World, Chenjie Cao, Xuhui Zuo, Zhenwei Wang, Yisu Zhang, Junta Wu, Zhenyang Liu, Yuning Gong, Yang Liu, Bo Yuan, et al. Hy-world 2.0: A multi-modal world model for reconstructing, generating, and simulating 3d worlds. *arXiv preprint arXiv:2604.14268*, 2026.

Huiqiang Jiang, Yucheng Li, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Zhenhua Han, Amir Abdi, Dongsheng Li, Chin-Yew Lin, et al. Minference 1.0: Accelerating pre-filling for long-context llms via dynamic sparse attention. *NeurIPS*, 37:52481–52515, 2024.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*, 2024.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. MapAnything: Universal feed-forward metric 3D reconstruction, 2025. arXiv preprint arXiv:2509.13414.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):1–14, 2023.

Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. Fifo-diffusion: Generating infinite videos from text without training. *arXiv preprint arXiv:2405.11473*, 2024.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *ICLR*, 2020.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Heejun Lee, Geon Park, Jaduk Suh, and Sung Ju Hwang. Infinitehip: Extending language model context up to 3 million tokens on a single gpu. *arXiv preprint arXiv:2502.08910*, 2025.

Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r, 2024.

Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. Vmem: Consistent interactive video scene generation with surfel-indexed view memory. *arXiv preprint arXiv:2506.18903*, 2025.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. Snapkv: Llm knows what you are looking for before generation. *Advances in Neural Information Processing Systems*, 37:22947–22970, 2024.

Feng Liu, Shiwei Zhang, Xiaofeng Wang, Yujie Wei, Haonan Qiu, Yuzhong Zhao, Yingya Zhang, Qixiang Ye, and Fang Wan. Timestep embedding tells: It's time to cache for video diffusion model. *arXiv preprint arXiv:2411.19108*, 2024a.

Hao Liu, Matei Zaharia, and Pieter Abbeel. Ring attention with blockwise transformers for near-infinite context. *arXiv preprint arXiv:2310.01889*, 2023.

Jinlai Liu, Jian Han, Bin Yan, Hui Wu, Fengda Zhu, Xing Wang, Yi Jiang, Bingyue Peng, and Zehuan Yuan. Infinitystar: Unified spacetime autoregressive modeling for visual generation. *arXiv preprint arXiv:2511.04675*, 2025a.

Songhua Liu, Zhenxiong Tan, and Xinchao Wang. Clear: Conv-like linearization revs pre-trained diffusion transformers up. *arXiv preprint arXiv:2412.16112*, 2024b.

Ziming Liu, Yifan Yang, Chengruidong Zhang, Yiqi Zhang, Lili Qiu, Yang You, and Yuqing Yang. Region-adaptive sampling for diffusion transformers. *arXiv preprint arXiv:2502.10389*, 2025b.

Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, et al. Step-video-t2v technical report: The practice, challenges, and future of video foundation model. *arXiv preprint arXiv:2502.10248*, 2025.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106, 2021.

Umair Nawaz, Ahmed Heakl, Ufaq Khan, Abdelrahman Shaker, Salman Khan, and Fahad Shahbaz Khan. Worldcache: Content-aware caching for accelerated video world models. *arXiv preprint arXiv:2603.22286*, 2026.

Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, Stephen Spencer, Jessica Yung, Michael Dennis, Sultan Kenjeyev, Shangbang Long, Vlad Mnih, Harris Chan, Maxime Gazeau, Bonnie Li, Fabio Pardo, Luyu Wang, Lei Zhang, Frederic Besse, Tim Harley, Anna Mitenkova, Jane Wang, Jeff Clune, Demis Hassabis, Raia Hadsell, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 2: A large-scale foundation world model, 2024. URL https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

Sand.ai. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

Meituan LongCat Team, Xunliang Cai, Qilong Huang, Zhuoliang Kang, Hongyu Li, Shijun Liang, Liya Ma, Siyu Ren, Xiaoming Wei, Rixu Xie, et al. Longcat-video technical report. *arXiv preprint arXiv:2510.22200*, 2025.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024.

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

Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21686–21697, 2024a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. *arXiv preprint arXiv:2501.12387*, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024b.

