VISUAL MEMORY FIELDS: A BRAIN-INSPIRED SPATIAL SCAFFOLD FOR EMBODIED MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Embodied agents need memory that can integrate growing streams of local, viewpoint-dependent observations into a persistent representation under finite capacity. Existing systems often trade off bounded storage against long-horizon retention and flexible reuse. Inspired by hippocampal-entorhinal systems, we propose Visual Memory Fields, which bind episode-specific sensory content to a reusable spatial scaffold represented as a continuously addressable field over position and heading. We instantiate this scaffold with a sparse Fourier basis shared across different maze layouts, and update episode-specific content online through compact sufficient statistics rather than storing observation history. A fixed Fourier support keeps the sufficient statistics—and therefore persistent memory—independent of experience length, while budget-constrained search allocates the available capacity across pose frequencies. This bounded representation can then be queried and reused after accumulation, rather than being tied to the objective under which experience was collected. We test this property in Memory Maze, first through view-centric recall over long histories and randomly withheld views, and then through scene-centric object localization using a separately supervised reader on the frozen field. We further show that the resulting scene belief can be re-expressed in candidate action frames for downstream control. Together, these results demonstrate Visual Memory Fields as a persistent, bounded, and reusable form of scaffolded embodied memory.

1 INTRODUCTION

An embodied agent sees an environment one local view at a time, but may learn only later what that experience will be needed for. A corridor passed during exploration may later need to be pictured from another heading; an object may become a goal after it has left sight. Memory must therefore preserve visual experience for questions that were not fully specified when the observations arrived, while its persistent state remains finite as experience accumulates.

Existing memory systems resolve this tension in different ways. Observation archives preserve direct access to the past, but grow with the trajectory. Bounded recurrent states avoid that growth, but what they retain is usually shaped by the readout for which they are trained. Biological memory suggests a different organization. Hippocampal–entorhinal accounts distinguish reusable relational structure from episode-specific sensory associations (Whittington et al., 2020; Chandra et al., 2025), while spatial-memory models emphasize that remembered content can be re-expressed across reference frames to support viewpoint-dependent imagery (Byrne et al., 2007; Bicanski & Burgess, 2018).

Drawing on biological accounts of memory, we consider a computational principle in which reusable structure is separated from episode-specific content, and use it to define a different memory organization: a shared spatial address system, online binding of episode-specific visual content to those addresses, and task-specific readers that can be attached after memory accumulation. The address system determines where a view is written and queried; it need not determine what a later task asks of the stored content. Our central hypothesis is that spatial addressing can be separated from both the content accumulated within an episode and the task-specific interpretation applied afterward.

We instantiate this organization as Visual Memory Fields (VMF). A sparse Fourier scaffold maps position and heading to continuous addresses and determines how evidence is shared across poses. As

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.

observations arrive, VMF accumulates pose moments and visual cross-statistics that define a pose-indexed regression field, without retaining the observation stream itself. Fourier difference moments compactly encode the regression geometry, while a fixed feature support keeps the persistent state independent of history length. The learned Fourier scaffold shapes how visual evidence is shared across positions and headings, with its support selected under a fixed memory budget. The resulting memory is bounded and continuously queryable; we then test whether its accumulated content can support tasks that were not specified during writing.

We test this organization in Memory Maze (Pasukonis et al., 2023), asking first whether early views remain accessible after later writes and whether the field can answer queries at unwritten viewpoints. The accumulated state is then frozen and evaluated through two distinct readouts: view-centric visual recall and scene-centric object localization with a separately supervised reader. Last-window and geometry-only controls test whether localization depends on retained visual content rather than recent observations or pose information alone. Finally, a frozen-field control study re-expresses the inferred scene belief in candidate action frames, testing whether the same stored experience can support action-centric control.

Our contribution is a brain-inspired scaffolded memory architecture that separates memory accumulation from downstream interpretation under a fixed persistent-state budget. VMF turns this principle into a testable organization in which a shared spatial structure can be learned once, populated online by episode-specific experience, and subsequently interpreted without rewriting the memory itself. More broadly, VMF treats memory capacity not only as a limit on what can be retained, but as a constraint on how experience is organized for later reuse.

