GRIDWM: GRID-LIKE SPATIAL INDEXING FOR LONG-HORIZON WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon world modeling in partially observable environments requires compact memories that preserve spatial structure despite accumulated localization drift. Inspired by the multi-periodic coding of entorhinal grid cells, we propose GridWM, which adopts a layer-wise organization of local temporal RoPE and multi-periodic spatial RoPE, together with content-conditioned address offsets, to capture recent temporal context and long-range spatial structure. A fixed-capacity memory keeps KV storage and per-step attention cost bounded as the sequence grows. On MemoryMaze, GridWM outperforms strong world-model baselines within the training sequence length and generalizes substantially better to prediction horizons far beyond those seen during training, while using a compact model. The temporal-spatial encoding remains effective across deeper Transformer stacks, with consistent gains as depth increases. These results show that combining local temporal context with structured spatial indexing provides an effective inductive bias for efficient long-horizon world modeling.

1 INTRODUCTION

First-person world models must maintain consistency over long interactions by integrating recent temporal context with spatial information from past trajectories.

Recent work suggests that long-term memory is not only a capacity problem but also an addressability problem. MemoryMaze (Pasukonis et al., 2023) and TECO (Yan et al., 2023) highlight the difficulty of maintaining visual consistency over long time spans and repeated scene revisits. CoME (Stapf et al., 2026) separates short-term, long-term, and spatial long-term memory into multiple experts for long-horizon visual generation. Structured memory-addressing mechanisms have also emerged: ReMind (Xu et al., 2026) retrieves historical frames using camera-phase RoPE, ReWorld (Chen et al., 2026) combines a fixed-budget KV cache with a pose-indexed landmark bank, and WorldTrace (Wu et al., 2026) shows that temporal positional encodings beyond the training range can impede access to long-term memories even when they remain cached. Together, these studies suggest that long-horizon world models require structured memory addresses rather than simply more historical tokens.

However, existing approaches typically rely on temporal indices, explicit camera poses, or separate landmark and cache mechanisms. When position is estimated through self-motion integration, localization errors inevitably accumulate. How local temporal structure and long-range spatial addresses should be coordinated within a unified Transformer memory therefore remains underexplored. This motivates a central question: **How should a world model organize recent temporal context and long-range spatial memory under a bounded memory budget?**

The biological nervous system offers an informative organizational principle. Entorhinal grid cells are widely thought to provide multi-scale spatial representations related to path integration and metric space (Hafting et al., 2005; Moser et al., 2008; 2015). Hippocampal representations, meanwhile, encode spatial position, behavioral context, and experiential content rather than merely visual templates of the current location (O'Keefe & Dostrovsky, 1971; Eichenbaum, 2017). Inspired by this hippocampal–entorhinal organization, recent computational models bind structured spatial representations to content (Whittington et al., 2020; Kymn et al., 2024; Zhang et al., 2026; Li et al., 2026b).

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.

Recent positional-encoding methods likewise incorporate structured spatial and temporal information. GridPE (Li et al., 2024) uses grid-cell-inspired Fourier features, nD-RoPE (Li et al., 2026a) introduces multi-scale, multidimensional rotary wave vectors, and multimodal variants such as M-RoPE (Wang et al., 2024) and VideoRoPE (Wei et al., 2025) explicitly encode spatiotemporal coordinates.

These methods primarily study how multiple coordinates are represented, with less attention to how local temporal and long-range spatial structure should be organized in a memory-constrained world model. Temporal continuity favors contiguous recent context, whereas spatial memory requires retrieval over sparse, long-range history. Applying both encodings uniformly in every Transformer layer may therefore underuse their complementary roles.

Building on this observation, we introduce GridWM, a compact hierarchical Transformer for long-horizon visual prediction. GridWM assigns temporal and spatial structure distinct computational roles. Temporal layers use local temporal RoPE to integrate contiguous observations and actions, while spatial layers use grid-cell-inspired multi-periodic spatial RoPE to retrieve long-range history through path-integrated addresses. Because an observation or local experience may contain information associated with multiple locations, GridWM introduces content-conditioned per-head offsets that adapt spatial addresses to the represented content. A fixed-capacity structured memory provides long-range access without unbounded cache growth.

On long-horizon visual prediction in MemoryMaze, GridWM, trained on 320-frame sequences, generalizes reliably to 1,000-frame rollouts and outperforms the baselines both within the training range and far beyond it. Further analysis shows that hierarchical temporal–spatial encoding benefits consistently from increased Transformer depth, while content-conditioned offsets learn scene-aligned spatial adjustments and improve retrieval of relevant history.

2 METHOD

Figure 1 presents GridWM, a hierarchical Transformer that combines local temporal processing with path-integrated, content-conditioned spatial addressing.

2.1 SPATIAL REPRESENTATION ROPE

A causal temporal convolution estimates the motion displacement $\hat{s}_t$ from the action and, when available, the previous visual representation. Architectural and training details are provided in Appendix A.2.

Let the ground-truth head direction be represented by the unit vector $\mathbf{d}_t = [\cos\theta_t, \sin\theta_t]^\top$. Given the predicted displacement $\hat{s}_t$, the corresponding displacement in world coordinates is $\Delta\mathbf{p}_t = \hat{s}_t\mathbf{d}_{t-1}$.

Position is represented using structured multi-period Fourier phases. The model uses p spatial periods spaced logarithmically in scale and w uniformly distributed plane-wave directions for each period, where w is a multiple of three. The wave vectors are organized into groups of three directions separated by 120°, following the directional structure of hexagonal spatial codes. Each plane wave maintains an independent complex phase and corresponding sine/cosine components, yielding $D_r = 2pw$ rotary dimensions.

For spatial period λ_g and base orientation α_g, the three wave vectors in a group are

$$\mathbf{k}_{g,j} = \frac{2\pi}{\lambda_g}\begin{bmatrix}\cos(\alpha_g + 2\pi j/3)\\ \sin(\alpha_g + 2\pi j/3)\end{bmatrix}, \qquad j \in \{0,1,2\}. \tag{1}$$

The corresponding unit-complex phase is updated through path integration:

$$c_{t,g,j} = c_{t-1,g,j}\exp\left(\mathrm{i}\,\mathbf{k}_{g,j}^\top\Delta\mathbf{p}_t\right), \qquad c_{0,g,j} = 1. \tag{2}$$

Using each unit-complex phase to rotate a two-dimensional channel pair of the query and key yields the spatial RoPE. This operation explicitly introduces the relative displacement between estimated positions into the attention computation, allowing the model to exploit spatial relationships directly.

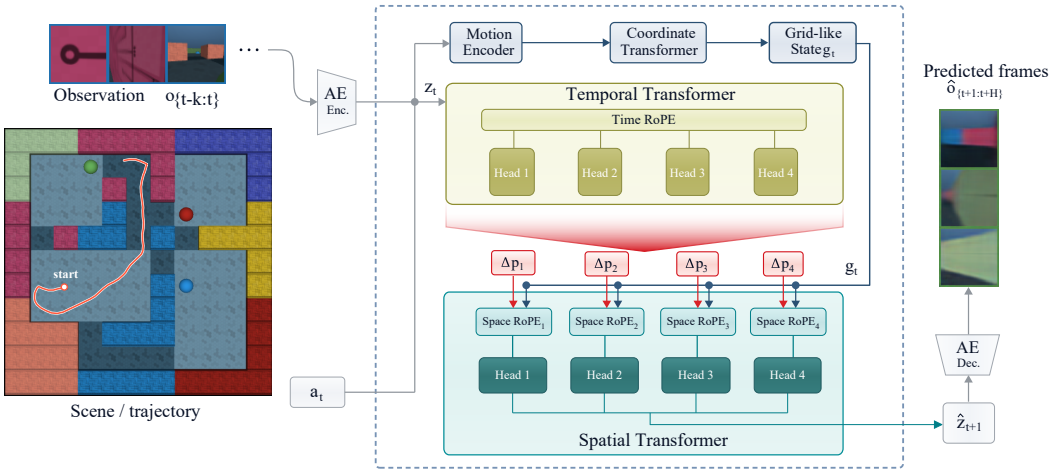

Figure 1: Overview of GridWM. Observation frames are encoded into compact visual representations, while the motion encoder and coordinate transformation update a path-integrated grid-like spatial state. The temporal Transformer first extracts local context using temporal RoPE. Content-conditioned per-head offsets then adjust the grid-like spatial addresses used by the spatial Transformer to retrieve historical experience. The resulting latent representations are decoded autoregressively into future frames.