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

Zun Wang, Han Lin, Jaehong Yoon, Jaemin Cho, Yue Zhang, and Mohit Bansal. Anchorweave: World-consistent video generation with retrieved local spatial memories. *arXiv preprint arXiv:2602.14941*, 2026.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. *arXiv preprint arXiv:2506.05284*, 2025.

Haocheng Xi, Shuo Yang, Yilong Zhao, Chenfeng Xu, Muyang Li, Xiuyu Li, Yujun Lin, Han Cai, Jintao Zhang, Dacheng Li, et al. Sparse videogen: Accelerating video diffusion transformers with spatial-temporal sparsity. *arXiv preprint arXiv:2502.01776*, 2025.

Yifei Xia, Suhan Ling, Fangcheng Fu, Yujie Wang, Huixia Li, Xuefeng Xiao, and Bin Cui. Training-free and adaptive sparse attention for efficient long video generation. *arXiv preprint arXiv:2502.21079*, 2025.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. *arXiv preprint arXiv:2504.12369*, 2025.

Desai Xie, Zhan Xu, Yicong Hong, Hao Tan, Difan Liu, Feng Liu, Arie Kaufman, and Yang Zhou. Progressive autoregressive video diffusion models. *arXiv preprint arXiv:2410.08151*, 2024.

Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025a.

Mingyu Yang, Junyou Li, Zhongbin Fang, Sheng Chen, Yangbin Yu, Qiang Fu, Wei Yang, and Deheng Ye. Playable game generation. *arXiv preprint arXiv:2412.00887*, 2024a.

Shuo Yang, Haocheng Xi, Yilong Zhao, Muyang Li, Jintao Zhang, Han Cai, Yujun Lin, Xiuyu Li, Chenfeng Xu, Kelly Peng, et al. Sparse videogen2: Accelerate video generation with sparse attention via semantic-aware permutation. *arXiv preprint arXiv:2505.18875*, 2025b.

Xiaoda Yang, Jiayang Xu, Kaixuan Luan, Xinyu Zhan, Hongshun Qiu, Shijun Shi, Hao Li, Shuai Yang, Li Zhang, Checheng Yu, et al. Omnicam: Unified multimodal video generation via camera control. *arXiv preprint arXiv:2504.02312*, 2025c.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024b.

Kaining Ying, Hengrui Hu, Siyu Ren, Jiamu Li, Fengjiao Chen, Ziwen Wang, Xuezhi Cao, Xunliang Cai, and Henghui Ding. Wbench: A comprehensive multi-turn benchmark for interactive video world model evaluation. *arXiv preprint arXiv:2605.25874*, 2026.

Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T. Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. *arXiv:2406.09394*, 2024a.

Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. Wonderjourney: Going from anywhere to everywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6658–6667, 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. *arXiv preprint arXiv:2506.03141*, 2025a.

Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactory: Creating new games with generative interactive videos. *arXiv preprint arXiv:2501.08325*, 2025b.

Mark YU, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. *arXiv preprint arXiv:2503.05638*, 2025.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024c.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297, 2020.

Haiyu Zhang, Xinyuan Chen, Yaohui Wang, Xihui Liu, Yunhong Wang, and Yu Qiao. Accvideo: Accelerating video diffusion model with synthetic dataset. *arXiv preprint arXiv:2503.19462*, 2025a.

Jintao Zhang, Haofeng Huang, Pengle Zhang, Jun Zhu, Jianfei Chen, et al. Sageattention: Accurate 8-bit attention for plug-and-play inference acceleration. *arXiv preprint arXiv:2410.02367*, 2024.

Jintao Zhang, Chendong Xiang, Haofeng Huang, Jia Wei, Haocheng Xi, Jun Zhu, and Jianfei Chen. Spargeattention: Accurate and training-free sparse attention accelerating any model inference. *Proceedings of the 42nd International Conference on Machine Learning*, 2025b.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023a.

Peiyuan Zhang, Yongqi Chen, Runlong Su, Hangliang Ding, Ion Stoica, Zhenghong Liu, and Hao Zhang. Fast video generation with sliding tile attention. *arXiv preprint arXiv:2502.04507*, 2025c.