2 Related work

Shared structure and episode content. The Tolman–Eichenbaum Machine separates reusable relational structure from sensory associations, while Vector-HaSH binds content to a spatial scaffold (Whittington et al., 2020; Chandra et al., 2025). VMF turns this organization into a supplied-pose Fourier kernel with episode-wise regression. Reference-frame models motivate its dependence on both position and heading (Byrne et al., 2007; Bicanski & Burgess, 2018).

Embodied memories and later readouts. Neural Map writes to a spatial tensor, and Scene Memory Transformer stores observation embeddings and attends over them for policy decisions(Parisotto & Salakhutdinov, 2018; Fang et al., 2019). Predictive and recurrent memories also support downstream control (Wayne et al., 2018; Samsami et al., 2024). VMF instead tests whether one bounded visual field, accumulated without object labels, can serve a separately trained object reader after the episode state is frozen. Its recall and localization experiments distinguish access to old visual content from access to geometry or a recent window.

Queryable fields under a budget. Fourier feature maps provide a natural continuous parameterization over coordinates(Rahimi & Recht, 2007; Sinha & Duchi, 2016), while DeepTime performs sequence-specific fitting of a continuous time-indexed predictor(Woo et al., 2023). VMF applies regression to pose-indexed visual content, retains Fourier difference moments in place of a full Gram matrix, and learns the frequency support within the persistent-state budget. Fixed-dimensional statistics yield the history-independent state size; support learning allocates that capacity rather than creating the bound. Further architectural comparisons appear in Appendix B.

3 Visual Memory Fields

VMF factorizes memory into a spatial address scaffold shared across episodes and environments within a task domain, an episode-specific writer state, and downstream readers of the accumulated field(Fig. 1). The visual codec and heading alignment are learned offline; scaffold learning starts from the zero-frequency (DC) mode and adds Fourier modes under the persistent-state budget, after which all learned shared parameters are frozen. During an episode, only the episode-specific state is updated by real observations, while downstream readers query the resulting field through the frozen scaffold. Here t counts observations, so N actions yield $N + 1$ observations.

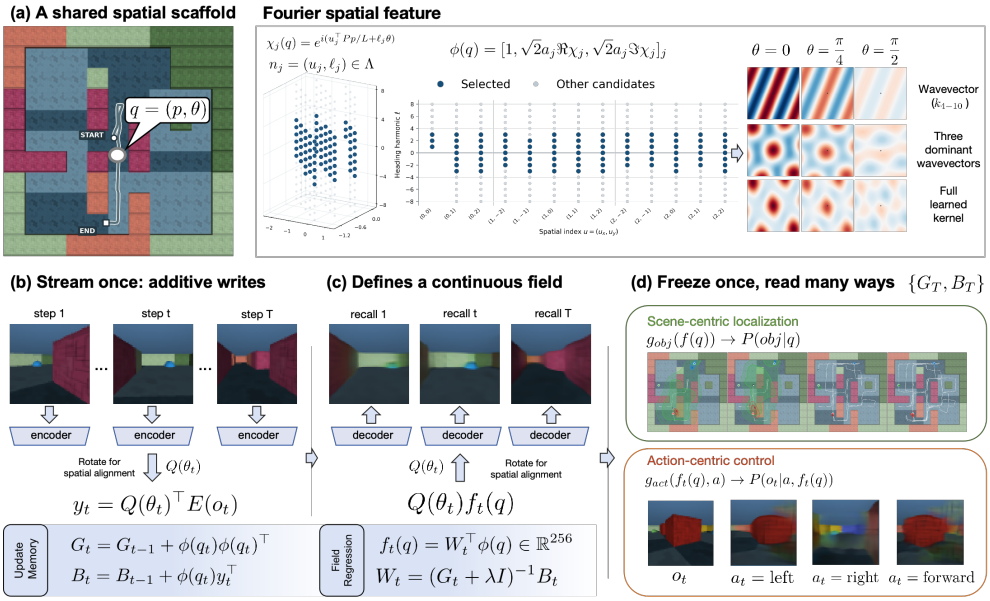

Figure 1: **From shared spatial addresses to reusable episode memory.** (a) A sparse set of joint spatial–heading Fourier modes maps pose $q = (p, \theta)$ to continuous addresses and defines the scaffold's pose-similarity structure; the rightmost slices show representative spatial components of the learned kernel. (b) Encoded observations are accumulated online into episode-specific regression statistics. (c) The accumulated state defines a continuous visual field that can be queried at arbitrary poses. (d) The same memory representation supports scene-centric object localization and action-centric candidate-pose reads for downstream control, without rewriting the stored experience.

