# SEAMEM: SEEN-AWARE IMPLICIT MEMORY FOR LONG-HORIZON CAMERA-CONTROLLED WORLD GENERATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Long-horizon world generation aims to enable controllable exploration of a scene beyond initial observations by generating views along user-specified camera trajectories. As the viewpoint changes, each target view may contain both previously seen content and newly revealed regions. The former must remain consistent with historical observations, while the latter requires plausible synthesis guided by learned world priors. The key challenge is therefore how to represent and use historical observations to preserve previously seen content while allowing plausible synthesis in newly revealed regions. Existing methods typically rely on either explicit 3D memory or implicit context memory. However, 3D reconstruction can accumulate persistent geometric errors when integrating imperfect generated views. In contrast, implicit context memory avoids this reconstruction error, but lacks awareness of seen and unseen regions. Consequently, previously observed content may be unnecessarily resynthesized, introducing hallucinations and degrading long-horizon consistency. To address these issues, we introduce **SeaMem**, a seen-aware implicit memory for long-horizon camera-controlled world generation. SeaMem couples implicit recovery of historical content with explicit identification of where the recovered content should constrain generation. Specifically, SeaMem queries compact historical representations with the target camera to recover content and predicts an unseen-region mask from attention statistics. Recovered content anchors seen regions, while a pretrained novel view synthesis (NVS) backbone synthesizes unseen regions. Within seen regions, a detail-deficient mask relaxes anchoring where historical detail is insufficient, enabling refinement under historical constraints. This design extends short-range NVS backbones to long-horizon world generation without requiring additional backbone fine-tuning. Extensive experiments demonstrate that SeaMem effectively extends NVS backbones to long-horizon world generation while achieving robust revisit consistency.

## 1 INTRODUCTION

Interactive world models generate future visual observations conditioned on user or agent actions (Ha & Schmidhuber, 2018; Yang et al., 2024; Bar et al., 2025). Building on advances in controllable video generation (Li et al., 2025a; Mao et al., 2025), we study long-horizon camera-controlled world generation: synthesizing coherent view sequences from initial observations along user-specified camera trajectories. The goal is to support continued exploration beyond the initial field of view while maintaining consistency with previously observed content.

To support such persistent exploration, a long-horizon world model requires two complementary capabilities. First, it must maintain a queryable representation of historical observations to ensure that subsequent observations remain consistent with past views. Second, it must plausibly synthesize regions newly exposed by the target camera using learned generative priors. Recent novel view synthesis (NVS) models have achieved remarkable progress in generating visually plausible views under varying camera poses (Jin et al., 2025; Zhou et al., 2025; Liang et al., 2026), providing strong priors for synthesizing newly revealed content. However, they lack an effective mechanism for maintaining persistent scene memory over long trajectories. A central challenge is therefore how to



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

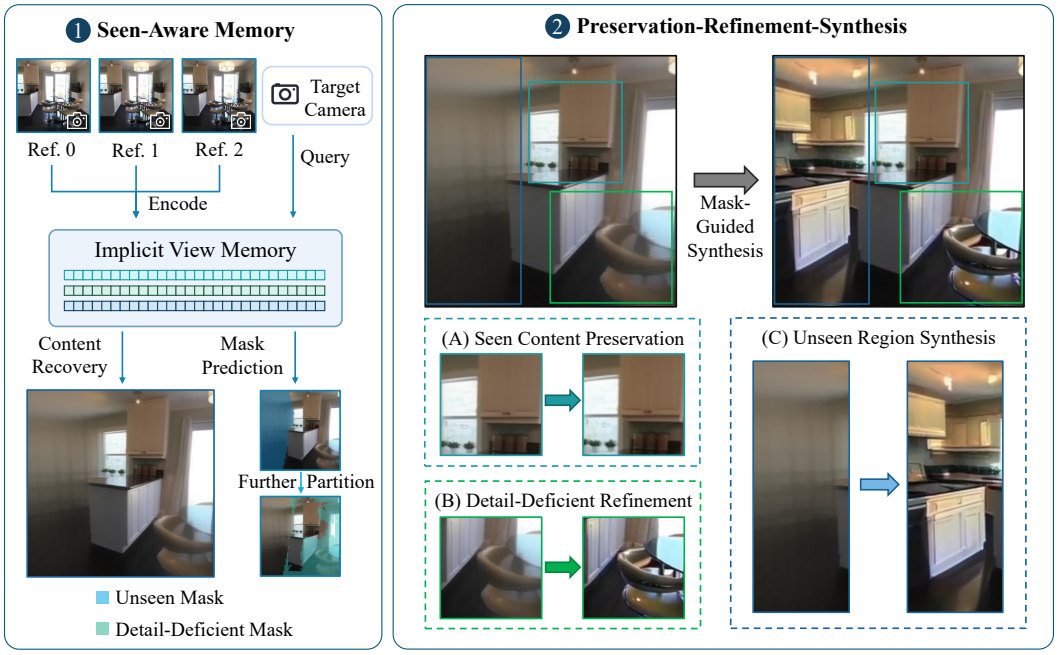



Figure 1: **From memory recovery to guided synthesis: preserving seen content, refining insufficient detail, and generating unseen content.** Given historical reference views and a target camera, SeaMem recovers target-view content and predicts unseen and detail-deficient region masks. In the right panel, arrows compare recovered content (left) with mask-guided NVS results (right): (A) sufficiently supported seen content is preserved; (B) detail-deficient seen content becomes sharper under historical constraints; and (C) blurred outputs in unseen regions are replaced with plausible, detailed content.

represent accumulated observations and use them to preserve previously seen content in subsequent target views over long horizons.

Existing world generation methods primarily address these problems through two world-memory paradigms, each with distinct limitations. **Explicit 3D Memory** reconstructs a 3D representation from historical views and renders it from the target viewpoint, thereby explicitly distinguishing seen from unseen regions (Ren et al., 2025; Li et al., 2025b; Zhao et al., 2026). However, repeatedly integrating imperfect generated views into the 3D memory can materialize cross-view inconsistencies as persistent geometric errors, which may propagate into subsequent generations (Zhao et al., 2026). In contrast, **Implicit Context Memory** encodes historical observations into latent context representations that condition target-view generation (Wu et al., 2026b; Yang et al., 2026; Chen et al., 2026), without relying on explicit 3D reconstruction and fusion. However, these methods usually condition generation over the entire target view without explicitly identifying the regions supported by historical observations, allowing hallucinations in known areas to introduce inconsistencies with earlier observations. Overall, these approaches either risk embedding cross-view inconsistencies into explicitly reconstructed geometry or lack an explicit mechanism for determining where historical content should constrain generation.

To overcome these limitations, we introduce **SeaMem**, a seen-aware implicit memory for long-horizon camera-controlled world generation. As illustrated in Figure 1, SeaMem encodes historical observations into compact latent representations and queries them via camera-conditioned attention to recover previously observed content in the target view, without reconstructing an explicit 3D scene representation. To determine where recovered content should constrain generation, SeaMem extracts spatial cues from attention patterns in the same memory-querying process. A lightweight visibility predictor decodes these cues into an explicit unseen-region mask, enabling recovered content to anchor seen regions while a pretrained NVS model synthesizes newly revealed content. Historical visibility, however, does not necessarily imply sufficient detail at the target-view scale: as the

camera approaches a previously observed surface, earlier observations may lack the detail required at the closer view. SeaMem therefore further identifies detail-deficient regions within seen areas and selectively relaxes memory anchoring, allowing the NVS model to refine them under historical constraints. Each generated view is subsequently encoded back into memory, continuously updating the scene history for future exploration and viewpoint revisits, without requiring additional fine-tuning of the NVS backbone.