Combining multiple spatial periods provides position-phase information at multiple scales, supporting both local positional discrimination and disambiguation over larger spatial ranges.

The head direction θ_t is repeated across 16 channel pairs and combined with the position phases above to form the spatial RoPE.

2.2 Hierarchical Spatiotemporal Transformer

Temporal and spatial RoPE are organized hierarchically. The temporal layers extract local contextual information from recent actions and observations, whereas the spatial layers retrieve historical information according to relative spatial relationships. At each step, the model takes the current visual representation $\mathbf{z}_t$ and the next action $\mathbf{a}_{t+1}$ as input and produces a token for predicting $\hat{\mathbf{z}}_{t+1}$. During observation, $\mathbf{z}_t$ is encoded from the ground-truth observation; during open-loop prediction, the model instead uses the predicted representation $\hat{\mathbf{z}}_t$ from the previous step.

Both layer types use standard pre-layer-normalization (Pre-LN) Transformer blocks composed of multi-head self-attention and feed-forward networks, with a residual connection around each sublayer. Each attention module uses four heads, with 256-dimensional query, key, and value representations per head.

The temporal layers combine standard temporal RoPE with sliding-window attention and, by default, attend to the 32 most recent tokens. The spatial layers use a bounded KV cache comprising these 32 recent tokens and up to 96 earlier observation memories. As shown in Appendix A.6, performance is insensitive to the specific memory-selection rule provided that broad temporal coverage is maintained.

Each retained spatial-layer entry already contains the local context extracted by the temporal layers and can therefore carry information beyond that of a single frame.

2.3 Content-Conditioned Spatial Addressing

A single observation frame may contain content associated with multiple physical locations, while a local experience representation aggregates information across multiple frames. To adapt spatial addressing to the represented content, we predict a two-dimensional positional offset for each

spatial-attention head h from the output of the temporal layers x_t^{temp}:

$$\boldsymbol{\delta}_t^h = f_{\text{offset}}^h\big(x_t^{\text{temp}}\big) \in \mathbb{R}^2. \tag{3}$$

Specifically, path integration produces the shared base phase $c_{t,g,j}$ for all spatial-attention heads. The content-conditioned offset then generates a head-specific spatial address through an additional rotation:

$$\widetilde{c}_{t,g,j}^h = c_{t,g,j} \exp\big(\mathrm{i}\, \mathbf{k}_{g,j}^\top \boldsymbol{\delta}_t^h\big)\,. \tag{4}$$

This design allows different attention heads to learn distinct content-conditioned spatial associations for the same experience representation.

2.4 Training Objective

Training is performed on chunks of $K = 16$ frames. After observing each chunk, the model autoregressively predicts the next K frames using the observation memory accumulated up to the current point, together with the actions and head directions from the subsequent chunk. The training objective consists of an image-reconstruction loss, an open-loop prediction loss, a pairwise representation loss, and a teacher–student motion loss. The pairwise loss encourages position representations of visually similar observations with similar head directions to be close, while pushing those of visually dissimilar observations apart. The complete objective is

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{student}} + 0.03\mathcal{L}_{\text{pair}}. \tag{5}$$

Detailed definitions of all losses are provided in Appendix A.

3 Experiments

We collect trajectories in the 9×9 Memory Maze and evaluate open-loop visual prediction conditioned on actions and head directions. Each model observes historical RGB frames, actions, and head directions, and autoregressively predicts future visual sequences from future actions and ground-truth head directions. All methods receive the same ground-truth head-direction information. All models are trained on continuous 320-frame trajectories using a $K = 16$ frame open-loop prediction objective for $30{,}000$ gradient steps; CoME instead retains its original diffusion-based training scheme. All models use a batch size of 16 and are trained over three random seeds. We consider two continuous open-loop evaluation protocols: predicting 64 frames from 256 context frames and predicting 500 frames from 500 context frames.

3.1 Baselines

We compare GridWM with recent world models equipped with spatial or long-term memory mechanisms, as well as general-purpose sequence models with comparable parameter counts. All baselines are adapted to a common observation, action, and head-direction interface, while their original architectures and training procedures are preserved whenever possible.

HPC–MEC. We reimplement and adapt HPC–MEC (Zhang et al., 2026), a hippocampal–entorhinal-inspired world model, using the same head-direction-based coordinate transformation and data interface as GridWM. The resulting model contains approximately 8.15 M parameters.

CoME. We adapt and train CoME on our dataset following its official implementation. To avoid discrepancies arising from different pose-estimation implementations, we provide ground-truth poses directly. CoME has substantially more parameters than GridWM. Appendix E further reports results under a longer training budget; the additional training improves performance but does not change the main conclusions.

General-purpose world models. We construct three latent-space world models based on Transformer, RSSM, and Mamba (Vaswani et al., 2017; Hafner et al., 2019; Gu & Dao, 2024; Hafner et al., 2025). They share the same visual encoder–decoder framework and input and output interfaces. All three models are trained with the same open-loop prediction and image-reconstruction losses, while retaining their architecture-specific auxiliary latent-space objectives. Each model contains approximately 8 M parameters.

4 RESULTS

4.1 LONG-HORIZON VISUAL PREDICTION

Table 1 compares visual prediction performance under the $256 \rightarrow 64$ and $500 \rightarrow 500$ protocols (Wang et al., 2004; Zhang et al., 2018). The $256 \rightarrow 64$ protocol fits within the 320-frame training length but uses a fourfold longer open-loop horizon. GridWM remains robust and reduces LPIPS by 9.4% relative to CoME. The $500 \rightarrow 500$ protocol exceeds the training regime in both total sequence length and prediction horizon. GridWM again achieves the best overall performance and widens its margin over most baselines, with HPC–MEC remaining the closest competitor. CoME remains competitive only in LPIPS, whereas GridWM provides balanced gains in both pixel-level fidelity and perceptual quality.

GridWM's fixed-capacity structured memory keeps KV storage and per-step attention cost bounded as the sequence grows. Appendix A.6 further shows that bounding the spatial cache prevents KV growth with negligible impact on prediction quality.

The qualitative comparison in Figure 2A further illustrates how prediction quality evolves with the rollout horizon. GridWM more consistently preserves scene layout, wall appearance, and viewpoint evolution. CoME produces sharp images, but some long-horizon predictions contain incorrect colors or scene geometry. Figure 2B reports per-step LPIPS under the $256 \rightarrow 64$ protocol. Beyond the 16-step open-loop horizon used during training, GridWM exhibits the slowest error accumulation, and its advantage increases with the rollout horizon. Although CoME achieves low LPIPS during the first few prediction steps, its error subsequently grows more rapidly.

To determine whether these long-horizon gains arise from exploiting additional history, we fix the prediction target to the next 64 frames and progressively increase the available context length (Figure 2C). GridWM's error decreases as the context grows and then remains stable when the context extends beyond the training sequence length, despite accumulated path-integration error and a fixed-capacity KV cache. This behavior indicates a stronger ability to extract useful information from long histories. HPC–MEC and CoME also benefit from additional context, whereas Mamba and RSSM show only limited improvement. Transformer benefits from longer context within the training range, but its performance degrades when the context is extended further.