3.1 A SHARED SCAFFOLD OVER POSITION AND HEADING

Fourier characters provide continuous addresses for observed and unobserved poses. Spatial plane waves combine with heading harmonics; their inner product yields a pose-similarity kernel, and their products close under frequency differences. This closure lets the writer accumulate regression statistics without storing poses. A learned sparse spectrum determines which pose distinctions fit within the budget.

A frozen codec maps image o_t to normalized latent $z_t = E(o_t) \in \mathbb{R}^c$, with $c = 256$. Supplied pose $q_t = (p_t, \theta_t)$ uses an episode-relative chart whose origin and forward/right axes are fixed by the first observation. Let $\Lambda = \{(u_j, \ell_j)\}_{j=1}^J$ contain one representative of each selected nonzero conjugate frequency pair, with $u_j \in \mathbb{Z}^2$ and $\ell_j \in \mathbb{Z}$. A learned matrix $P \in \mathbb{R}^{2\times 2}$, positive amplitudes a_j, and configured scale L define

$$\chi_j(q) = \underbrace{e^{\mathrm{i}u_j^\top Pp/L}}_{\text{spatial plane wave}} \underbrace{e^{\mathrm{i}\ell_j\theta}}_{\text{heading harmonic}},$$

$$\phi(q) = [1, \{\sqrt{2}a_j \operatorname{Re}\chi_j(q), \sqrt{2}a_j \operatorname{Im}\chi_j(q)\}_{j=1}^J]^\top, \qquad d = 1 + 2J. \tag{1}$$

Each selected mode can jointly encode position and heading; the address is not a full tensor product of all possible modes. The address inner product gives the scaffold's similarity kernel:

$$k(q, q') = 1 + 2\sum_j a_j^2 \cos\big(u_j^\top P(p - p')/L + \ell_j(\theta - \theta')\big). \tag{2}$$

The kernel governs pose interactions: pure spatial modes distinguish locations, pure angular modes distinguish headings, and mixed modes make addresses view-selective. Its spectrum is shared; each episode binds new visual content to it. The selected spectrum and its spatial cross-section are analyzed in Appendix C.

To share visual content across headings, we align the latent before writing: $y_t = Q(\theta_t)^\top z_t$, where

$$Q(\theta) = I + V[\operatorname{blockdiag}_r \operatorname{Rot}(m_r\theta) - I]V^\top, \quad V^\top V = I, \quad m_r \in \mathbb{Z}. \tag{3}$$

Orthonormal column pairs in V make $Q(\theta)Q(\theta') = Q(\theta+\theta')$ and $Q(\theta)^\top = Q(-\theta)$. Thus alignment handles predictable latent rotation, while mixed address modes retain viewpoint-dependent content.

3.2 Accumulating episodic content with sufficient statistics

Given the shared address and alignment, the episode writer fits the observed visual content by ridge regression:

$$W_t = \arg\min_W \sum_{i\le t} \|W^\top \phi(q_i) - y_i\|^2 + \lambda \|W\|_F^2,$$

$$f_t(q) = Q(\theta) W_t^\top \phi(q), \qquad \widehat{o}_t(q) = D(f_t(q)). \tag{4}$$

The field depends on past observations through a Gram matrix $G_t = \sum_i \phi(q_i)\phi(q_i)^\top$ and content cross-statistic B_t. We retain B_t directly and encode G_t with Fourier difference moments. Define the signed support

$$\Omega = \{\mathbf{0}\} \cup \Lambda \cup (-\Lambda)$$

and its frequency-difference set

$$\Gamma = \Omega - \Omega := \{n - n' : n, n' \in \Omega\}.$$

. Character products reduce to frequency differences, so

$$h_t(\delta) = \sum_{i\le t} \chi_\delta(q_i), \quad \delta \in \Gamma, \qquad B_t = \sum_{i\le t} \phi(q_i) y_i^\top. \tag{5}$$

recover G_t and hence $W_t = (G_t + \lambda I)^{-1} B_t$. Each real observation adds one term to each statistic, with no replay of images or poses. Conjugate symmetry makes h_t equivalent to $|\Gamma|$ real scalars. The implementation stores unweighted content statistics and applies shared amplitudes when solving, an equivalent parameterization of Eq. 4.