In summary, our contributions are as follows:

- We introduce SeaMem, a seen-aware implicit memory that couples target-view recovery of historical content with explicit estimation of its spatial support. Without constructing an explicit 3D scene representation, SeaMem provides both recovered content and region-specific guidance for preserving historical observations during long-horizon camera-controlled generation.
- We reveal that query-to-memory attention in implicit memory naturally contains region-visibility cues, enabling explicit distinction between seen and unseen regions. This allows seen content to be anchored by memory, while selectively relaxing the anchoring of detail-deficient regions for refinement under historical constraints.
- We evaluate SeaMem on short-range generation and long-horizon revisit evaluation. Across different NVS backbones, SeaMem consistently improves revisit consistency over long horizons while retaining competitive short-range synthesis quality.

## 2 Related Work

**Novel View Synthesis.** Novel view synthesis (NVS) aims to synthesize target views from reference observations and their corresponding camera poses. Scene-specific methods such as NeRF and 3D Gaussian Splatting optimize a representation for each scene (Mildenhall et al., 2020; Kerbl et al., 2023), while generalizable methods directly predict novel views or reconstruct scene representations from sparse observations (Yu et al., 2021; Wang et al., 2021; Charatan et al., 2024; Chen et al., 2024; Jin et al., 2025). Generative NVS methods leverage image and video diffusion priors to synthesize plausible content beyond the coverage of reference views (Liu et al., 2023; Sargent et al., 2024; Voleti et al., 2024; Gao et al., 2024). Recent approaches further incorporate geometric guidance and camera conditioning for controllable scene-level view synthesis (Seo et al., 2024; Yu et al., 2025b; Zhou et al., 2025; Bahmani et al., 2025; Liang et al., 2026). However, these paradigms face distinct challenges when extended to long-horizon world generation. Scene-specific reconstruction methods are designed to represent observed scenes rather than support generative exploration beyond captured content. Reference-conditioned NVS models can synthesize novel views, but conditioning on all accumulated views incurs growing computational and memory costs, whereas restricting the reference set risks discarding observations needed for future revisits. Persistent exploration therefore requires both plausible synthesis beyond observed content and effective management of accumulated observations.

**Explicit 3D Memory for World Generation.** Explicit 3D memory organizes historical observations in a shared 3D coordinate system to provide spatially grounded conditioning for target-view generation. Existing methods maintain depth-lifted point clouds or geometric caches (Ren et al., 2025; Huang et al., 2025; Li et al., 2026), retrieve historical observations from surfel-indexed memory (Li et al., 2025b), or construct updatable Gaussian and spatially organized latent representations (Zhao et al., 2026; Wang et al., 2026). These representations support target-dependent rendering or retrieval and provide an explicit estimate of historical coverage through geometric projection. However, their reliability depends on accurate geometry and camera poses. Even with accurate camera poses, newly generated views may remain geometrically inconsistent with the accumulated scene state. Repeatedly reconstructing or fusing such views can embed these discrepancies into the maintained geometry, causing them to persist through target-view rendering and propagate across subsequent autoregressive updates (Zhao et al., 2026).

**Implicit Context Memory for World Generation.** Implicit context memory preserves historical observations as latent states, cached features, or retrieved observations to condition generation without maintaining an explicit 3D reconstruction. Existing methods retain information beyond the

recent context window by compressing historical frames into compact representations or retrieving relevant past observations to condition subsequent generation (Zhang et al., 2025; Yu et al., 2025a; Xiao et al., 2026; Hong et al., 2025; Zhang et al., 2026). Recent camera-controlled world models further organize historical context to support persistent scene exploration (Wu et al., 2026b; Wei et al., 2026; Yang et al., 2026). While these mechanisms extend access to historical observations, they do not explicitly determine which target-view regions should be constrained by the recovered content. The generator must implicitly infer where historical context should constrain the target view and where novel synthesis is required. Consequently, previously observed content may be unnecessarily resynthesized, potentially introducing appearance or structural drift over autoregressive generation and degrading revisit consistency.

# 3 METHOD

## 3.1 OVERVIEW OF SEAMEM

SeaMem is a seen-aware implicit view memory that preserves historical observations and recovers their content at target viewpoints. As illustrated in Figure 2, SeaMem encodes historical images and camera parameters into compact view representations, which are queried by the target camera to recover previously observed content. Attention statistics from the same query process are leveraged to identify unseen regions. For seen regions, SeaMem further determines where historical observations lack sufficient detail at the target-view scale.

The recovered content and predicted masks provide spatial guidance for a pretrained NVS model. Recovered content anchors seen regions with sufficient detail, while memory anchoring is selectively relaxed in detail-deficient regions to allow refinement under historical constraints. Unseen regions are synthesized by the NVS prior without direct memory anchoring. Each generated view is subsequently encoded back into memory, continuously updating the historical context for future exploration and viewpoint revisits without requiring additional fine-tuning of the NVS backbone. We next describe the implicit view memory for historical content recovery, followed by attention-based estimation of historical support and its use in guiding a frozen NVS model.

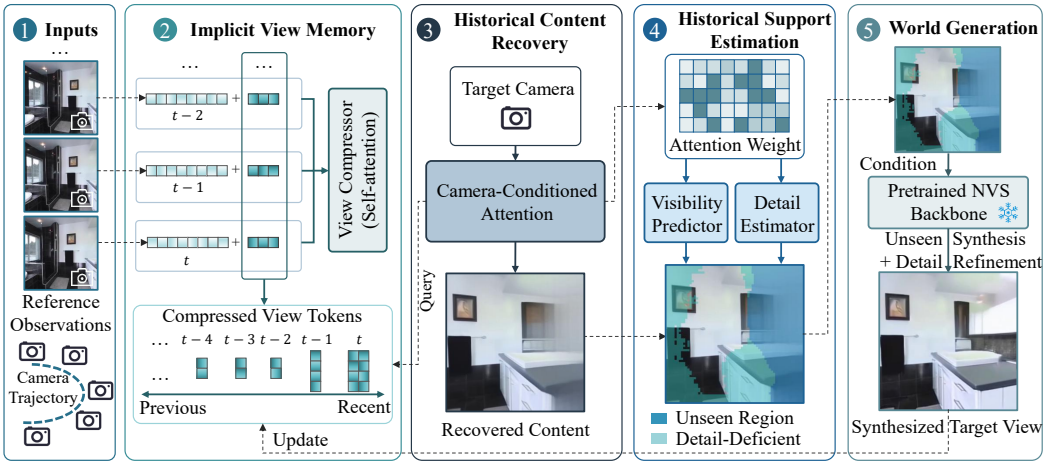



Figure 2: **Overview of SeaMem for long-horizon camera-controlled world generation.** (1–3) Historical observations are compressed into implicit view memory and queried with the target camera to recover historically observed content. (4) Attention statistics from the query process support the prediction of unseen (blue) and detail-deficient (green) regions. (5) The recovered content and region masks guide a pretrained NVS backbone: sufficiently supported seen regions remain anchored to memory, detail-deficient seen regions are refined under historical constraints, and unseen regions are synthesized without direct memory anchoring. Generated views are encoded back into memory for subsequent generation. Colored masks are schematic illustrations of the two region types.