Yifan Zhang, Chunli Peng, Boyang Wang, Puyi Wang, Qingcheng Zhu, Zedong Gao, Eric Li, Yang Liu, and Yahui Zhou. Matrix-game: Interactive world foundation model. *arXiv*, 2025d.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, et al. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in Neural Information Processing Systems*, 36:34661–34710, 2023b.

Appendices

A Additional Experimental Evidence

A.1 Soft-Fusion Calibration

We calibrate the fixed fusion strength on pooled query-key distributions and frustum slab overlap matrices from two navigation trajectories. The calibration contains 48 grids spanning four representative layers, two denoising steps, and three rollout lengths. Positive overlap values have median 0.211, which motivates rescaling rather than multiplying by raw IoU. Among the tested settings, $\lambda = 2$ and $g_{\mathrm{ref}} = 0.2$ are the weakest pair that recovers at least 85% of positive-overlap blocks in the fused top-80 while changing at most 25% of the model-only selections. Concretely, positive-overlap recall increases from 0.602 to 0.852, and the selected set changes by 9.0% on average. Values at or below $\tau = 0.1$ receive no boost.

B Experimental Implementation Details

Velox is instantiated on the LingBot-Fast-Cam inference stack (Gao et al., 2026) with a Wan-style video diffusion transformer. Its routing backend uses frustum slab matching: stored chunks provide source-camera token rectangles and robust depth intervals, target query blocks obtain Q-depth through multiwarp, and Eq. (7) provides the world-space OBB overlap score.

B.1 Reported Routing Configuration

The reported configuration applies the fixed fusion in Eq. (11) only when $\rho_i \geq 0.9$. It uses $\lambda = 2$, $g_{\mathrm{ref}} = 0.2$, $\tau = 0.1$, and $K_{\mathrm{cov}} = 80$. Rows below the coverage threshold retain the model-only cumulative-mass route and are capped at $K_{\mathrm{unc}} = 150$ without a minimum route-size floor. The previous chunk is reserved inside these budgets. Global Gilbert neighbors and reference-frame blocks are then added as safeguards. The same sparse routing rule is applied at every transformer layer.

B.2 Camera Pose Alignment for Frustum Slab Geometry

The frustum slab table is reconstructed in the LingBot-map camera-to-world frame, while the chunk cameras used by the video generator come from ViPE poses. LingBot-map does not accept external camera poses during streaming reconstruction. Velox therefore maps each target ViPE camera into the reconstructed coordinate frame before multiwarp, OBB construction, frustum slab IoU, and coverage estimation.

Let (R_*^V, p_*^V) and (R_*^L, p_*^L) denote the most recent matched ViPE and LingBot-map camera-to-world poses. The alignment fit over seen camera centers supplies a positive scale α, and the latest pose pair defines the bridge rotation $R_b = R_*^L (R_*^V)^\top$. For a target ViPE camera (R_q^V, p_q^V), the reported relative alignment constructs

$$\widehat{R}_q = R_b R_q^V, \qquad \widehat{p}_q = p_*^L + R_b\big(\alpha(p_q^V - p_*^V)\big)\,. \tag{18}$$

This construction maps the latest ViPE pose exactly to the latest LingBot-map pose. It also preserves the target camera's relative rotation and scaled translation from that anchor. The frustum slab builder consumes $\widehat{C}_q^{-1}$, where $\widehat{C}_q = [\widehat{R}_q \mid \widehat{p}_q]$, as the target world-to-camera matrix. Before scoring, the implementation refreshes historical slab cameras under the current bridge so that cached geometry and target queries share one coordinate frame.

B.3 Recency-Priority Multiwarp for Query-Block Depth

The current noisy chunk does not provide a reconstructed depth map, yet each query block needs a target-camera depth interval for geometric matching. Using only the latest reconstructed frame leaves holes after camera rotation or scene revisitation. Pooling all historical warps is also undesirable because stale and recent observations can place different surfaces in the same target token. Multiwarp

resolves this ambiguity by selecting the newest source frame with enough support before computing any depth statistic.