Finally, we separately evaluate predictions at spatially revisited and spatially novel locations (Figure 2D). A target is revisited if a context observation occurred at least 32 frames earlier within 2 m of its position and with a heading difference of at most 45°; it is novel if every context observation is more than 2 m away. Other targets are excluded. GridWM achieves the lowest MSE on both subsets and outperforms most baselines in LPIPS, although its LPIPS difference from CoME on novel targets is not statistically significant. These results demonstrate robust prediction at both revisited and novel locations under a constrained memory budget.

As a non-comparative adaptation study, we scale GridWM to first-person trajectories derived from HM3D (Ramakrishnan et al., 2021). The model preserves major layouts and viewpoint changes in some rollouts beyond its 16-frame training horizon; Appendix C reports the setup, qualitative predictions, and held-out metrics.

4.2 HIERARCHICAL TEMPORAL–SPATIAL ENCODING ENABLES SCALABLE LONG-HORIZON PREDICTION

Local temporal continuity and long-range spatial relationships exhibit different structures, motivating us to examine whether these two types of positional information should be organized differently within a Transformer. We systematically compare temporal RoPE (T), spatial RoPE (S), within-layer joint encoding (J) (Wang et al., 2024), no positional encoding (N), and different layer-wise temporal–spatial arrangements, evaluating both in-range performance and long-horizon generalization.

All configurations use the same 128-token memory budget. Temporal layers use a contiguous FIFO window, whereas spatial layers use recent-plus-coverage (R+C): 32 recent tokens and 96 earlier memories retained for temporal coverage. Unless cache organization is under study, each layer uses its default policy; content-conditioned spatial offsets are disabled to isolate positional encoding and

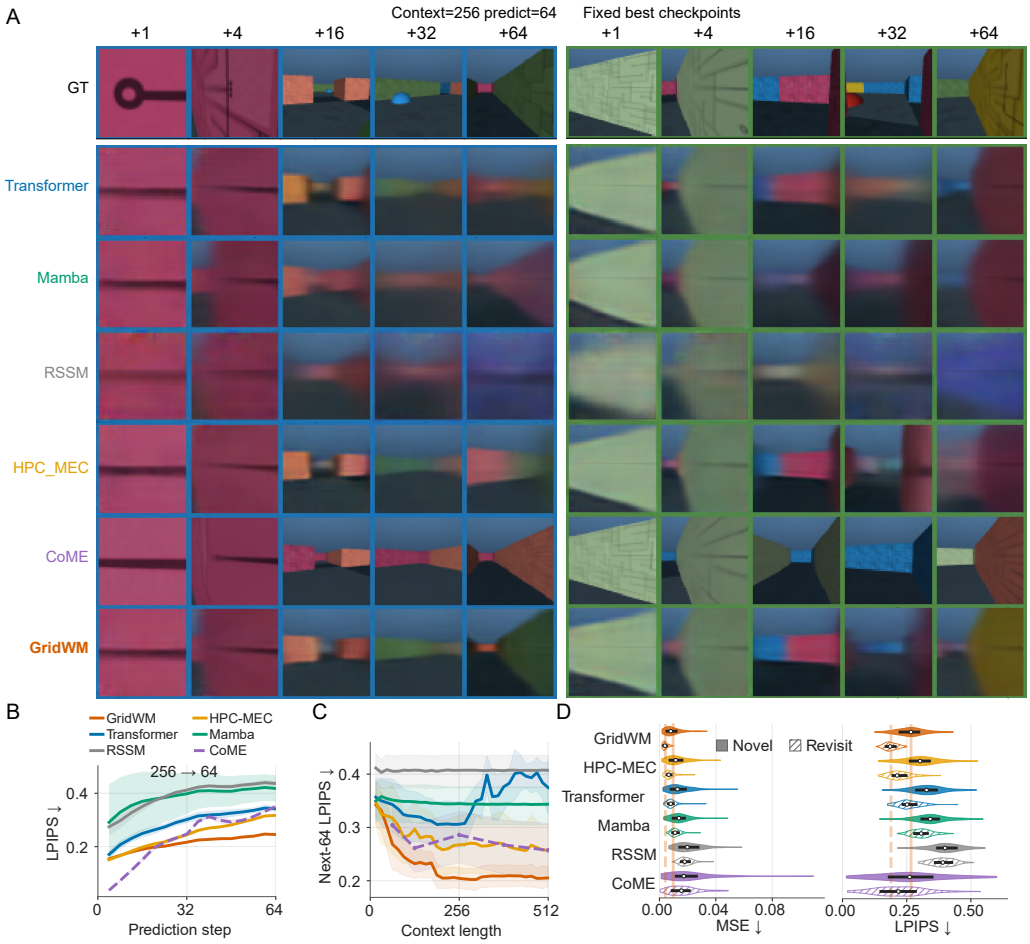

Figure 2: Analysis of long-horizon visual prediction in Memory Maze. **A**: Qualitative predictions at selected rollout horizons under the $256 \rightarrow 64$ protocol. **B**: Per-step LPIPS over the 64-frame prediction horizon. **C**: LPIPS for the next 64 predicted frames as the available context length increases. **D**: MSE and LPIPS on spatially novel and revisited prediction frames.

layer-wise organization. We report LPIPS under the $256 \rightarrow 64$ and $500 \rightarrow 500$ protocols to assess long-horizon generalization within and beyond the training sequence length.

Temporal and spatial positional structures are complementary. Figure 3A compares the main encoding schemes at the same two-layer depth. TT, which uses only temporal RoPE, and NN, which uses no positional encoding, both perform poorly on long-horizon prediction. In contrast, SS, which uses only spatial RoPE, substantially reduces the error under the $500 \rightarrow 500$ protocol, highlighting the importance of explicit spatial indexing for long-range prediction. JJ, which directly combines temporal and spatial signals within each layer, fails to realize the full benefit of both and performs between the purely temporal and purely spatial variants. When temporal and spatial relations are instead assigned to different Transformer layers, both TS and ST achieve stronger overall performance, with TS—temporal contextualization followed by spatial organization—performing better. These results indicate that temporal and spatial positional structures provide complementary information, but the benefit depends on how they are organized within the network; simple within-layer mixing is insufficient to fully exploit their complementarity.

Hierarchical organization consistently benefits from increased depth. We further compare temporal–spatial Transformers with different depths and layer arrangements (Figure 3B). As the network is scaled from 2 to 4 and 6 layers, LPIPS consistently decreases under both the $256 \rightarrow 64$ and

Table 1: Open-loop visual prediction on 9×9 Memory Maze. We report mean $\pm$ sample standard deviation over three training seeds. Best and second-best results are shown in bold and underlined, respectively.

256 context frames → 64 predicted frames				
Model	MSE ↓	SSIM ↑	LPIPS ↓	PSNR ↑
Transformer	0.0100 ± 0.0006	0.7092 ± 0.0094	0.2886 ± 0.0137	22.02 ± 0.47
Mamba	0.0171 ± 0.0060	0.6568 ± 0.0348	0.3873 ± 0.0813	19.01 ± 2.14
RSSM	0.0184 ± 0.0011	0.6269 ± 0.0024	0.3914 ± 0.0035	18.64 ± 0.08
HPC–MEC	$\underline{0.0082 \pm 0.0001}$	$\underline{0.7392 \pm 0.0015}$	0.2447 ± 0.0030	$\underline{23.42 \pm 0.11}$
CoME	0.0171 ± 0.0004	0.5738 ± 0.0017	$\underline{0.2345 \pm 0.0050}$	17.67 ± 0.10
GridWM (ours)	$\mathbf{0.0059 \pm 0.0002}$	$\mathbf{0.7639 \pm 0.0026}$	$\mathbf{0.2124 \pm 0.0027}$	$\mathbf{24.72 \pm 0.16}$
500 context frames → 500 predicted frames				
Model	MSE ↓	SSIM ↑	LPIPS ↓	PSNR ↑
Transformer	0.0281 ± 0.0022	0.5931 ± 0.0097	0.4063 ± 0.0069	16.95 ± 0.33
Mamba	0.0244 ± 0.0003	0.6171 ± 0.0027	0.4354 ± 0.0379	17.02 ± 0.38
RSSM	0.0269 ± 0.0038	0.5752 ± 0.0292	0.4451 ± 0.0160	16.70 ± 0.57
HPC–MEC	$\underline{0.0159 \pm 0.0004}$	$\underline{0.6657 \pm 0.0012}$	0.3573 ± 0.0064	$\underline{19.54 \pm 0.09}$
CoME	0.0263 ± 0.0007	0.4880 ± 0.0059	$\underline{0.3169 \pm 0.0020}$	15.80 ± 0.12
GridWM (ours)	$\mathbf{0.0133 \pm 0.0005}$	$\mathbf{0.7013 \pm 0.0047}$	$\mathbf{0.2920 \pm 0.0037}$	$\mathbf{21.15 \pm 0.17}$