### 3.2 Implicit View Memory for Historical Content Recovery

SeaMem builds upon the camera-conditioned view synthesis architecture of LVSM (Jin et al., 2025) to maintain a compact, queryable memory of historical observations. Each historical image and its camera parameters are first encoded into dense view tokens. A learnable compressor jointly processes the dense view tokens and appended learnable tokens through self-attention, retaining the outputs at the learnable-token positions as a compact latent representation. During training, we randomly select a token budget $L$ and retain the first $L$ compressed tokens (a prefix), enabling each view to be represented under different token budgets.

Given a target camera, SeaMem queries the memory to recover previously observed content at the target viewpoint. Following LVSM, the target camera is encoded as Plücker-ray tokens that interact with historical memory tokens through self-attention. The resulting target tokens are decoded into a recovered image:

$$\hat{I} = R(M, c), \tag{1}$$

where $M$ denotes the historical view memory, $c$ the target camera, and $R$ the camera-conditioned recovery network. During autoregressive generation, SeaMem maintains historical views at different token budgets, retaining more tokens for recent views and representing older views more compactly. Each newly generated view is encoded together with its camera parameters and incorporated into memory for subsequent queries. The updated memory thus retains information from both initial observations and generated views, providing historical context for continued exploration and viewpoint revisits.

### 3.3 Attention-Based Historical Support Estimation

The recovered image $\hat{I}$ does not explicitly indicate where historical observations provide sufficient support for target-view synthesis. SeaMem estimates this support using attention statistics extracted across multiple Transformer layers from the same memory-query process. For each target patch, these statistics characterize its allocation to historical memory, the strength of historical matches, and the spatial distribution of target-side attention. For example, attention mass over historical tokens measures memory allocation, while neighborhood aggregation and entropy characterize the concentration and spread of target-side attention. These complementary statistics form spatial feature maps for estimating historical visibility and detail sufficiency. Complete feature definitions and layer selections are provided in Appendix B.1, with feature-selection and channel ablations reported in Appendix B.3.

To identify unseen regions, a lightweight visibility predictor combines selected attention statistics with memory-budget and camera-relation features. It estimates a seen probability for each target patch, with probabilities below a fixed threshold defining the unseen-region mask. Supervision is constructed offline from VGGT geometry by projecting historical observations into the target view and computing patch-level coverage ratios.

Within seen regions, however, historical coverage does not necessarily imply sufficient detail at the target-view scale. As the camera approaches a previously observed surface, its increased image footprint may require details that were unresolved in earlier observations. SeaMem therefore employs a separate detail predictor that combines selected attention statistics with memory-budget, camera-relation, and recovered-image features. Its supervision is likewise derived from VGGT geometry, using geometric magnification within valid seen regions as a proxy for insufficient historical detail at the target scale. At inference, the predicted detail-deficient mask is restricted to regions classified as seen. The two predictors are trained independently with separate feature normalization; supervision and optimization details are provided in Appendix B.2.

The resulting masks guide a pretrained NVS backbone through region-dependent latent anchoring. We encode $\hat{I}$ into the backbone's latent space and anchor seen regions to the corresponding timestep-matched noisy latent during denoising. In detail-deficient seen regions, anchoring is released earlier, allowing subsequent denoising steps to refine missing details from a historically constrained state. Unseen regions receive no direct memory anchoring and are synthesized using the backbone's native image and camera conditioning. This interface integrates SeaMem with the NVS backbone without requiring additional backbone fine-tuning.

# 4 EXPERIMENTS

## 4.1 SETTINGS

### 4.1.1 TRAINING DATA.

Our training data consists of 65K video clips from the train split of RealEstate10K dataset (Zhou et al., 2018), covering diverse real-world indoor and outdoor scenes, with no overlap with the evaluation set. For each training clip, we construct two types of camera trajectories. The first samples temporally consecutive frames from a randomly selected starting point, producing smooth camera motions with limited viewpoint changes. The second samples frames with randomly varying temporal intervals within the full clip, resulting in larger viewpoint changes and diverse camera motion patterns. Together, these two trajectory constructions expose SeaMem to diverse camera motion patterns, enabling robust long-horizon generation across different exploration trajectories.

### 4.1.2 BENCHMARKS AND METRICS.

We evaluate on RealEstate10K (RE10K) and ACID (Liu et al., 2021) under three settings: Short-Range Smooth Motion, Short-Range Random Motion, and Long-Horizon Revisit. All methods receive the same three initial ground-truth views and target camera trajectories. During rollout, they may update their internal states using generated views but have no access to future ground-truth observations.

Both short-range settings contain 3 reference and 60 target views drawn from the dataset sequences. Smooth Motion uses temporally consecutive views, whereas Random Motion samples views at randomly varying temporal intervals. Generated views are evaluated against the corresponding ground-truth images to measure target-view quality. For long-horizon evaluation, following the cycle-trajectory protocol adopted by recent world-generation methods (Team HunyuanWorld, 2025; Wu et al., 2026a;b), we construct revisit subsets on both datasets. Each rollout follows 120 target camera poses and then traverses the same trajectory in reverse, forming a 240-frame out-and-back sequence. At each camera pose, we compare the generated views from the outgoing and returning passes to measure consistency across repeated visits to the same viewpoint.

We report PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018). Images are center-cropped to a square, with ground-truth images resized to the corresponding output resolution where applicable. Metrics are averaged within each trajectory and then macro-averaged across trajectories. Short-range metrics measure target-view quality, whereas revisit metrics measure consistency between generated views rather than fidelity to ground truth.

### 4.1.3 IMPLEMENTATION DETAILS

We build SeaMem on LVSM using input views at $512 \times 512$ resolution. The compressor uses 512 learnable tokens and supports budgets of 128, 256, and 512 by retaining variable-length prefixes. We train the compression and view-synthesis Transformers while freezing the image and camera tokenizers and the image decoder. Final-stage fine-tuning uses AdamW with learning rate $2 \times 10^{-5}$, weight decay 0.05, and bfloat16 mixed precision. Overall, SeaMem is trained on eight NVIDIA RTX 4090 GPUs for approximately two weeks.

The two lightweight region predictors are trained independently on cached features. Both use supervision derived from VGGT geometry: projected historical coverage supervises unseen-region mask prediction, while geometric magnification provides detail-deficiency labels within valid seen regions. Prediction thresholds are selected on validation data. During generation, both NVS backbones remain frozen, and recovered content provides region-dependent anchoring as described in Section 3.3. Generated views are incorporated into memory after each segment or block. Memory configurations and predictor training details are provided in Appendices A and B, respectively.

### 4.1.4 BASELINES

We compare SeaMem with both its underlying NVS backbones and world-generation methods representing different memory paradigms. SEVA (Zhou et al., 2025) and ViewCrafter (Yu et al., 2025b) are evaluated with and without SeaMem to assess the benefit of memory augmentation across

backbones. For world-generation baselines, LIVE (Huang et al., 2026) represents implicit context memory, conditioning autoregressive generation on recent latent frames, whereas VMem (Li et al., 2025b) and Gen3C (Ren et al., 2025) represent explicit 3D memory through geometrically organized historical observations.

### 4.2 Main Results

We evaluate short-range generation quality and long-horizon revisit consistency quantitatively in Table 1, with qualitative examples of region-guided synthesis and revisit consistency in Figures 3 and 4, respectively.

**Short-Range Generation Quality.** Both SEVA and ViewCrafter provide capable short-range NVS backbones, and SeaMem further improves their target-view quality without additional backbone fine-tuning. As shown in Table 1, SeaMem improves overall target-view quality for both backbones. Notably, integrating SeaMem with SEVA achieves state-of-the-art performance among the compared methods. Figure 3 illustrates how unseen-region masking enables synthesis beyond historical coverage and how relaxing anchoring in detail-deficient seen regions improves local generation quality.