For FP32 content/origin and FP64 moments, the retained state occupies

$$\mathcal{B}(\Lambda) = 4dc + 8|\Gamma| + 12 \quad \text{bytes.} \tag{6}$$

For the illustrated 9×9 scaffold, the persistent writer state occupies 177,156 bytes ($d = 165$, $|\Gamma| = 1023$), 36.4% less than a packed FP64 Gram representation of the same field. This allocation does not grow with episode length, even though every real observation is written. The moments are sufficient for the ridge field in exact arithmetic (Appendix A.7).

3.3 Learning the scaffold under a persistent-state budget

The fixed allocation makes the choice of modes consequential: adding a conjugate pair increases the visual coefficients and can enlarge the difference-moment bank. Starting with the DC component alone, we optimize the frequency support and continuous parameters under the actual writer-byte constraint:

$$\min_{\Lambda, P, \log a} \ \tfrac{1}{2}\mathcal{L}_{\text{recall}} + \tfrac{1}{2}\mathcal{L}_{\text{heldout}} \quad \text{subject to} \quad \mathcal{B}(\Lambda) \le \mathcal{B}_{\max}. \tag{7}$$

The two losses reward retention of written views and prediction at withheld headings. Both use visual trajectories alone; object labels and rewards do not train the memory. With the codec and calibrated Q fixed, we constrain amplitudes to $10^{-6} \le a_j \le 16$, alternate continuous refinement of P, a with feasible additions, deletions and swaps of frequency pairs, and select the scaffold using TRAIN data only. We then freeze it before fitting any task reader. Appendix A.4 and Algorithms 1–2 specify the search and the per-observation writer.

3.4 Reusing the frozen episode field

View recall evaluates and decodes $f_t(q)$ at a requested pose. For 9×9 object localization, a separate reader samples the same field at 256 spatial Sobol nodes and 32 headings, predicts object-location Gaussians in each query-centered frame, transports them into the common episode chart, and mixes them using a nonnegative visitation measure $\rho_t(q)$ computed from the same Fourier moments h_t, requiring no additional persistent state. The resulting mixture defines a scene-centric belief over object location. This fixed quadrature exposes both visual content and pose coverage without replaying observations. Object labels train only the reader; the codec, scaffold and writer remain frozen (Appendix A.5).

3.5 SCENE-CENTRIC BELIEF IN CANDIDATE ACTION FRAMES

Let G denote the object-location random variable under the scene-centric mixture. For a candidate pose $\widetilde{q} = (\widetilde{p}, \widetilde{\theta})$, the object mixture's displacement becomes $R_{\widetilde{\theta}}^{\top}(G - \widetilde{p})$. We provide the relative mean together with characteristic-function values at the scaffold's pure-spatial frequencies, retaining information about dispersion and multimodality beyond the mixture mean(Appendix A.6). For each candidate action, a motion predictor proposes future poses at which the memory can be queried. Global supplies these action-relative descriptors; Local supplies recalled content and ridge leverage; PPO receives neither readout. All three share an executed-action external-reward prediction head. The control study keeps all learned VMF and object-reader parameters frozen throughout policy training; real observations continue to update the episode state, while candidate poses only query it. (Appendix F).

4 EXPERIMENTAL EVALUATION

Can a fixed-size field preserve visual experience for later queries and downstream use? We evaluate VMF on three settings: long-history and withheld-view recall, scene-centric object localization, and action-relative control. The scaffold is learned from visual trajectories under a fixed byte budget and frozen before task-specific readers or policies are trained. All experiments use Memory Maze with supplied poses and a frozen visual codec; object labels supervise readers, not memory. Full protocols and uncertainty conventions appear in the appendix.