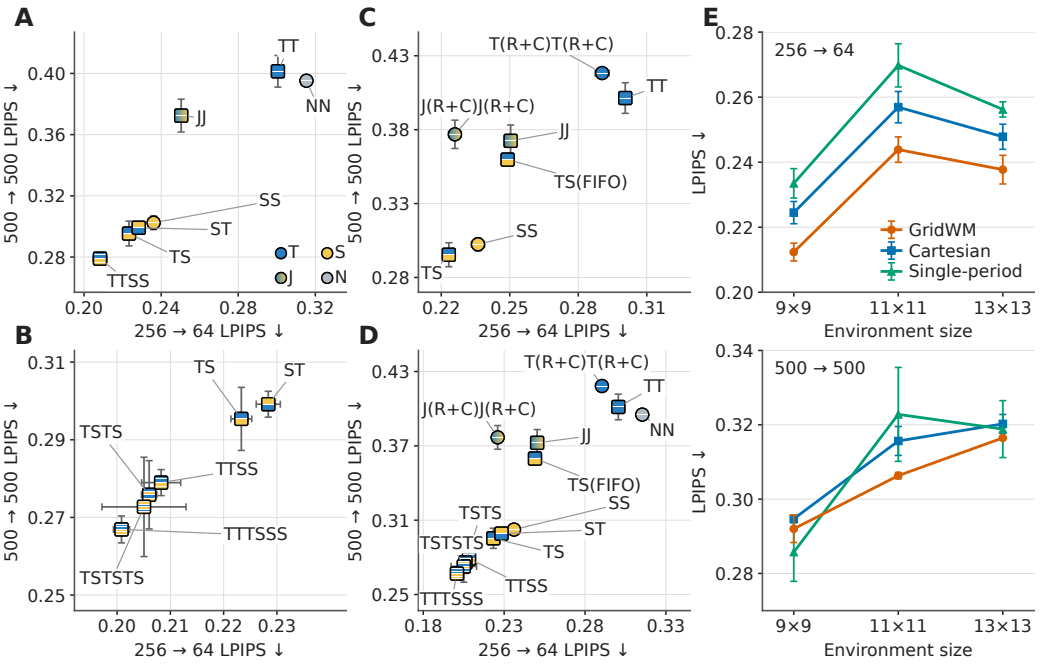

Figure 3: Analysis of hierarchical temporal–spatial encoding. **A**: Comparison of positional encodings in two-layer models. **B**: Temporal–spatial layer arrangements at depths of 2, 4, and 6 layers. **C**: Effects of exchanging the default cache organizations of temporal and spatial layers. **D**: Combined comparison of positional encoding, hierarchy, and cache-organization variants. **E**: Zero-shot generalization of different spatial encodings from the 9×9 training environment to larger environments under the $256 \rightarrow 64$ and $500 \rightarrow 500$ protocols.

$500 \rightarrow 500$ protocols, indicating that the hierarchical temporal–spatial design continues to benefit from increased model depth. At six layers, the blockwise TTTSSS arrangement achieves lower prediction error than the alternating TSTSTS arrangement. At both four and six layers, the blockwise

arrangements (TTSS and TTTSSS) also exhibit smaller variation across random seeds than their alternating counterparts (TSTS and TSTSTS). Within the tested range of two to six layers, these results show that placing local temporal processing before long-range spatial organization provides stable depth scaling.

Temporal and spatial layers benefit from different memory organizations. In the experiments above, temporal layers access a contiguous window of recent history by default, whereas spatial layers combine a recent window with long-term memories selected to provide broad temporal coverage. To examine the effect of this asymmetric design, we further swap the cache organization used by the two layer types (Figure 3C). Applying R+C coverage sampling to temporal layers does not yield consistent improvements and leads to a clear degradation in long-horizon prediction. Likewise, replacing the coverage-based memory of spatial layers with a pure FIFO cache degrades the performance of TS. In contrast, the default configuration—contiguous recent history for temporal layers and long-range coverage for spatial layers—achieves the best overall performance. This asymmetry suggests that, under a limited memory budget, local temporal modeling relies more strongly on contiguous context, whereas spatial retrieval benefits more from broad coverage over long histories. Figure 3D summarizes all variants discussed above for a complete comparison.

Taken together with the spatial-cache organization experiments in Appendix A.6, these results show that spatial indexing allows the model to exploit information spanning long temporal ranges without retaining the full history. Consequently, KV storage and per-step attention cost remain bounded as the sequence grows.

Multi-periodic grid encoding strengthens spatial discrimination and scale transfer. Keeping all other settings unchanged, we compare the multi-periodic grid encoding with Cartesian-coordinate RoPE and a single-period three-direction encoding (Figure 3E). We evaluate zero-shot transfer from the 9×9 training environment to larger 11×11 and 13×13 environments, averaging results over three training seeds. Under the $256\rightarrow64$ protocol, the multi-periodic encoding achieves the lowest LPIPS across all environment sizes, suggesting that its multi-scale phases distinguish physical locations more precisely. Under the more challenging $500\rightarrow500$ setting, the differences among encodings narrow, while the multi-periodic representation remains stable and competitive across environment scales. Overall, its balance of prediction accuracy, spatial-scale transfer, and long-horizon stability motivates its use as GridWM's default spatial representation.

4.3 Learnable Address Offsets Improve Spatial Organization

Temporal layers aggregate visual information across frames, so their contextual representations may not correspond to a single agent-centered position from path integration. GridWM therefore predicts a content-conditioned two-dimensional offset from the temporal output for each spatial-attention head, adapting the base path-integrated address to the represented content.

Learnable address offsets consistently improve visual prediction. We compare the full GridWM with learned address offsets, a No-offset variant trained from scratch with offsets disabled, and three variants obtained by modifying the learned offsets only at test time after normal training: taking their mean, shuffling them, or setting them to zero (Figure 4A). The full GridWM achieves the lowest LPIPS under both the $256\rightarrow64$ and $500\rightarrow500$ protocols, reducing LPIPS by approximately 9.0% and 5.5%, respectively, relative to the No-offset model. All three test-time interventions degrade performance. These results show that content-conditioned offsets improve the learned representations and actively adjust spatial addresses during inference.

The learned address offsets align with visible scene geometry. We measure the distance from both the path-integrated base address and the head-specific adjusted address to the wall intersected by the current viewing direction (Figure 4B). After applying the learned offset, 92.4% of samples move closer to the corresponding wall location, despite receiving no direct supervision of wall positions. Together with the performance degradation under offset interventions in Figure 4A, this result indicates that the offsets are structured, content-dependent spatial adjustments rather than arbitrary perturbations.