**Long-Horizon Revisit Consistency.** Despite their strong short-range synthesis capabilities, both NVS backbones lag behind existing world-generation methods in long-horizon revisit consistency. Integrating SeaMem substantially improves both backbones on RE10K and ACID, bringing their revisit performance close to that of state-of-the-art world-generation models across both datasets. Notably, on RE10K, the dataset used to train SeaMem, SeaMem + SEVA even surpasses Gen3C in revisit consistency. These results demonstrate that SeaMem effectively extends short-range NVS backbones toward consistent long-horizon world generation without requiring additional backbone fine-tuning. Figure 4 qualitatively demonstrates how SeaMem improves SEVA's preservation of scene layout and local appearance between first visits and revisits at the same camera poses.

Table 1: Quantitative comparison on RE10K and ACID. Short-range settings evaluate generation quality against ground-truth target views; long-horizon revisit measures self-consistency between outgoing and returning generated views. PSNR is reported in dB. Cell colors indicate the best and second best results within each dataset and setting.

| | Short-Range Generation | | | | | | Long-Horizon Revisit | | |
|---|---|---|---|---|---|---|---|---|---|
| | Smooth Motion | | | Random Motion | | | Self-Consistency | | |
| Method | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ |
| **RE10K** | | | | | | | | | |
| ViewCrafter (Yu et al., 2025b) | 17.65 | 0.634 | 0.359 | 13.62 | 0.569 | 0.591 | 11.33 | 0.263 | 0.632 |
| SEVA (Zhou et al., 2025) | 18.92 | 0.690 | 0.349 | 16.01 | 0.611 | 0.463 | 14.61 | 0.533 | 0.490 |
| LIVE (Huang et al., 2026) | 17.71 | 0.556 | 0.337 | 15.17 | 0.480 | 0.434 | 16.64 | 0.525 | 0.420 |
| VMem (Li et al., 2025b) | 16.52 | 0.623 | 0.370 | 13.96 | 0.570 | 0.499 | 14.89 | 0.589 | 0.480 |
| Gen3C (Ren et al., 2025) | 18.82 | 0.682 | 0.349 | 15.58 | 0.601 | 0.461 | 20.97 | 0.760 | 0.284 |
| **SeaMem + ViewCrafter** | 19.32 | 0.691 | 0.353 | 15.89 | 0.609 | 0.468 | 20.79 | 0.745 | 0.307 |
| **SeaMem + SEVA** | 20.39 | 0.726 | 0.310 | 16.22 | 0.613 | 0.439 | 21.17 | 0.743 | 0.281 |
| **ACID** | | | | | | | | | |
| ViewCrafter (Yu et al., 2025b) | 17.87 | 0.522 | 0.433 | 15.03 | 0.478 | 0.556 | 11.74 | 0.277 | 0.651 |
| SEVA (Zhou et al., 2025) | 20.02 | 0.538 | 0.364 | 18.21 | 0.544 | 0.431 | 13.58 | 0.401 | 0.474 |
| LIVE (Huang et al., 2026) | 20.33 | 0.525 | 0.348 | 17.47 | 0.460 | 0.451 | 20.24 | 0.594 | 0.423 |
| VMem (Li et al., 2025b) | 20.54 | 0.598 | 0.269 | 17.62 | 0.564 | 0.430 | 18.56 | 0.560 | 0.344 |
| Gen3C (Ren et al., 2025) | 21.75 | 0.621 | 0.334 | 17.98 | 0.534 | 0.464 | 23.69 | 0.743 | 0.271 |
| **SeaMem + ViewCrafter** | 20.30 | 0.613 | 0.397 | 17.34 | 0.555 | 0.510 | 20.46 | 0.665 | 0.381 |
| **SeaMem + SEVA** | 23.06 | 0.669 | 0.316 | 19.26 | 0.589 | 0.445 | 21.63 | 0.756 | 0.284 |

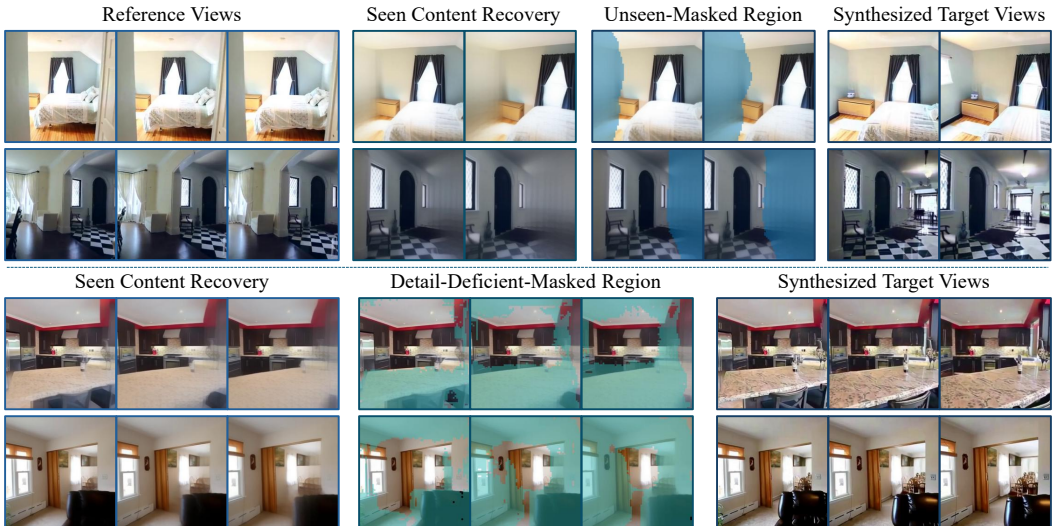



Figure 3: **Qualitative examples of unseen- and detail-deficient-region masking. Top:** Unseen masks identify regions outside reference coverage, where recovery lacks reliable detail. Excluding these regions from anchoring allows NVS to synthesize plausible, detailed content. **Bottom:** Relaxing anchoring in detail-deficient seen regions enables finer details under historical guidance. Masks are shown separately; blurred regions outside the displayed mask may belong to the other category.