Bounded source cache. For each reconstructed chunk, Velox caches the pixel depths, camera poses, intrinsics, confidence values, and validity masks of its first and last frames. A ring buffer retains the most recent K chunks, giving at most $2K$ source frames. For source frame f, a pixel p is retained only when its depth $D_f(p)$ is finite and exceeds ϵ_z, its confidence lies above the configured lower-percentile cutoff, and it passes the available static-region mask. An optional sky mask can further reject unreliable distant regions. The reported configuration uses $K = 30$, $\epsilon_z = 0.01$, and no sky rejection. It drops the bottom 20% of confidence values among pixels with valid depth. Each cached depth map remains paired with its own LingBot-map pose, which is the camera pose under which that depth was reconstructed.

Warping into target latent cameras. Let $\mathbf{K}_f$ and $[R_f \mid \mathbf{t}_f]$ be the source intrinsics and world-to-camera pose. For a valid source pixel with homogeneous coordinate $\bar{\mathbf{p}}$, multiwarp first recovers its world point:

$$\mathbf{x}_p^f = D_f(p)\mathbf{K}_f^{-1}\bar{\mathbf{p}}, \qquad \mathbf{X}_{f,p}^w = R_f^\top\left(\mathbf{x}_p^f - \mathbf{t}_f\right). \tag{19}$$

For target latent frame t, let $\mathbf{K}_t^Q$ and $[R_t^Q \mid \mathbf{t}_t^Q]$ denote the intrinsics and aligned world-to-camera pose from Eq. (18). The source point becomes

$$\mathbf{x}_{f,p,t}^Q = R_t^Q\mathbf{X}_{f,p}^w + \mathbf{t}_t^Q, \qquad \widetilde{\mathbf{p}}_{f,p,t}^Q = \pi\Big(\mathbf{K}_t^Q\mathbf{x}_{f,p,t}^Q\Big), \tag{20}$$

where π applies perspective division. Samples behind the target camera or outside the target image are discarded. If one latent token spans $h_c \times w_c$ target-image pixels, a projected pixel $(\widetilde{x}, \widetilde{y})$ is assigned to token coordinates $(\lfloor \widetilde{y}/h_c \rfloor, \lfloor \widetilde{x}/w_c \rfloor)$. The third coordinate of $\mathbf{x}_{f,p,t}^Q$ supplies its target-camera depth. We write the depth multiset assigned to token u as $\mathcal{Z}_{f,t,u}$ and its size as $n_{f,t,u}$.

Newest reliable source. Cached frames are ordered from oldest to newest. Multiwarp selects the newest frame that contributes at least k_{warp} samples to a target token:

$$f_{t,u}^\star = \max\left\{f \mid n_{f,t,u} \geq k_{\mathrm{warp}}\right\}. \tag{21}$$

Only samples from $f_{t,u}^\star$ contribute to that token's statistics. The implementation records the quantile ladder

$$z_Q^\tau(t,u) = \mathrm{Quantile}_\tau\Big(\mathcal{Z}_{f_{t,u}^\star,t,u}\Big), \quad \tau \in \{5,10,20,50,80,90,95\}, \tag{22}$$

and uses $z_Q(t,u) = z_Q^{50}(t,u)$ as the token depth. If the newest frame lacks enough samples, an older frame may fill the hole. If no source reaches the sample floor, the token remains invalid. Selecting the source before taking quantiles prevents depth samples from different observations and surfaces from being mixed.

From token depth to a query frustum slab. A global Gilbert query block can cross latent-frame boundaries, so Velox first divides it into frame-local rectangular fragments. For fragment $B_{i,t}$, define its valid token set as $\mathcal{V}_{i,t} = \{u \in B_{i,t} \mid f_{t,u}^\star \text{ exists}\}$. Its target-camera depth interval is

$$z_{i,t}^- = \mathrm{Quantile}_{p_{\mathrm{lo}}}(\{z_Q(t,u) : u \in \mathcal{V}_{i,t}\}), \qquad z_{i,t}^+ = \mathrm{Quantile}_{p_{\mathrm{hi}}}(\{z_Q(t,u) : u \in \mathcal{V}_{i,t}\}). \tag{23}$$

The implementation enforces $z_{i,t}^+ \geq z_{i,t}^- + \epsilon_{\mathrm{slab}}$. The fragment's token rectangle and this robust interval define the query frustum slab used for world-space OBB matching. An empty $\mathcal{V}_{i,t}$ contributes no geometric evidence. If the resulting query block does not reach the coverage threshold, the router follows the model-derived path in Eq. (13).