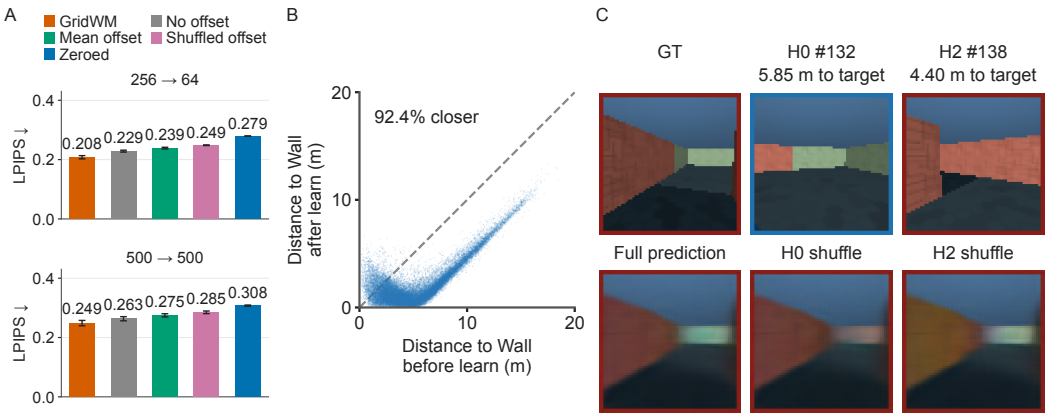

Figure 4: Analysis of content-conditioned learnable address offsets. **A**: LPIPS of the full GridWM, a model trained without address offsets, and three test-time interventions applied after normal training: replacing the learned offsets with their mean, shuffling them, or setting them to zero. **B**: Distance from the base and offset-adjusted spatial addresses to the wall intersected by the current viewing direction. **C**: A qualitative example showing the ground-truth target, high-attention memories retrieved by spatial heads H0 and H2, the full prediction, and predictions obtained after shuffling the retrieval results of either head.

Adjusted spatial addresses retrieve historical visual cues relevant to the current prediction. Finally, we examine how these address adjustments contribute to visual prediction in practice. To reduce the effect of directly retrieving historical views from the same location, we select target frames whose positions are more than 4 m away from all ground-truth context observations. Figure 4C shows a representative example. High-attention memory entries retrieved by spatial heads H0 and H2 are located 5.85 m and 4.40 m from the target position, respectively, and contain partial information about the target wall observed from different viewpoints, while neither contains the complete target view. The full model reconstructs the wall structure and color of the target scene reasonably well, whereas shuffling the retrieval results of either H0 or H2 leads to clear prediction errors. This example suggests that content-conditioned addressing helps the model retrieve prediction-relevant information from spatially distinct historical experiences that provide complementary visual cues, and incorporate these cues into subsequent visual generation.

Overall, these results indicate that the learnable address offsets do more than simply introduce additional degrees of freedom. Instead, they provide content-dependent spatial adjustments to the base positions obtained through path integration. These adjustments yield consistent prediction gains, exhibit interpretable structure aligned with scene geometry, and influence the historical memories actually used by the model, thereby enabling a more flexible binding between temporal context and spatial memory.

5 Conclusion

We introduced GridWM, a hierarchical Transformer that combines temporal contextualization, multi-periodic grid-like spatial indexing, fixed-capacity memory, and content-conditioned address offsets. This design provides efficient access to long-range history without unbounded storage and improves prediction within and beyond the training sequence length.

6 Limitations

GridWM assumes reliable head-direction inputs and is evaluated mainly on MemoryMaze. The non-comparative HM3D study provides only preliminary evidence beyond this setting; broader benchmarks and downstream tasks remain future work.

AI USE STATEMENT

We used generative AI tools to assist with research ideation, literature discovery, drafting of some paper sections, and language refinement. AI tools were used to help explore research directions, identify related work, and improve the clarity and presentation of the manuscript. All AI-assisted content, including generated text and suggestions, was carefully reviewed and verified by the authors. The authors take full responsibility for the final content, claims, experimental results, and conclusions presented in this paper.

REPRODUCIBILITY STATEMENT

We provide implementation details of GridWM, training configurations, and evaluation protocols in the appendix. The source code, configuration files, and scripts required to reproduce the reported experiments are released in the anonymous supplementary repository at `https://anonymous.4open.science/r/GridWM-B30D/`. We also provide detailed descriptions of the dataset construction, baseline implementations, and experimental settings to facilitate reproduction of our results.

REFERENCES

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

Howard Eichenbaum. The role of the hippocampus in navigation is memory. *Journal of neurophysiology*, 117(4):1785–1796, 2017.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*, 2024.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pp. 2555–2565. PMLR, 2019.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

Torkel Hafting, Marianne Fyhn, Sturla Molden, May-Britt Moser, and Edvard I. Moser. Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052):801–806, 2005. doi: 10.1038/nature03721.

Christopher J. Kymn, Sonia Mazelet, Anthony Thomas, Denis Kleyko, E. Paxon Frady, Friedrich T. Sommer, and Bruno A. Olshausen. Binding in hippocampal–entorhinal circuits enables compositionality in cognitive maps. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1235. URL `https://proceedings.neurips.cc/paper_files/paper/2024/hash/4526cfacdbca6b6e184568dac91bf070-Abstract-Conference.html`.

Boyang Li, Yulin Wu, Nuoxian Huang, and Wenjia Zhang. Gridpe: A grid cell-inspired unified position embedding for arbitrary-dimensional spaces. *arXiv preprint arXiv:2406.07049*, 2024.

Boyang Li, Yulin Wu, Sizhe Xu, Nuoxian Huang, Zhonghang Yuan, Shangyi Guo, Shu Yang, and Takahiro Yabe. nd-rope: A generalized rope for n-dimensional position embedding. *arXiv preprint arXiv:2606.12146*, 2026a.

Tiantian Li, Xingxing Cao, Yifei Wang, Xiaojiao Yang, Xiaolong Zou, and Bo Hong. Hippoformer: Integrating hippocampus-inspired spatial memory with transformers. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL `https://openreview.net/forum?id=hxwV5EubAw`.

Edvard I Moser, Emilio Kropff, and May-Britt Moser. Place cells, grid cells, and the brain's spatial representation system. *Annu. Rev. Neurosci.*, 31(1):69–89, 2008.

May-Britt Moser, David C Rowland, and Edvard I Moser. Place cells, grid cells, and memory. *Cold Spring Harbor perspectives in biology*, 7(2):a021808, 2015.

John O'Keefe and Jonathan Dostrovsky. The hippocampus as a spatial map: preliminary evidence from unit activity in the freely-moving rat. *Brain research*, 1971.

Jurgis Pasukonis, Timothy Lillicrap, and Danijar Hafner. Evaluating long-term memory in 3d mazes. In *International Conference on Learning Representations*, 2023.

Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X. Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Sebastian Stapf, Pablo Acuaviva Huertos, Aram Davtyan, and Paolo Favaro. Composition of memory experts for diffusion world models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=sUEdpZCHdp.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

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.

Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Jian Tong, Haodong Duan, Qipeng Guo, Jiaqi Wang, et al. Videorope: What makes for good video rotary position embedding? *arXiv preprint arXiv:2502.05173*, 2025.

James C. R. Whittington, Timothy H. Muller, Shirley Mark, Guifen Chen, Caswell Barry, Neil Burgess, and Timothy E. J. Behrens. The tolman–eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5):1249–1263.e23, 2020. doi: 10.1016/j.cell.2020.10.024.

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

Tianshuo Xu, Yichen Xie, Depu Meng, Chensheng Peng, Quentin Herau, Bo Jiang, Yihan Hu, and Wei Zhan. Teaching video generators to remember: Eliciting dynamic memory for out-of-sight state evolution. *arXiv preprint arXiv:2605.25333*, 2026.

Wilson Yan, Danijar Hafner, Stephen James, and Pieter Abbeel. Temporally consistent transformers for video generation. In *International Conference on Machine Learning*, pp. 39062–39098. PMLR, 2023.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Tianqiu Zhang, Muyang Lyu, Xiao Liu, and Si Wu. Structure abstraction and generalization in a hippocampal–entorhinal inspired world model. In *Forty-Third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=AYXgo5FjYz.

A Implementation and Training Details

A.1 Visual Encoder and Decoder

GridWM uses a convolutional encoder–decoder. The encoder takes a $64\times64\times3$ RGB image as input and consists of four downsampling convolutional blocks. Each block contains a 3×3 convolution with stride 2, followed by GroupNorm and SiLU. The channel dimensions are 32, 64, 128, and 128, respectively.