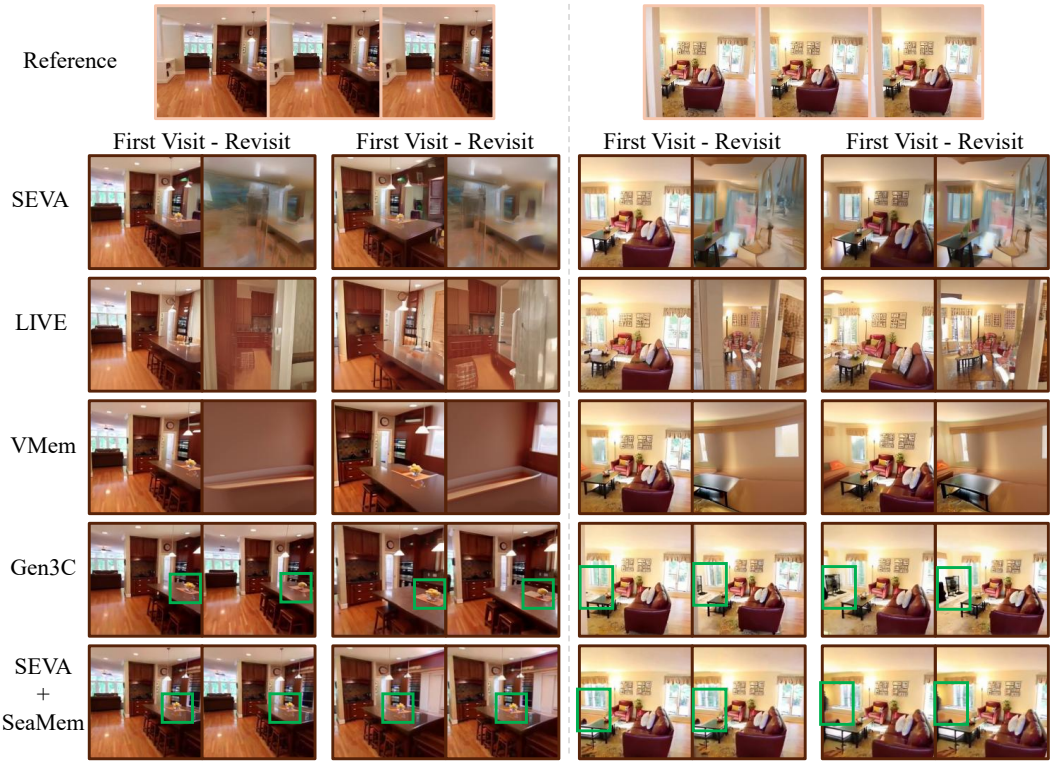



Figure 4: **Qualitative comparison of long-horizon revisit consistency.** Given the same initial reference views, each pair shows a method's generated images at the same camera pose during the first visit (left) and revisit (right). SeaMem substantially improves SEVA's preservation of scene layout and appearance across revisits, achieving consistency competitive with explicit-memory methods. Colored boxes highlight corresponding local regions for closer comparison.

### 4.3 Ablation Studies

Table 2 evaluates SeaMem and its components using frozen SEVA, measuring short-range generation quality under smooth camera motion and long-horizon revisit consistency.

Table 2: Component ablation of SEVA + SeaMem with SEVA frozen in all variants. PSNR is reported in dB. Colors indicate the best and second best results within each setting.

| | Smooth motion | | | Revisit | | |
|---|---|---|---|---|---|---|
| Variant | PSNR↑ | SSIM↑ | LPIPS↓ | PSNR↑ | SSIM↑ | LPIPS↓ |
| SEVA Only (w/o SeaMem) | 18.92 | 0.690 | 0.349 | 14.61 | 0.533 | 0.490 |
| w/o Unseen Mask | 20.15 | 0.714 | 0.322 | 32.79 | 0.945 | 0.072 |
| w/o Detail-Deficient Mask | 19.93 | 0.703 | 0.314 | 21.58 | 0.735 | 0.272 |
| w/o Compressor | 20.47 | 0.741 | 0.297 | *OOM* | *OOM* | *OOM* |
| **SEVA + SeaMem** | 20.39 | 0.726 | 0.310 | 21.17 | 0.743 | 0.281 |

**SeaMem Augmentation.** To evaluate the overall contribution of SeaMem, we remove it and use the original SEVA backbone without memory recovery or region-dependent anchoring. This reduces short-range PSNR from 20.39 to 18.92 dB and revisit PSNR from 21.17 to 14.61 dB, with corresponding degradation in SSIM and LPIPS. These results demonstrate that SeaMem substantially strengthens long-horizon consistency while also improving short-range generation quality, without additional backbone fine-tuning.

**Unseen-Region Masking.** To evaluate the contribution of the unseen-region mask, we next remove it and anchor unseen regions despite their lack of historical support. Although this variant improves revisit consistency, it reduces short-range PSNR from 20.39 to 20.15 dB, with corresponding degradations in SSIM and LPIPS. As shown in Figure 3, recovered content in unseen regions lacks reliable detail, whereas releasing these regions from memory anchoring enables plausible and detailed synthesis. The higher revisit scores therefore reflect the easier reproduction of under-detailed recovered content rather than improved synthesis of unseen regions. These results demonstrate that the unseen-region mask prevents unsupported recovery outputs from constraining novel synthesis.

**Detail-Deficient-Region Masking.** To assess the benefit of detail-deficient-region masking, we disable the detail-deficient-region mask while retaining the unseen-region mask, applying the standard anchoring schedule uniformly to all seen content. Restoring detail-dependent release improves short-range PSNR from 19.93 to 20.39 dB, with higher SSIM and lower LPIPS. Figure 3 illustrates how relaxing anchoring permits finer detail in regions insufficiently supported at the target-view scale. Revisit results are mixed, indicating that this improvement in generation quality does not uniformly improve self-consistency. The detail-deficient-region mask thus complements unseen-region masking by allowing selective refinement within historically observed content.

**Memory Compression.** To examine the trade-off between generation quality and historical-view capacity, we replace compressed historical representations with dense view tokens. On a single RTX 4090 24GB GPU, our recovery-stage capacity test accommodates 1,854 historical views with compression, compared with 142 using dense tokens (Appendix A.2). Compression therefore increases historical-view capacity by approximately $13\times$ at a modest cost to short-range quality, making substantially longer histories accessible for continued exploration and revisits.

## 5 Conclusion

We introduced SeaMem, a seen-aware implicit memory that recovers historical content and derives explicit region masks to guide long-horizon camera-controlled generation. These masks preserve historically supported content while enabling unseen-region synthesis and selective refinement of detail-deficient seen regions. Experiments across two NVS backbones and two datasets demonstrate improved revisit consistency and competitive generation quality without additional backbone fine-tuning.

## AI USE STATEMENT

We used generative AI primarily for language polishing and translation, with additional assistance in checking literature and references and checking manuscript formatting.

## REFERENCES


Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B. Lindell, and Sergey Tulyakov. AC3D: Analyzing and Improving 3D Camera Control in Video Diffusion Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22875–22889, 2025. doi: 10.1109/CVPR52734.2025.02130.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15791–15801, 2025.

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelSplat: 3D Gaussian Splats from Image Pairs for Scalable Generalizable 3D Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19457–19467, 2024. doi: 10.1109/CVPR52733.2024.01840.

Shuo Chen, Cong Wei, Sun Sun, Ping Nie, Kai Zhou, Ge Zhang, Ming-Hsuan Yang, and Wenhu Chen. Context Forcing: Consistent Autoregressive Video Generation with Long Context, 2026.

Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. MVSplat: Efficient 3D Gaussian Splatting from Sparse Multi-View Images. In *Computer Vision – ECCV 2024*, pp. 370–386, 2024. doi: 10.1007/978-3-031-72664-4_21.

Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan Barron, and Ben Poole. CAT3D: Create Anything in 3D with Multi-View Diffusion Models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 75468–75494, 2024.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

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, 2025.

Junchao Huang, Ziyang Ye, Xinting Hu, Tianyu He, Guiyu Zhang, Shaoshuai Shi, Jiang Bian, and Li Jiang. LIVE: Long-horizon interactive video world modeling. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.03747.

Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson W. H. Lau, Wangmeng Zuo, and Chunchao Guo. Voyager: Long-Range and World-Consistent Video Diffusion for Explorable 3D Scene Generation, 2025.

Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. LVSM: A Large View Synthesis Model with Minimal 3D Inductive Bias. In *International Conference on Learning Representations*, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023. doi: 10.1145/3592433.

Guangyuan Li, Bo Li, Jinwei Chen, Xiaobin Hu, Lei Zhao, and Peng-Tao Jiang. MagicWorld: Towards Long-Horizon Stability for Interactive Video World Exploration, 2026.