4.1 LONG-HISTORY RETENTION AND WITHHELD-VIEW ACCESS

We first test whether early visual evidence remains accessible after long write histories and at viewpoints that were never written. Recall is evaluated on 64 held-out 9×9 maps over 2048 actions. All methods share observations, poses, codec, evaluation masks, and query poses; matched-Q controls additionally share VMF's heading alignment and visual supervision. We report latent NMSE on both written views and target observations withheld from the writer (Appendix D).

We compare VMF with three learned visual-memory baselines spanning complementary mechanisms: Fourier ridge, a Fourier-feature regression baseline; DeepTime, a sequence-specific regression baseline; and Scene Memory Transformer (SMT), an attention-based memory over observation embeddings. All are local implementations trained on the same 500 training and 100 validation episodes. Matched-supervision variants additionally use VMF's heading alignment and written-view/withheld-heading objective, so recall differences are not attributable to a decoder or target set available only to VMF.

The learned comparison memories are local implementations trained on the same 500 training and 100 validation episodes, rather than scores transcribed from prior papers. They use the same frozen visual codec. Matched-supervision variants additionally receive the VMF heading alignment and the same written-view/withheld-heading objective, so differences in the recall table are not caused by a decoder or target set seen only by VMF.

Table 1 summarizes recall of written and withheld views after 2048 actions. VMF has the lowest mean NMSE on both query sets, including target views that never enter the writer.

Table 1: **View recall after 2048 actions.** Latent NMSE (↓); mean ± sample SD across three fits.

Memory	Written	Withheld 20%
Fourier	0.519 ± 0.012	0.598 ± 0.015
DeepTime	0.557 ± 0.017	0.642 ± 0.020
SMT	0.974 ± 0.010	0.975 ± 0.008
VMF	$\mathbf{0.403 \pm 0.002}$	$\mathbf{0.497 \pm 0.002}$

Figure 2 brings the quantitative and visual tests together. Early views remain queryable as the episode grows, even when later revisits to their locations are removed (b). A separate 40° heading gap, anchored at action 500, writes only its endpoints and queries the unseen interior after further accumulation (c). Thus the two curves test retention of old content and interpolation at an unwritten viewpoint, respectively. Both pool squared residuals and target energy over the same 64 maps; points and whiskers summarize three independent fits per method. The early-view cohort uses the first oc-