The resulting $4\times4\times128$ feature map is flattened and passed through a multilayer perceptron (MLP), producing an $S_D = 234$-dimensional visual representation:

$$\mathbf{z}_t = E_\phi(I_t). \tag{6}$$

The decoder first maps the latent representation back to a $4 \times 4 \times 128$ feature map through a linear layer, followed by four stages of nearest-neighbor upsampling and convolution to recover a 64×64 RGB image. A sigmoid activation is applied at the output:

$$\widehat{I}_{t+1} = D_\psi(\widehat{\mathbf{z}}_{t+1})\,. \tag{7}$$

The encoder, predictive model, and decoder are trained jointly from scratch.

A.2 Motion Encoder and Teacher–Student Alignment

The motion encoder uses a causal temporal convolution with kernel size 9 and dilation 1, covering the current input and the preceding eight steps. Given the action $\mathbf{a}_t$ and, when available, the visual observation from the previous step, the observation branch predicts the displacement s_t^{obs}. During open-loop prediction, an action-only branch predicts s_t^{act}.

The observation branch acts as the teacher and the action-only branch as the student. Their outputs are aligned using a smooth L_1 loss, with gradients stopped through the teacher:

$$\mathcal{L}_{\mathrm{student}} = \mathrm{SmoothL1}\big(s_t^{\mathrm{act}}, \mathrm{sg}\big[s_t^{\mathrm{obs}}\big]\big)\,, \tag{8}$$

where $\mathrm{sg}[\cdot]$ denotes the stop-gradient operation.

A.3 Reconstruction and Open-Loop Prediction Losses

The reconstruction loss for an observed frame t is

$$\mathcal{L}^t_{\mathrm{recon}} = \mathrm{MSE}(D_\psi(\mathbf{z}_t)\,, I_t)\,. \tag{9}$$

Let b denote the final observed frame of the current chunk. The model then autoregressively predicts the following $K = 16$ frames, conditioned on the memory accumulated up to frame b. The open-loop prediction loss for this chunk is

$$\mathcal{L}^{(b)}_{\mathrm{pred}} = \frac{1}{K}\sum_{r=1}^{K} \mathrm{MSE}\big(D_\psi\big(\widehat{\mathbf{z}}_{b+r|b}\big)\,, I_{b+r}\big)\,, \tag{10}$$

where $\widehat{\mathbf{z}}_{b+r|b}$ denotes the latent state predicted r steps ahead using observations only up to frame b. Aggregation over frames, chunks, and batch elements is omitted for clarity.

A.4 Pairwise Spatial-Representation Loss

The pairwise loss provides supervision for the path-integration and motion modules. We consider temporal pairs (i, j) satisfying

$$j - i > K, \tag{11}$$

We further require their circular head-direction difference to be at most 15°.

Let V_{ij} denote the similarity between the corresponding visual representations. We define the positive and negative pair sets as

$$\mathcal{P}^+ = \big\{(i,j) : i < j,\ j - i > K,\ \Delta\theta_{ij} \le 15^\circ,\ V_{ij} \ge \tau^+\big\}\,, \tag{12}$$

and

$$\mathcal{P}^- = \left\{(i,j) : i < j,\ j - i > K,\ \Delta\theta_{ij} \le 15^\circ,\ V_{ij} < \tau^-\right\}. \tag{13}$$

We compute the spatial-representation similarity S_{ij} between the two corresponding path-integrated representations. The pairwise loss is

$$\mathcal{L}_{\text{pair}} = \frac{1}{|\mathcal{P}^+|} \sum_{(i,j)\in\mathcal{P}^+} [m_+ - S_{ij}]_+ + \frac{1}{|\mathcal{P}^-|} \sum_{(i,j)\in\mathcal{P}^-} [S_{ij} - m_-]_+, \tag{14}$$

where $[x]_+ = \max(x, 0)$. We use the following thresholds and margins:

$$\tau^+ = 0.98, \quad \tau^- = 0.1, \quad m_+ = 0.9, \quad m_- = 0.7. \tag{15}$$

Positive pairs are encouraged to reach a spatial-representation similarity of at least m_+, while negative pairs are encouraged to remain below m_-. When either $\mathcal{P}^+$ or $\mathcal{P}^-$ is empty, its corresponding loss term is defined as zero.

A.5 Main Model Configuration

The main GridWM model reported in Table 1 contains approximately 7.72 million trainable parameters and uses a two-layer TS Transformer, with one temporal layer followed by one spatial layer. The temporal layer maintains a 32-entry FIFO KV cache.

For each spatial-attention head, the spatial layer applies RoPE to 128 of the 256 query and key dimensions. Positional RoPE occupies 96 dimensions, using eight spatial periods with six rotary channel pairs per period. Head direction occupies a further 32 dimensions by repeating the corresponding rotation across 16 channel pairs. The remaining 128 dimensions are not rotated.

The spatial layer receives the complete output of the temporal layer. For each token, a content-conditioned address network with dimensions $384 \to 128 \to 8$ produces eight scalars, which are reshaped into four two-dimensional position offsets, one for each attention head. These offsets adjust the spatial rotary phases used by the corresponding heads.

A.6 Long-Term Memory Management

By default, GridWM manages its fixed-capacity long-term memory using temporal-redundancy-based eviction. Let the retained entries be ordered by their timestamps. When the long-term memory is full, we assign each interior candidate at timestamp τ_j the temporal-redundancy score

$$C_j = (\tau_j - \tau_{j-1})(\tau_{j+1} - \tau_j)(\tau_{j+1} - \tau_{j-1}). \tag{16}$$

The entry with the smallest C_j is removed, while the two temporal endpoints are always retained. This strategy preferentially removes temporally redundant entries and maintains coverage over a broad temporal range.

Sensitivity to the memory-management strategy. To assess whether GridWM depends on a particular eviction heuristic, we keep the checkpoints trained with the default temporal-coverage strategy fixed and modify only the memory-management policy used at test time. The fixed-interval strategy retains one long-term entry every k frames, with $k = 4$ for the $256 \to 64$ protocol and $k = 8$ for the $500 \to 500$ protocol. The random strategy discards 80% of the candidate entries. As shown in Table 2, prediction quality remains stable despite substantial differences in the number and distribution of retained memories. Relative to the default strategy, MSE changes by at most 4.3% under the $256 \to 64$ protocol and 0.9% under the $500 \to 500$ protocol. These results indicate that GridWM can flexibly exploit its spatial memory and is not strongly tied to a specific eviction rule.

Bounded versus full spatial memory. Beyond changing the cache policy only at test time, we also conduct a from-scratch comparison using the SS model. The full-memory variant uses the complete spatial KV cache during both training and inference, whereas the bounded variant uses recent-plus-coverage (R+C) memory with at most 128 cached entries in both stages. As shown in Table 3, the four image-prediction metrics exhibit negligible degradation overall, with several metrics showing modest improvements. R+C also keeps the peak spatial KV cache at 1.000 MiB under both protocols, whereas the full cache grows from 2.492 to 7.805 MiB as the sequence length increases.

Table 2: Sensitivity to the test-time long-term memory-management strategy. All results use checkpoints trained with the default temporal-coverage strategy. The number of long-term entries is averaged over evaluation sequences.

Protocol	Strategy	Long-term entries	MSE ↓	SSIM ↑	LPIPS ↓	PSNR ↑
$256 \rightarrow 64$	Temporal coverage (default)	96	0.006216	0.7587	0.2205	24.400
	Fixed interval ($k = 4$)	56	0.006142	0.7582	0.2218	24.323
	Random eviction (80%)	43.56	0.006480	0.7531	0.2294	24.095
$500 \rightarrow 500$	Temporal coverage (default)	96	0.013600	0.6989	0.2926	21.166
	Fixed interval ($k = 8$)	59	0.013712	0.6961	0.2973	21.037
	Random eviction (80%)	89.58	0.013623	0.6985	0.2945	21.100

Thus, explicit spatial memory management bounds KV storage without sacrificing prediction quality, highlighting its empirical importance for long-horizon prediction. We further hypothesize that such sparse memory may help mitigate memory interference caused by accumulated path-integration errors, although directly validating this effect remains future work.