Jiaqi Li, Junshu Tang, Zhiyong Xu, Longhuang Wu, Yuan Zhou, Shuai Shao, Tianbao Yu, Zhiguo Cao, and Qinglin Lu. Hunyuan-GameCraft: High-Dynamic Interactive Game Video Generation with Hybrid History Condition, 2025a.

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*, pp. 25690–25699, 2025b.

Jinglin Liang, Zijian Zhou, Rui Huang, Shuangping Huang, and Yichen Gong. OrbitNVS: Harnessing Video Diffusion Priors for Novel View Synthesis, 2026.

Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite Nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14458–14467, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Infinite_Nature_Perpetual_View_Generation_of_Natural_Scenes_From_a_ICCV_2021_paper.html.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-Shot One Image to 3D Object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9298–9309, 2023. doi: 10.1109/ICCV51070.2023.00853.

Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, Tong He, Jiangmiao Pang, Yu Qiao, and Kaipeng Zhang. Yume-1.5: A Text-Controlled Interactive World Generation Model, 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. In *European Conference on Computer Vision*, pp. 405–421, 2020. doi: 10.1007/978-3-030-58452-8_24.

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*, pp. 6121–6132, 2025. doi: 10.1109/CVPR52734.2025.00574.

Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, and Jiajun Wu. ZeroNVS: Zero-Shot 360-Degree View Synthesis from a Single Image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9420–9429, 2024. doi: 10.1109/CVPR52733.2024.00900.

Junyoung Seo, Kazumi Fukuda, Takashi Shibuya, Takuya Narihira, Naoki Murata, Shoukang Hu, Chieh-Hsin Lai, Seungryong Kim, and Yuki Mitsufuji. GenWarp: Single Image to Novel Views with Semantic-Preserving Generative Warping. In *Advances in Neural Information Processing Systems*, volume 37, pp. 80220–80243, 2024. doi: 10.52202/079017-2550.

Team HunyuanWorld. HY-World 1.5: A Systematic Framework for Interactive World Modeling with Real-Time Latency and Geometric Consistency. Technical report, 2025. URL https://3d-models.hunyuan.tencent.com/world/world1_5/HYWorld_1.5_Tech_Report.pdf.

Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. SV3D: Novel Multi-View Synthesis and 3D Generation from a Single Image Using Latent Video Diffusion. In *European Conference on Computer Vision*, pp. 439–457, 2024. doi: 10.1007/978-3-031-73232-4_25.

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*, pp. 5294–5306, 2025. doi: 10.1109/CVPR52734.2025.00499.

Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P. Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. IBRNet: Learning Multi-View Image-Based Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2021.

Weijie Wang, Haoyu Zhao, Yifan Yang, Feng Chen, Zeyu Zhang, Yefei He, Zicheng Duan, Donny Y. Chen, Yuqing Yang, and Bohan Zhuang. Latent Spatial Memory for Video World Models, 2026.

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

Zhengxuan Wei, Xu Guo, Xinghui Li, Xunzhi Xiang, Min Wei, Yiran Zhu, Qiulin Wang, Xintao Wang, Pengfei Wan, Xiangwang Hou, and Qi Fan. Geometry-Aware Implicit Memory for Video World Models, 2026.

Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry Forcing: Marrying Video Diffusion and 3D Representation for Consistent World Modeling. In *International Conference on Learning Representations*, 2026a.

Ruiqi Wu, Xuanhua He, Meng Cheng, Tianyu Yang, Yong Zhang, Zhuoliang Kang, Xunliang Cai, Xiaoming Wei, Chunle Guo, Chongyi Li, and Ming-Ming Cheng. Infinite-World: Scaling Interactive World Models to 1000-Frame Horizons via Pose-Free Hierarchical Memory, 2026b.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. WorldMem: Long-Term Consistent World Simulation with Memory, 2026.

Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators, 2024.

Ying Yang, Zhengyao Lv, Tianlin Pan, Haofan Wang, Binxin Yang, Hubery Yin, Chen Li, Ziwei Liu, and Chenyang Si. StableWorld: Towards Stable and Consistent Long Interactive Video Generation, 2026.

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural Radiance Fields from One or Few Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4578–4587, 2021. doi: 10.1109/CVPR46437.2021.00455.

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 *ACM SIGGRAPH Asia Conference Papers*, 2025a.

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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–18, 2025b. doi: 10.1109/TPAMI.2025.3613256.

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*, volume 38, 2025.

Lvmin Zhang, Shengqu Cai, Muyang Li, Chong Zeng, Beijia Lu, Anyi Rao, Song Han, Gordon Wetzstein, and Maneesh Agrawala. TinyHistory: Lightweight Video History Embeddings via Two-Stage Context Learning, 2026.

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*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

Jinjing Zhao, Fangyun Wei, Zhening Liu, Hongyang Zhang, Chang Xu, and Yan Lu. Spatia: Video Generation with Updatable Spatial Memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4245–4257, 2026.

Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable Virtual Camera: Generative View Synthesis with Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12405–12414, 2025. doi: 10.1109/ICCV51701.2025.01153.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Transactions on Graphics*, 37(4):1–12, 2018. doi: 10.1145/3197517.3201323.

# APPENDIX

## A IMPLICIT VIEW MEMORY

### A.1 MEMORY QUERY CONFIGURATION

Memory querying uses view representations with varying token budgets. Each $8 \times 8$ image patch is encoded into a dense token, resulting in 4096 tokens per input view. The compressor jointly processes these dense tokens and a set of appended learnable tokens, retaining the outputs corresponding to the learnable-token positions as the compressed view representation. During generation, an angular field-of-view overlap heuristic selects historical views for each generation segment or block. Selected views are assigned query budgets of 4096, 4096, 512, 256, and 128 tokens from newest to oldest, respectively, capped by the number of tokens available in their stored representations.

### A.2 MEMORY CAPACITY

We test historical-view capacity on a single 24-GB RTX 4090 using bfloat16 and one $512 \times 512$ target query. Pre-encoded tokens from one view and its camera are replicated to vary history length. Dense memory uses 4096 tokens per view; compressed memory uses the query budget schedule in Appendix A.1, with all test views included. The test excludes cache construction, region prediction, NVS generation, and online updates; it does not measure end-to-end rollout capacity.

Table 3: Historical-view capacity on one 24-GB RTX 4090. GPU memory is peak allocation; the final column gives the largest successful tested history. Dense memory at 160 views is marked OOM based on its failure at 143 views.

| | At 160 historical views | | Largest successful |
|---|---|---|---|
| Memory | History tokens | GPU memory (GiB) | history (views) |
| Dense (4096 tokens/view) | 655,360 | OOM | 142 |
| SeaMem compressed | 28,928 | 3.14 | 1,854 |

## B REGION MASK PREDICTORS

### B.1 ARCHITECTURE AND INPUTS

Each predictor uses a $1 \times 1$ input projection to 32 channels, followed by parallel $3 \times 3$ convolutions with dilation rates 1, 2, and 4 and 16 output channels each. The concatenated features pass through $1 \times 1$ projections to 16 channels and one logit, with GELU after every hidden convolution. Visibility and detail have 15,217 and 15,345 parameters and independent weights.

The visibility and detail predictors use 16 and 20 input channels, respectively. Tables 5 and 6 list all channels and their layer selections alongside the ablation results. Layer indices are zero-based; spatial features occupy a $64 \times 64$ target-patch grid, and scalar features are broadcast over this grid.