Implementation path. The GPU procedure has two logical passes. The first counts valid samples for every source-frame, target-frame, and target-token tuple and selects $f_{t,u}^\star$. The second compacts only winner-frame samples and computes segmented quantiles. This ordering makes the source cache bounded, avoids cross-frame depth mixing, and supplies block-aligned Q-depth without reconstructing the noisy target chunk. The reported configuration uses $k_{\mathrm{warp}} = 2$ and $(p_{\mathrm{lo}}, p_{\mathrm{hi}}) = (10, 90)$. It does not apply the optional source-to-target pose gate or the experimental warp occluder.

B.4 Runtime Path and Timing Scope

The measured sparse path is a Triton indexed block-sparse attention implementation, not a Python masking wrapper around dense attention. Queries are reordered once per chunk into the global-aligned Gilbert token order, and KV caches are written directly in that layout so that the hot kernel loads contiguous KV tiles without a per-call host permutation. The kernel supports rectangular current-query versus growing-history attention, consumes the routed block lists produced by the fused-score scheduler, and uses a runtime route-length bound so later chunks do not trigger new compiler specializations. Section D gives the full kernel path.

Frustum slab geometry is built asynchronously. After a chunk is decoded, a worker thread on a dedicated CUDA stream updates reconstruction state, records chunk cameras, builds stored frustum slabs, refreshes multiwarp frame caches, and prepares the Velox state for a later chunk while the main stream continues DiT denoising. The implementation uses a two-chunk freshness lag. Chunk c normally consumes geometry through chunk $c-2$, which is sufficient because frustum slab overlap changes gradually. The worker is the sole writer to the geometry objects. The main stream waits only on a completion event when a future state is needed. Section C gives the full implementation path, including the worker stream, VAE side stream, prewarming, and tail handling.

All speed numbers are measured on one H200 from the 48-chunk timing traces in Figure 3. We report per-chunk attention-call latency and per-chunk denoising critical-path time rather than a single aggregate wall-clock number because early chunks have small dense-attention baseline cost, while late chunks expose the scaling behavior that matters for long video generation.

C Asynchronous Frustum Slab Construction

The geometry pipeline would be expensive if it were inserted serially between generated chunks. In a synchronous implementation, chunk c must be denoised, decoded, reconstructed into depth and pose state, inserted into the frustum slab table and multiwarp cache, aligned with the next target cameras, and scored before chunk $c+1$ can denoise. The reconstruction and frustum slab preparation path is on the same order as one DiT denoise step in our setting, so serial execution would hide the attention-kernel gains.

Velox relaxes this dependency with a two-chunk freshness lag. Target chunk c consumes a Velox state built from frustum slabs through chunk $c-2$, while the worker prepares states for later chunks. The state remains self-consistent: each worker task first inserts exactly one decoded chunk into the frustum slab table and multiwarp cache, then scores that table against the target chunk's latent cameras. Chunks before the first available Velox state run dense attention, so the method never consumes an uninitialized route.

The implementation uses one Python worker thread and one dedicated CUDA stream. The worker is the sole writer to the frustum slab table, pose aligner, multiwarp frame cache, and sparse-attention context. The main denoising thread submits a geometry task after a chunk decode and receives a future keyed by the target chunk. Immediately before chunk c denoises, the main stream waits on the future's CUDA event if the worker has not already finished, then consumes the prepared geometric and importance states. This wait is a device-side stream dependency. There is no host-wide synchronization in the hot path.

Each geometry task performs four operations. First, the reconstructed frames are passed through the streaming reconstruction backend to update depths, poses, confidence, and masks. Second, the relative alignment maps target camera-control poses into the reconstructed frame and refreshes stored slab cameras under the current bridge. Third, valid depth samples are converted into per-block frustum slab primitives and inserted into the versioned frustum slab table, while first and last frame depths are cached for multiwarp. Fourth, the sparse context warps cached depths into the target cameras, builds Q frustum slabs, scores OBB overlap against historical frustum slabs, computes coverage, and materializes the geometric state consumed by soft fusion at each attention layer.

The implementation also moves VAE decode off the denoise critical path with a side stream, prewarms the worker thread's cuDNN handle, prewarms segmentation and sparse-attention kernels outside the measured interval, and skips frustum slab scoring for the final chunk because no later denoise

consumes it. Frames are moved to CPU after decode and copied back on the worker stream, avoiding extra GPU residency while the main stream denoises.