Table 3: From-scratch comparison of full and bounded spatial KV caches using the SS model. R+C retains at most 128 entries during both training and inference. Peak spatial KV-cache memory is reported in MiB.

Protocol	Training / inference cache	Peak KV (MiB) ↓	MSE ↓	SSIM ↑	LPIPS ↓	PSNR ↑
$256 \rightarrow 64$	Full / Full	2.492	0.006856	0.747846	0.234693	23.844
	R+C / R+C	1.000	0.006602	0.749573	0.231666	23.922
$500 \rightarrow 500$	Full / Full	7.805	0.014121	0.689286	0.307988	20.648
	R+C / R+C	1.000	0.014261	0.689550	0.304276	20.582

B Dataset Details

We collect $2{,}368$ training episodes and 112 test episodes in Memory Maze. Each episode is generated from an environment initialized with a different random seed and contains $1{,}000$ frames. Figure 6 shows representative layouts and sampled trajectories for the 9×9, 11×11, and 13×13 Memory Maze environments.

In the training set, 4.4% of actions result in wall collisions, leading to relatively large discrepancies between the intended action-induced motion and the actual displacement. Consequently, path-integration errors are difficult to avoid over long trajectories, testing the model's ability to integrate spatial memories and resolve ambiguity as the underlying spatial representation drifts.

Figure 5 compares the path-integration error decoded from the main GridWM model on our collected dataset and the official dataset. For both datasets, we decode the estimated position from the largest-period component of the grid-like representation and measure its deviation from the ground-truth position over the first 500 observed frames. Although the error accumulates with context length in both settings, the drift is substantially more pronounced on our collected trajectories, making long-horizon visual prediction more challenging. The multi-periodic representation, content-conditioned offsets, and visual content inputs provide complementary mechanisms that may partially mitigate the resulting performance loss by helping disambiguate drifting spatial states, although they do not eliminate the underlying localization error.

The trajectories also contain frequent revisits: for 70.98% of frames, the agent had previously visited a location within 1 m of its current position more than 32 frames earlier. This setting therefore places stronger demands on how models organize and retrieve spatial information.

Unless otherwise specified, we report experiments on our collected 9×9 Memory Maze dataset. All experiments in Section 4.3 instead use the official Memory Maze dataset. Its less severe path-

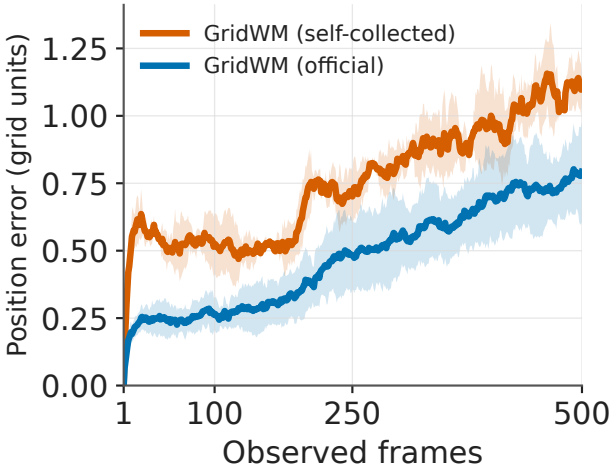

Figure 5: Accumulation of path-integration error in the main GridWM model on our collected and the official datasets. Position is decoded from the largest-period component of the grid-like representation and compared with the ground-truth position over the first 500 observed frames. The larger error on our collected trajectories indicates more severe localization drift and a more challenging setting for long-horizon visual prediction.

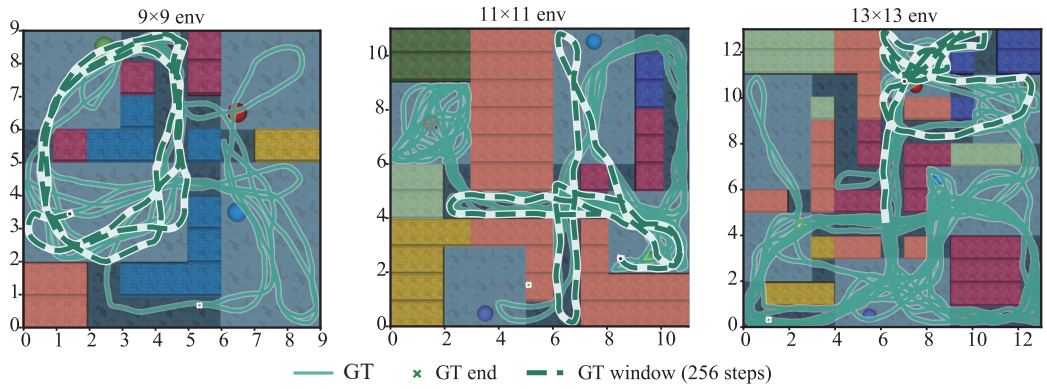

Figure 6: Representative layouts and sampled trajectories in 9×9, 11×11, and 13×13 Memory Maze environments. The dashed segment marks a 256-step ground-truth window, and the cross marks its endpoint.

integration drift provides a cleaner setting for isolating the effect of the learnable address offsets and also results in lower overall prediction error.

C Adaptation to Complex Indoor Environments

To assess GridWM's adaptability to more complex indoor environments, we conduct additional experiments on a first-person trajectory dataset built from HM3D. Visual observations are represented using the pretrained Stable Diffusion VAE (Rombach et al., 2022) `sd-vae-ft-mse`. We freeze its encoder and decoder and train only the latent-space prediction model. The model uses a TTSS layer arrangement with eight attention heads and a head dimension of 512, resulting in 106.49 million trainable parameters; the spatial RoPE design remains unchanged. Ground-truth x–z positions and headings are provided during both observation and prediction, while future RGB frames are never given to the predictor. Training uses 16-step autoregressive rollouts with latent-space MSE and runs for 100,000 gradient updates.

Figure 7 shows predictions across multiple test scenes. GridWM captures the major spatial layouts and viewpoint changes, although fine details remain blurrier than the frozen VAE reconstructions. When rollouts extend beyond the 16-frame training horizon to 32 and 64 frames, recognizable structure persists in some scenes. These results provide preliminary evidence that GridWM can adapt to complex indoor environments, while fine-grained appearance modeling and long-horizon fidelity remain open challenges. The retained local structure may reflect the model's use of historical observations.

Table 4 reports image-quality metrics on an independent test set comprising 20 episodes from 10 HM3D minival scenes. A protocol $X \rightarrow Y$ denotes X observed frames followed by Y autoregressively predicted frames.

D Baseline Implementation Details

D.1 HPC–MEC

HPC–MEC uses the same visual encoding framework and motion-processing module as GridWM. The main differences lie in the MEC component and memory mechanism, both of which follow the design of the original method.

For the MEC component, we compare two variants: *(i)* continuously propagating the MEC state within each chunk, and *(ii)* re-inferring the MEC state from the HPC state at every step. Based on a preliminary single-seed experiment, we use the second variant in the reported experiments. During open-loop prediction, the HPC state is driven by the model's predicted visual representations and does not access future ground-truth images.

D.2 CoME

The main text reports CoME under a training budget approximately matched to that of GridWM. Specifically, ImageVAE, Base, STM, and Spatial Base are trained for $20{,}000$, $10{,}000$, $30{,}000$, and $30{,}000$ gradient steps, respectively. Following the original setup, ImageVAE is frozen after training. LTM and SLTM perform online LoRA adaptation using the observed context of each trajectory.

Extended-training results are reported in Appendix E. Under the $500 \rightarrow 500$ evaluation protocol, evaluating a single sequence takes approximately 278 seconds.

D.3 General-Purpose World Models

Transformer, RSSM, and Mamba take the previous visual latent, the current transition action, and the ground-truth head direction at the target time as inputs. Each model updates its latent state and autoregressively predicts the subsequent visual latents.