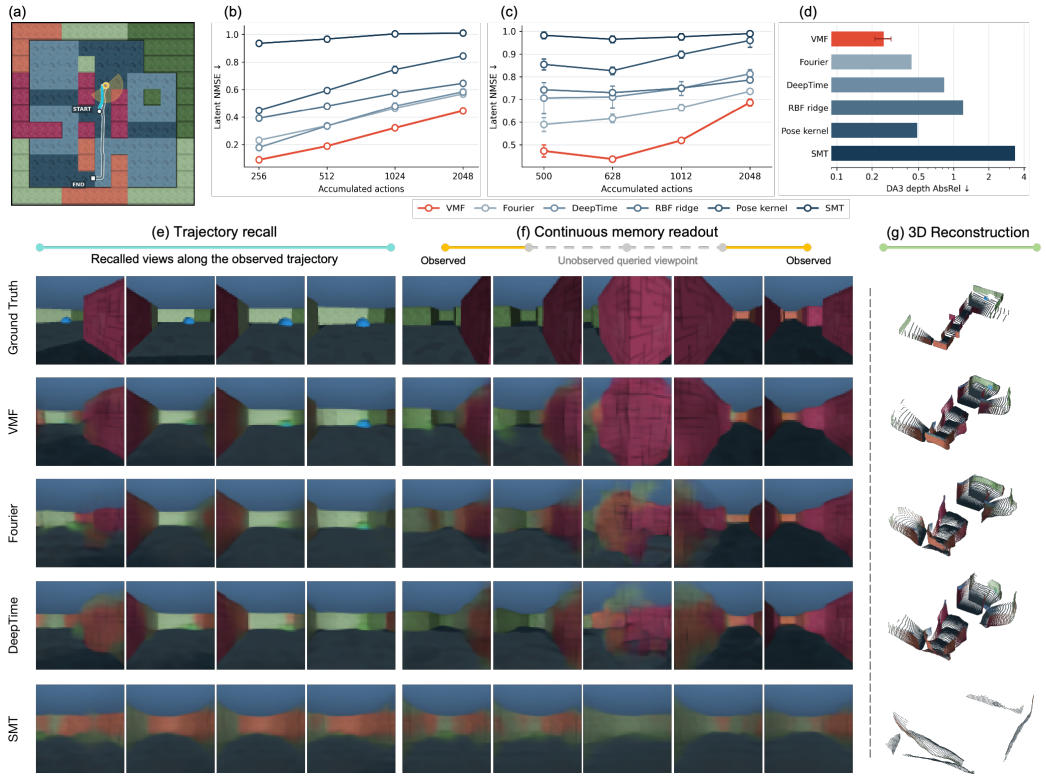

Figure 2: **Visual memory after accumulation.** (a) A selected Memory Maze episode and the path used for qualitative reads. (b) Recall of early views with refreshing writes removed; (c) delayed interpolation across a $40°$ heading gap anchored at action 500. Curves show mean latent NMSE and pointwise sample SD across three independent fits per method on 64 maps. (d) Depth Anything 3 AbsRel on 32 shared views of one map; the VMF whisker spans three fits, while comparison bars are single fits. (e) Recalled views along an observed trajectory; (f) continuous readout between observed endpoints at unwritten viewpoints; (g) ground-truth geometry and reconstructions from VMF, Fourier, DeepTime and SMT under shared cameras. The examples in (a,e–g) use one episode state frozen at action 1000; the eight-view 3D illustration is separate from the 32-view depth score in (d).

currence of each pose, and the heading-gap protocol removes nearby route writes before inserting its endpoints (Appendix D).

The lower panels show what these errors mean visually. Along one recorded path, VMF recalls observed views (e) and synthesizes interior viewpoints that were never written (f); the same query poses are used for every method. The 3D comparison (g) illustrates how recalled appearance carries across viewpoints, beyond individual frames. A frozen depth estimator receives recalled RGB under shared cameras and intrinsics; simulator depth supplies the reference and common display mask (Appendix D.4).

4.2 Reinterpreting one frozen state

We next ask whether an accumulated visual memory can support a later inference task without being re-encoded. Unlike view recall, object localization requires integrating evidence across multiple viewpoints into a scene-centric belief over object position.

At action 1000, we freeze a VMF episode state and apply two readouts without intervening writes: the visual decoder and a separately trained object-location reader. Reversing the order of the two readouts leaves both the memory state and the outputs unchanged, providing a simple state-invariance check.

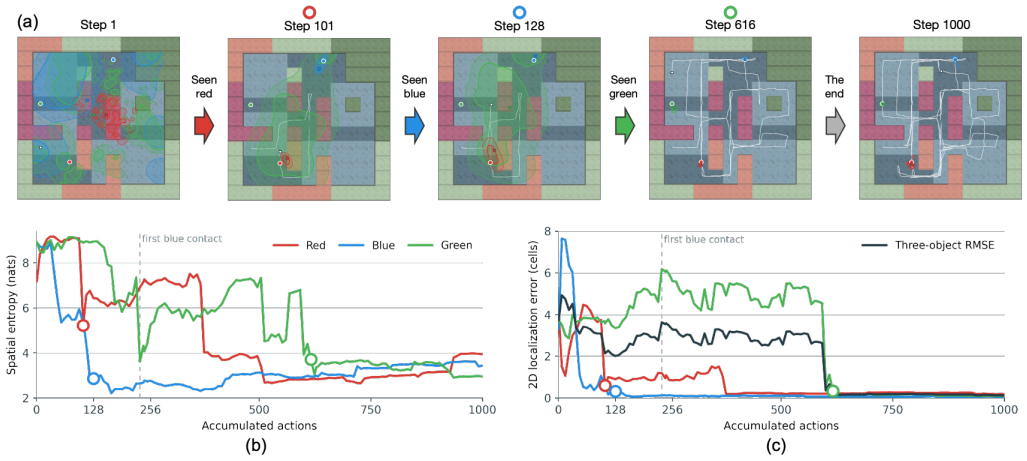

Figure 3: **Object belief during one held-out episode.** Top: three-object Gaussian-mixture readouts at five labeled actions from 1 to 1000. Filled markers denote target locations; crosses denote predicted mixture means; contours enclose 50% and 80% of each density within the displayed arena. Bottom: spatial grid entropy (left) and object-position error (right) as observations accumulate. The dark curve is RMSE across the three objects; open circles mark intermediate readouts, and the dashed line marks first physical contact with the blue object.