All reported timing results use a single NVIDIA H200. The async worker and the DiT denoiser share that one GPU. Memory reconstruction is not offloaded to a second device. A second GPU could reduce CUDA-stream contention and further improve end-to-end time, but it is not part of the reported measurement protocol.

D VELOX ATTENTION KERNEL AND GLOBAL GILBERT LAYOUT

The sparse route is implemented as a native indexed attention path. The router produces block lists, the KV cache is stored in the order consumed by those lists, and the Triton kernel performs only the selected block interactions.

D.1 GLOBAL-ALIGNED GILBERT TOKEN ORDER

Let a latent chunk contain T frames, spatial grid height H, spatial grid width W, and $S = THW$ tokens. The row-major token index is

$$\lambda(t, y, x) = tHW + yW + x, \qquad 0 \le t < T,\ 0 \le y < H,\ 0 \le x < W.$$

Here t, y, and x are the frame, row, and column indices within the chunk. Velox uses a global-aligned Gilbert map $\Gamma : \{0, \dots, (c+1)S-1\} \to \{0, \dots, (c+1)S-1\}$ over the visible prefix, where Γ maps a global row-major token index to its global Gilbert rank. The implementation builds a single-chunk Gilbert rank map γ once, then fills each chunk with a parity-dependent reflection. If (cT) is even, chunk c uses the single-chunk map directly. If (cT) is odd, the chunk-local coordinates are reflected by the base-frame exit quadrant before ranks are offset by cS. For the chunk-local reflection F_c, the global rank is:

$$\Gamma(cS + \lambda(t, y, x)) = cS + \gamma(F_c(\lambda(t, y, x))) .$$

This incremental rule is rank-identical to building the full-sequence Gilbert curve over all visible frames. With the same token block size $b_{\text{tok}} = 128$ used in Section 3.1, the global block id is:

$$\beta_{\text{glob}}(c, \lambda) = \left\lfloor \frac{\Gamma(cS + \lambda)}{b_{\text{tok}}} \right\rfloor .$$

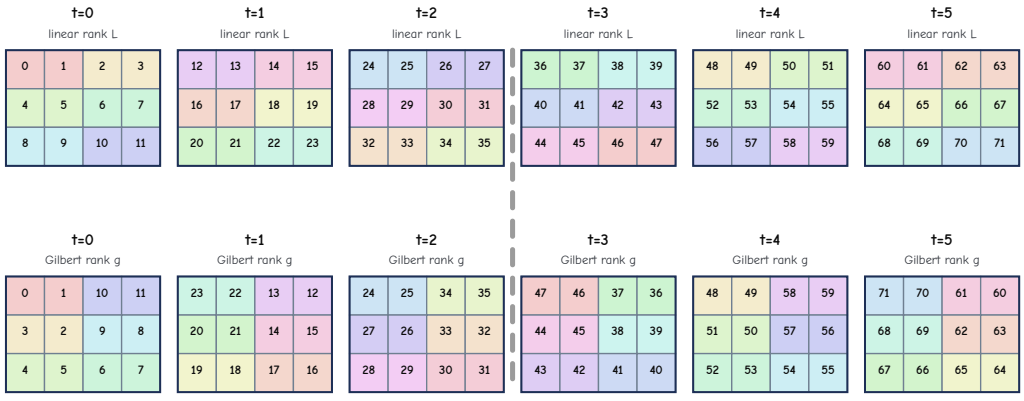


Figure 4: **THW linear order versus global Gilbert order.** Row-major THW ranking scans each frame independently, whereas a single global Gilbert curve preserves spatial locality across six frames and continues across the chunk boundary.

D.2 GLOBAL GILBERT KV STORAGE

Velox writes each chunk's KV cache directly in its global-aligned Gilbert order. Let $\mathbf{K}_{c,\lambda,a,:}$ and $\mathbf{V}_{c,\lambda,a,:}$ denote the linear-layout key and value vectors for chunk c, token λ, and attention head a.