The Transformer baseline is a two-layer causal model with eight attention heads and a hidden width of 480. It uses learnable absolute positional embeddings along the temporal axis. The Mamba baseline uses three Mamba-1 layers with a hidden width of 464. The RSSM baseline uses a 768-dimensional deterministic state together with a 234-dimensional continuous Gaussian latent state. All three models contain approximately 8 million parameters.

Because learnable absolute positions may be unfavorable for extrapolation beyond the training sequence length, we additionally consider a more controlled relative-position comparison. The TT configuration in Figure 3A is a two-layer temporal-only Transformer using relative temporal RoPE and the same 128-token memory budget as the other positional-encoding variants. Despite removing the absolute-position extrapolation issue, TT still performs poorly under long-horizon evaluation, particularly the $500 \rightarrow 500$ protocol. Thus, replacing absolute embeddings with temporal RoPE alone does not yield strong long-horizon generalization; the gains of GridWM additionally rely on explicit spatial indexing and layer-wise temporal–spatial organization.

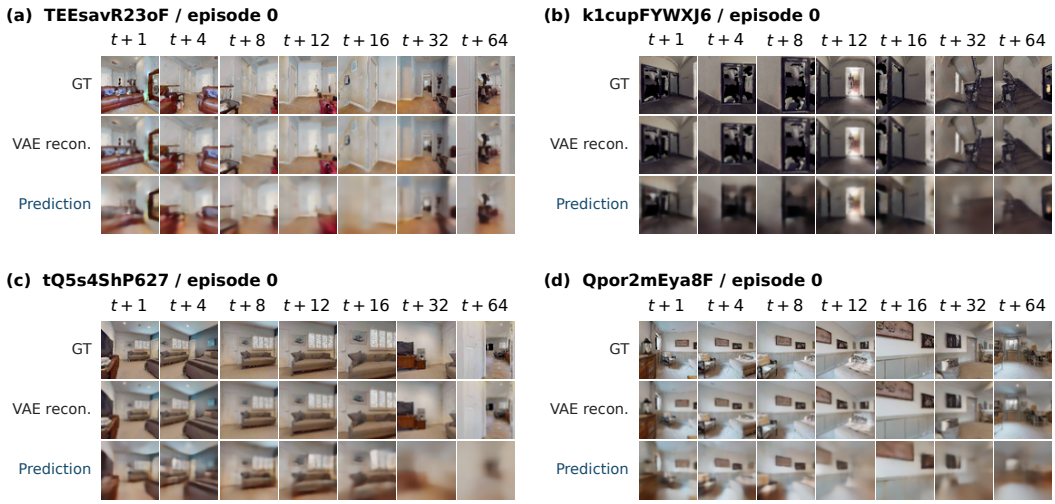

Figure 7: Qualitative predictions on HM3D-based trajectories. GridWM preserves major scene structure and viewpoint evolution across multiple test scenes. The 32- and 64-frame predictions extend beyond the 16-frame training horizon; frozen-VAE reconstructions are shown as a reference for representation quality.

Table 4: Open-loop prediction performance on the HM3D-based test set.

Protocol	MSE ↓	LPIPS ↓	SSIM ↑	PSNR ↑
$256 \rightarrow 16$	0.011407	0.4480	0.6010	21.47
$256 \rightarrow 64$	0.011929	0.4624	0.5971	21.37

E Extended Baseline Experiments

E.1 HPC–MEC

The original HPC–MEC formulation combines observation feedback with recurrent MEC-state propagation. Our setting instead requires continuous open-loop prediction after a fixed observation history, making it necessary to specify how the MEC state is updated when no new image observations are available. We therefore compare two update protocols: CI-R, which preserves recurrent MEC-state propagation, and SI, which re-estimates the MEC state at every step from the current HPC state. Under SI, the HPC state is updated autoregressively using the model's own predicted visual representations and never accesses future ground-truth images.

Both variants are trained from scratch for 10,000 gradient steps using identical training configurations and the same random seed. As shown in Table 5, SI achieves lower MSE under both the short- and long-horizon protocols and consistently improves the long-horizon perceptual metrics. We therefore adopt SI in the main experiments, while reporting CI-R to make the effect of this implementation choice explicit.

Table 5: Comparison of HPC–MEC update protocols. Both variants are trained from scratch for 10,000 gradient steps using identical configurations and the same random seed. LPIPS, SSIM, and PSNR are reported under the $500 \rightarrow 500$ protocol.

MEC update	MSE ($256 \rightarrow 64$) ↓	MSE ($500 \rightarrow 500$) ↓	LPIPS ↓	SSIM ↑	PSNR ↑
CI-R	0.012053	0.018827	0.410430	0.647472	17.2523
SI	**0.009785**	**0.015406**	**0.369950**	**0.662986**	**18.1231**

E.2 CoME

The original CoME study uses a substantially larger training budget, with up to one million gradient steps. To examine the effect of additional optimization, we extend the training of Base, STM, and Spatial Base to $30{,}000$, $50{,}000$, and $50{,}000$ gradient steps, respectively, while keeping the remaining setup unchanged. We report results under the original $200 \rightarrow 17$ protocol as well as the $256 \rightarrow 64$ and $500 \rightarrow 500$ protocols used in our main evaluation.

As shown in Table 6, the extended budget improves all four prediction metrics across the three protocols. Nevertheless, CoME does not surpass GridWM under either of the main evaluation protocols. CoME also contains more than 293 million trainable parameters, substantially more than GridWM. Thus, even with additional optimization and a much larger model, the performance gap is not closed, indicating lower training efficiency than GridWM.

Table 6: CoME performance under the matched and extended training budgets. The extended setting trains Base, STM, and Spatial Base for $30{,}000$, $50{,}000$, and $50{,}000$ gradient steps, respectively.

Protocol	Training setting	MSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓
$200 \rightarrow 17$	Matched budget	0.006815	21.672	0.6869	0.1143
	Extended training	**0.005252**	**22.797**	**0.7257**	**0.0927**
$256 \rightarrow 64$	Matched budget	0.017093	17.672	0.5738	0.2345
	Extended training	**0.015293**	**18.155**	**0.5954**	**0.2150**
$500 \rightarrow 500$	Matched budget	0.026328	15.797	0.4880	0.3169
	Extended training	**0.025996**	**15.851**	**0.4909**	**0.3120**

E.3 Transformer with Spatial Content Inputs

To determine whether GridWM benefits merely from access to a grid-like spatial representation, rather than from using it through spatial RoPE, we train two additional Transformer variants. HEX96 + Full injects the same 96-dimensional, path-integrated grid-like representation used by GridWM into the token content, whereas XY + Full injects the path-integrated Cartesian coordinates directly. Both variants use relative temporal RoPE and a full KV cache, following a standard Transformer setting and providing access to the complete history. All results are averaged over three training seeds.

Table 7 shows that spatial content inputs provide modest improvements in MSE and short-horizon LPIPS over the original Transformer. However, neither variant improves long-horizon LPIPS, and both remain substantially behind GridWM under both protocols. Thus, simply adding grid-like or Cartesian position information to token content is insufficient. The grid-like representation is most effective when used as a rotary spatial address that directly structures attention.

Table 7: Comparison of spatial RoPE with spatial representations injected as Transformer content. HEX96 + Full and XY + Full use relative temporal RoPE and a full KV cache. Values are means over three training seeds.

Model	MSE ($256 \rightarrow 64$) ↓	LPIPS ($256 \rightarrow 64$) ↓	MSE ($500 \rightarrow 500$) ↓	LPIPS ($500 \rightarrow 500$) ↓
GridWM (ours)	**0.005850**	**0.212404**	**0.013286**	**0.292006**
Transformer	0.010017	0.288581	0.028135	0.406308
HEX96 + Full	0.009399	0.280305	0.025517	0.411359
XY + Full	0.009054	0.276214	0.025456	0.418372