**Historical attention.** Let $A_{qj}^{\ell,h}$ denote post-softmax attention from target patch $q$ to token $j$ at layer $\ell$ and head $h$. Historical and target tokens form sets $\mathcal{I}$ and $\mathcal{T}$, with $|\mathcal{T}| = 4096$. Define $s_I = \sum_{j\in\mathcal{I}} A_{qj}^{\ell,h}$ and $s_T = \sum_{j\in\mathcal{T}} A_{qj}^{\ell,h}$, so $s_I + s_T = 1$. Statistics are computed per head before arithmetic averaging, denoted $\langle\cdot\rangle_h$, with $\epsilon = 10^{-12}$. Let $\mathcal{K}(q,h,\ell)$ index the largest $k = \min(16, |\mathcal{I}|)$ weights across all historical tokens. We compute

$$
\begin{aligned}
S^\ell(q) &= \langle s_I \rangle_h, & Q^\ell(q) &= \left\langle \frac{1}{k} \sum_{j\in\mathcal{K}} A_{qj}^{\ell,h} \right\rangle_h, \\
R^\ell(q) &= \left\langle \frac{s_T}{\max(s_I, \epsilon)} \right\rangle_h, & F_I^\ell(q) &= \left\langle \frac{\sum_{j\in\mathcal{K}} A_{qj}^{\ell,h}}{\max(s_I, \epsilon)} \right\rangle_h.
\end{aligned}
\tag{2}
$$

Here $S$ is historical attention mass, $Q$ the mean of the strongest historical matches, $R$ the target/history mass ratio, and $F_I$ the fraction of historical mass assigned to those matches. The detail predictor applies $\log(1+R)$ after head averaging. We fix $k = 16$ without a separate hyperparameter search.

**Local target attention.** For the neighborhood $\mathcal{N}_{a:b}(q)$ at Euclidean patch-grid distances $[a, b)$, we compute average attention per neighbor and its fraction of target-side attention:

$$D^{\ell}_{a:b}(q) = \left\langle \frac{\sum_{j\in\mathcal{N}_{a:b}(q)} A^{\ell,h}_{qj}}{\max(|\mathcal{N}_{a:b}(q)|, 1)} \right\rangle_h, \qquad F^{\ell}_{a:b}(q) = \left\langle \frac{\sum_{j\in\mathcal{N}_{a:b}(q)} A^{\ell,h}_{qj}}{\max(s_T, \epsilon)} \right\rangle_h. \tag{3}$$

Neighborhoods are clipped at image boundaries, and $[0, 2)$ includes the query patch. Visibility uses $D_{4:8}$; detail uses both densities and fractions.

**Target-attention distribution.** For $P_j = A^{\ell,h}_{qj} / \max(s_T, \epsilon)$, normalized entropy and its inter-layer change are

$$E^{\ell}(q) = \left\langle -\frac{\sum_{j\in\mathcal{T}} P_j \log\max(P_j, \epsilon)}{\log 4096} \right\rangle_h, \qquad \Delta E^{18}(q) = E^{18}(q) - E^{17}(q). \tag{4}$$

With $\Sigma$ the covariance of target-grid coordinates under $P$, anisotropy is $U^{\ell}(q) = \langle(\lambda_{\max}(\Sigma) - \lambda_{\min}(\Sigma))/\max(\operatorname{tr}\Sigma, \epsilon)\rangle_h$.

**Dense-reference correspondence.** These features use only the two historical views retaining dense tokens. For reference $r$, define $M_{h,r} = \sum_{j\in r} A^{\ell,h}_{qj}$, $P_{h,r}(q, j) = A^{\ell,h}_{qj}/M_{h,r}$, mean reference coordinate $\mu_{h,r}(q) = \sum_{j\in r} P_{h,r}(q, j)c_j$, and normalized weight $w_{h,r} = M_{h,r}/\sum_{r'\in\mathcal{R}_d} M_{h,r'}$. Finite differences give $J_{h,r}(q) = [\partial_x\mu_{h,r}, \partial_y\mu_{h,r}]$, using central differences internally and one-sided differences at boundaries. The encoded minimum singular value is

$$J^{\ell}_{\min}(q) = \log\left(1 + \left\langle \sum_{r\in\mathcal{R}_d} w_{h,r}(q)\, \sigma_{\min}(J_{h,r}(q)) \right\rangle_h\right). \tag{5}$$

Neighboring-attention similarity $C^{\ell}(q)$ uses the same reference weighting and head averaging of cosine similarities between $P_{h,r}(q, \cdot)$ and its valid four-connected neighbors. References with negligible attention mass are excluded. These descriptors characterize attention correspondences, not reconstructed geometry.

**Memory and camera features.** The compressed-memory budget is $b = \sum_{r:n_r<4096} n_r/4096$, where $n_r$ is the retained token count. Relative rotations use $\theta_r = \arccos[\mathrm{clip}((\operatorname{tr}(R_r^T R_t) - 1)/2, -1, 1)]$ for camera-to-world rotations $R_t, R_r$, with the minimum, latest, or mean angle divided by $\pi$. Detail additionally uses $\min_r \|C_t - C_r\|_2$ and $[R^T_{\text{latest}}(C_t - C_{\text{latest}})]_z$: minimum camera-center distance and signed displacement along the latest reference's optical axis. Translations use trajectory-normalized scene coordinates before backend-specific scaling.

**Recovered-image features.** RGB is clamped to $[0, 1]$ and converted to luminance $Y = 0.299R + 0.587G + 0.114B$. Over each $8 \times 8$ patch, we compute mean gradient magnitude with Sobel filters divided by eight and the high-frequency ratio

$$g_{\text{hf}} = \frac{\text{mean}_{\text{patch}}\,|Y - G_1 * Y|}{\text{mean}_{\text{patch}}\,|Y - G_2 * Y| + 10^{-6}}, \tag{6}$$

Gaussian filters $G_1, G_2$ have standard deviations 1 and 2 and kernel sizes 7 and 13. Sobel and Gaussian filtering use reflected padding.

### B.2 Supervision and Training

The following steps construct the two supervision signals from offline VGGT geometry (Wang et al., 2025).

**Historical coverage.** Reference world points are projected into the target camera, excluding invalid or out-of-image points. A nearest-depth buffer and radius-three splatting ($7 \times 7$ max pooling) produce a coverage mask; $8 \times 8$ average pooling gives soft patch targets $v(q)$. Training uses these continuous values. The threshold $v(q) \geq 0.2$ is used only for evaluation and validation-based selection. This coverage construction does not require target-depth agreement.

**Detail deficiency.** Target depth defines the target-to-reference projection $\phi_r(u)$. Valid projections lie inside the reference and agree with its depth within 5%. For $J_r^{\mathrm{geo}} = \partial \phi_r / \partial u$, define

$$s_r(u) = \sqrt{|\det J_r^{\mathrm{geo}}(u)|} = \sqrt{\sigma_1 \sigma_2}, \qquad m(q) = \frac{1}{\max_{r \in \mathcal{R}_{\mathrm{valid}}(q)} \overline{s}_r(q)}, \tag{7}$$

Here $\overline{s}_r(q)$ averages stable pixel scales within patch $q$. Stability requires relative depth gradient at most 0.05, condition number at most 4, and finite scale and minimum singular value above $10^{-6}$. Candidate references must cover at least half the patch with stable valid pixels. Detail labels are $\mathbf{1}[m(q) \geq 1.3]$, restricted to patches with depth-consistent visibility at least 0.8 and finite positive magnification; unseen or invalid patches are excluded from the loss. VGGT labels are computed at $518 \times 518$ and pooled with stride eight to $64 \times 64$, excluding the final six rows and columns. This training-validity mask differs from the predicted seen-region gate used at inference.