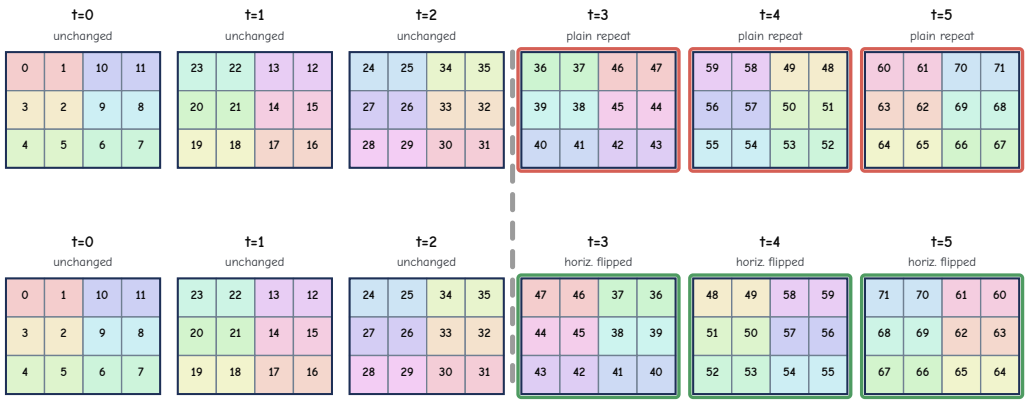


Figure 5: **Chunk-wise construction of the global Gilbert order.** Repeating the single-chunk order without reorientation creates a mismatch after the chunk boundary. Applying the parity-dependent horizontal reflection to the second chunk reproduces the corresponding segment of the full-sequence Gilbert order.

Let κ denote a within-chunk Gilbert rank, and let $G_c(\kappa)$ be the gather index that maps κ back to the corresponding chunk-local linear token. The stored tensors are:

$$\widetilde{\mathbf{K}}_{c,\kappa,a,:} = \mathbf{K}_{c,G_c(\kappa),a,:}, \qquad \widetilde{\mathbf{V}}_{c,\kappa,a,:} = \mathbf{V}_{c,G_c(\kappa),a,:}, \qquad 0 \le \kappa < S.$$

Each chunk slot is front-padded as needed so the current chunk's query blocks align to the global block ids in the visible prefix, and padded positions are zero-filled. The sparse kernel therefore sees a chunked contiguous address range whose block boundaries match β_{glob}.

The implementation uses separate query and KV sequence lengths, because current-chunk queries attend to a growing history cache. This rectangular interface is what lets Velox replace only historical visual attention while leaving the pretrained chunked generator interface unchanged.

D.3 Indexed Sparse Attention Kernel

Fix a current chunk c and layer ℓ. Let $\mathcal{J}_i^\ell$ be the materialized block list consumed by the kernel for query block i. It is the implementation-level form of $\overline{\mathcal{R}}_i^\ell$ in Eq. (14). For the query tokens in block i, the kernel computes:

$$\mathbf{O}_i = \mathrm{softmax}\left(\frac{\mathbf{Q}_i \mathbf{K}_{\mathcal{J}_i^\ell}^\top}{\sqrt{d}}\right) \mathbf{V}_{\mathcal{J}_i^\ell}.$$

The block list is different for each query block, but the route-length bound is a runtime value rather than a compiler specialization. This avoids per-chunk Triton recompilation as the retained history grows.

The kernel evaluates the softmax with online normalization. For routed KV tile step q with score matrix $\mathbf{S}^{(q)}$, let $m^{(q)}$, $w^{(q)}$, and $\mathbf{A}^{(q)}$ denote the row-wise maximum, softmax normalizer, and value accumulator before processing that tile. The update is:

$$m^{(q+1)} = \max\left(m^{(q)}, \max_v \mathbf{S}_v^{(q)}\right), \qquad \alpha^{(q)} = \exp\left(m^{(q)} - m^{(q+1)}\right),$$

$$w^{(q+1)} = \alpha^{(q)} w^{(q)} + \sum_v \exp\left(\mathbf{S}_v^{(q)} - m^{(q+1)}\right),$$

$$\mathbf{A}^{(q+1)} = \alpha^{(q)} \mathbf{A}^{(q)} + \exp\left(\mathbf{S}^{(q)} - m^{(q+1)}\right) \mathbf{V}^{(q)}.$$

After the final routed tile, $\mathbf{O} = \mathbf{A}/w$. In the Gilbert-storage path, KV columns are contiguous, so the kernel can use contiguous memory hints and fuse the accumulator scaling into the matrix multiply update. The gathered linear-storage path remains available for validation and ablation, but the reported timings use the contiguous Gilbert-storage path.

Velox uses a small routed-tile group in the inner loop. If $L_i = |\mathcal{J}_i^\ell|$ and $B_{\mathrm{grp}} = 4$, the loop has $\lfloor L_i / B_{\mathrm{grp}} \rfloor$ full groups and $L_i \bmod B_{\mathrm{grp}}$ tail blocks. The full groups are predicate-free, and only the final tail is masked. This two-phase loop recovered the runtime-bound implementation from the scheduling overhead caused by checking the tail predicate in every unrolled group.

D.4 Route Materialization and Caches

The route materializer reserves the previous chunk before selecting older history. For covered rows, it ranks the remaining blocks by the fused distribution:

$$\mathcal{J}_i^\ell = \mathrm{unique}\Big(\mathcal{P}_{c-1} \cup \mathrm{Top}_{K_{\mathrm{cov}} - |\mathcal{P}_{c-1}|}\left(\{\widetilde{p}_{i,r}^\ell : r \notin \mathcal{P}_{c-1}\}\right) \cup \mathcal{L}_i^{\mathrm{glob}} \cup \mathcal{F}_0\Big), \qquad \rho_i \geq \eta_{\mathrm{cov}}.$$

For uncovered rows, it applies the same reservation to the capped model-only route:

$$\mathcal{J}_i^\ell = \mathrm{unique}\Big(\mathcal{P}_{c-1} \cup \mathrm{Prefix}_{\eta_{\mathrm{imp}}, K_{\mathrm{unc}} - |\mathcal{P}_{c-1}|}\left(\{p_{i,r}^\ell : r \notin \mathcal{P}_{c-1}\}\right) \cup \mathcal{L}_i^{\mathrm{glob}} \cup \mathcal{F}_0\Big), \qquad \rho_i < \eta_{\mathrm{cov}}.$$

The implementation floors the remaining selection budget to avoid an empty older-history route. The global-neighbor and reference-frame safeguards are added outside the long-history budget, which is why a materialized route can contain more than K_{cov} blocks.

The route builder memoizes frustum slab overlap tables, coverage, global/window local-neighbor adjacency, and pooled historical key summaries. It builds the block mask by direct indexed scatter from row and column tensors, avoiding synchronization-heavy coordinate extraction. Cross-chunk boundary neighbors are already present in the global/window 26-neighbor adjacency, so no separate seam stamping appears in the current route materialization.

D.5 Failure Modes and Validation Scope

Velox is most reliable when the frustum slab table and multiwarp Q-depth explain the current view. The expected failure cases are therefore geometry-driven. If reconstructed depth drifts, if the multiwarp cache fills a newly disoccluded token with stale far geometry, or if OBB approximation overestimates overlap for a thin structure, soft fusion can promote stale or misleading history. The current router mitigates these cases by applying geometry only to well-covered rows, bounding the geometric multiplier, retaining the layer-specific model distribution, and preserving previous-chunk and local-neighbor safeguards. The goal is to remove history that is spatially irrelevant while retaining enough model evidence for noisy generated-video geometry.

The WBench evaluation also suggests where future quality regressions should be inspected. Velox improves spatial and photometric consistency over dense attention, but perspective and subject consistency decrease. This is consistent with the design risk: static background regions are usually supported by frustum slabs, while object- or camera-sensitive regions can be dynamic, partially reconstructed, or semantically important despite weak geometric correspondence. A stronger quality study should therefore report object-centric cases separately from background revisits and regions where Q-depth or frustum slab overlap assumptions are stressed.

Finally, speed and quality should be validated under matched budgets. A random same-size route can test whether any sparsity helps. A recent-window route can test whether temporal locality is enough. A frame/FOV route can test whether view-level retrieval is sufficient. A model-derived route can test whether geometry adds value beyond attention statistics. Velox should outperform these baselines specifically when current regions have different relevant histories. This validation target is sharper than only comparing against dense attention, because it isolates the contribution of frustum slab geometry as the routing signal.