**Training data.** Visibility uses 1,000 scenes, half continuous and half random trajectories, with a 400/50/50 train/validation/test split per domain. Detail uses the 500 random-trajectory scenes with the same split. This yields 48,000/6,000/6,000 target examples for visibility and 24,000/3,000/3,000 for detail. Each 63-view sequence supplies three local-reference windows: references 0–2, 3–22, and 23–42 condition targets 3–22, 23–42, and 43–62. The full camera trajectory sets coordinate normalization, but only the specified references provide image content.

**Optimization.** Binary cross-entropy and Dice losses are equally weighted. Visibility uses batch-wise positive weight $\mathrm{clip}(\text{unseen mass}/\text{seen mass}, 0.1, 10)$; detail uses fixed positive weight 5.7611 on valid seen patches. Table 4 summarizes optimization and selection.

Table 4: Predictor training and validation-based selection. Both predictors use a single fixed random seed.

| Setting | Visibility | Detail |
|---|---|---|
| Optimizer | AdamW | |
| Learning rate / weight decay | $10^{-4}$ / $10^{-4}$ | |
| Batch size / maximum steps | 128 / 10,000 | |
| Random seed | 20260731 | |
| Validation interval | 50 steps | |
| Early-stopping patience / minimum gain | 10 checks / 0.002 | |
| Selection metric | Macro unseen $F_2$ | Macro positive IoU |
| Retained checkpoint (step) | 2,200 | 4,650 |

Visibility selection averages over both trajectory domains and all three stages, scanning thresholds from 0.1 to 0.8. Detail selection averages over the three stages, breaking ties by lower false-positive rate and then higher recall.

**Mask construction.** For predicted seen and detail probabilities $p_s, p_d$ and validation-selected thresholds $\tau_s, \tau_d$, the masks are

$$U(q) = \mathbf{1}[p_s(q) < \tau_s], \qquad D(q) = \mathbf{1}[p_s(q) \geq \tau_s]\, \mathbf{1}[p_d(q) \geq \tau_d]. \tag{8}$$

The detail mask is restricted to predicted seen regions. The full model uses the same 20-channel detail predictor in the main experiments, component ablations, and predictor analysis below.

### B.3 Feature Selection and Ablation

We select predictor inputs in two stages, using the same training and validation splits as the final leave-one-channel-out experiments. First, we rank statistic–layer pairs by their individual mask-segmentation performance and retain a broad set of high-ranking candidates. Together with auxiliary

inputs, these form the initial channel set for each predictor. We then perform channel-removal ablations on this set to assess each input's contribution in combination with the others and identify the best-performing subset among the evaluated configurations.

To assess the contribution of each retained input channel, we conduct leave-one-channel-out ablations for both predictors, as reported in Tables 5 and 6. Each variant is retrained with the same data split, initialization source, random seed, and optimization budget as the corresponding full model. Checkpoints and thresholds for the two predictors are selected separately on the validation set: the visibility predictor uses unseen-region $F_2$, while the detail-deficiency predictor uses IoU. The former emphasizes unseen-region recall to reduce erroneous anchoring in unsupported regions, whereas the latter balances missed detail-deficient regions against unnecessary relaxation of anchoring in sufficiently supported regions. The results indicate contributions from both attention statistics and auxiliary inputs.

Table 5: Leave-one-channel-out ablation of the 16-channel visibility predictor. Test unseen $F_2$ is averaged over six domain/stage groups; $\Delta$ is ablated minus full-model $F_2$.

| Removed channel | Test $F_2 \uparrow$ | $\Delta F_2$ |
|---|---|---|
| None (full 16 channels) | 0.8159 | 0.0000 |
| *Attention features* | | |
| Historical attention mass $S^9$ | 0.8072 | -0.0087 |
| Historical attention mass $S^{13}$ | 0.8113 | -0.0045 |
| Historical attention mass $S^5$ | 0.8149 | -0.0009 |
| Historical attention mass $S^{14}$ | 0.8115 | -0.0043 |
| Historical attention mass $S^{10}$ | 0.8106 | -0.0053 |
| Strong-match mean $Q^{16}$ | 0.8118 | -0.0040 |
| Strong-match mean $Q^{15}$ | 0.8156 | -0.0002 |
| Strong-match mean $Q^{18}$ | 0.8150 | -0.0009 |
| Strong-match mean $Q^9$ | 0.8149 | -0.0010 |
| Target-local density $D^9_{4:8}$ | 0.8156 | -0.0003 |
| Target-local density $D^{16}_{4:8}$ | 0.8119 | -0.0039 |
| Target-local density $D^{18}_{4:8}$ | 0.8151 | -0.0007 |
| *Memory and camera features* | | |
| Compressed-memory budget $b$ | 0.8042 | -0.0117 |
| Minimum relative rotation | 0.8130 | -0.0029 |
| Latest relative rotation | 0.8140 | -0.0019 |
| Mean relative rotation | 0.8133 | -0.0026 |

Table 6: Leave-one-channel-out ablation of the 20-channel detail predictor. Test IoU is averaged over three stages within VGGT-valid seen regions; $\Delta$ is ablated minus full-model IoU.

| Removed channel | Test IoU$\uparrow$ | $\Delta$IoU |
|---|---|---|
| None (full 20 channels) | 0.6675 | 0.0000 |
| *Attention features* | | |
| Target-local density $D^{15}_{2:4}$ | 0.6585 | -0.0090 |
| Target/history ratio $\log(1+R^{22})$ | 0.6592 | -0.0083 |
| Target-local fraction $F^9_{0:2}$ | 0.6603 | -0.0072 |
| Target-local fraction $F^9_{4:8}$ | 0.6674 | -0.0001 |
| Target entropy $E^9$ | 0.6594 | -0.0081 |
| Entropy change $\Delta E^{18}$ | 0.6570 | -0.0105 |
| Strong-match mean $Q^{22}$ | 0.6606 | -0.0069 |
| Target anisotropy $U^{17}$ | 0.6568 | -0.0107 |
| Strong-match fraction $F_I^{19}$ | 0.6613 | -0.0062 |
| Correspondence Jacobian $J^{15}_{\min}$ | 0.6485 | -0.0190 |
| Correspondence Jacobian $J^{16}_{\min}$ | 0.6593 | -0.0082 |
| Correspondence Jacobian $J^{18}_{\min}$ | 0.6566 | -0.0109 |
| Neighboring-attention similarity $C^{16}$ | 0.6530 | -0.0145 |
| *Memory and camera features* | | |
| Compressed-memory budget $b$ | 0.6610 | -0.0065 |
| Minimum relative rotation | 0.6623 | -0.0052 |
| Mean relative rotation | 0.6656 | -0.0019 |
| Minimum camera distance | 0.6519 | -0.0156 |
| Latest axial displacement | 0.6548 | -0.0127 |
| *Recovered-image features* | | |
| Recovered-image gradient | 0.6588 | -0.0087 |
| Recovered-image high frequency | 0.6624 | -0.0051 |

## C Supplementary Videos

The supplementary material contains two video folders. `ShortTermQuality` pairs recovered content with predicted masks (blue: unseen; green: detail-deficient) and final synthesis for comparison at matching timestamps. `LongTermRevisit` compares SeaMem + SEVA against SEVA, LIVE, VMem, and Gen3C on the same camera trajectories. Compare first visits with subsequent revisits within each video, and synchronize videos from the same sample folder to assess revisit consistency across methods.