Figure 3 illustrates how the scene-centric readout develops as evidence accumulates. The same trained reader is evaluated on successive prefixes of one held-out episode. Its object-location distributions sharpen at different times, and the blue-object belief becomes concentrated before first physical contact with that object. The accompanying position-error curves distinguish increasing confidence from increasing accuracy: low entropy alone does not imply a correct location estimate. This single-episode visualization is illustrative; the cross-map evaluation is reported in Table 2.

To test whether localization depends specifically on retained visual history, Table 2 compares full-history VMF with controls that vary the amount and form of information available to the reader: the current observation alone, the last 64 observations, geometry/visitation without visual content, task-supervised sequence models, and alternative full-history visual memories. Each restricted-information condition is fitted to its own inputs rather than applying a VMF-trained reader to degraded information.

We evaluate both the action-1000 endpoint and the aggregate over actions 501–1000 on two held-out cohorts. Paired-64 is geometry-disjoint from training, while Clean-865 excludes official test trajectories whose wall geometry overlaps training modulo rotation or reflection. The two cohorts are reported separately and are not pooled (Appendix E).

Full-history VMF has lower mean RMSE than the restricted-information controls, indicating that the object reader exploits visual evidence retained beyond recent observations or visitation structure alone. VMF has the lowest mean RMSE in both action-1000 columns and on the Clean-865 action-window aggregate, while matched-Q Fourier has the lowest mean on the Paired-64 aggregate. Object labels never update the visual memory (Appendix E).

As a within-size extension, we separately fit three 15×15 VMF scaffolds and paired six-object readers. Over actions 501–1000 on the official 1000-maze evaluation, the three fits obtain 5.224 ± 0.246 cells RMSE (27.295 ± 2.546 cells MSE). The codec and heading frame remain frozen, while the spatial scaffold is refitted for the larger arena. For scale, published offline probes trained with 29,000 trajectories report $9 \times 9/15 \times 15$ Objects MSE of $1.8/25.3$ for S5WM and $8.9/34.8$ for S4WM (Pasukonis et al., 2023; Deng et al., 2023).

4.3 Action-Relative Scene Access for Control

Recall and localization establish what can be read from an accumulated field. We next ask whether those readouts can support action selection. In the 9×9 Memory Maze control task, a color cue identifies the target object and the agent navigates for up to 1000 actions. The visual codec, VMF

Table 2: **Object localization.** RMSE in cells (↓) at action 1000 and over actions 501–1000. Full-history VMF reports mean ± sample SD across three independently fitted memory–reader pairs; restricted-history and comparison conditions are fitted separately.

	Action 1000		Actions 501–1000	
Memory / input	**Clean-865**	**Paired-64**	**Clean-865**	**Paired-64**
Untrained references				
Arena center	3.695	3.571	3.695	3.571
Uniform guess	5.211	5.123	5.211	5.123
Restricted information				
Current observation	4.643 ± 0.014	4.411 ± 0.024	4.682 ± 0.012	4.315 ± 0.004
Geometry / visitation	$3.703^{\dagger}$	$3.548^{\dagger}$	$3.709^{\dagger}$	$3.542^{\dagger}$
Task-supervised sequence baselines				
GRU	4.044 ± 0.028	3.850 ± 0.063	4.031 ± 0.020	3.871 ± 0.020
Bounded attention	3.921 ± 0.014	3.896 ± 0.022	3.904 ± 0.010	3.840 ± 0.008
Full-history visual-memory baselines				
Pose kernel	1.761 ± 0.051	1.209 ± 0.041	1.808 ± 0.050	1.369 ± 0.035
Matched-Q Fourier	1.454 ± 0.032	0.777 ± 0.112	1.566 ± 0.009	$\mathbf{0.928 \pm 0.027}$
Ours				
VMF, last 64	3.527 ± 0.013	3.245 ± 0.033	3.507 ± 0.004	3.218 ± 0.005
VMF (full history)	$\mathbf{0.933 \pm 0.098}$	$\mathbf{0.732 \pm 0.184}$	$\mathbf{1.209 \pm 0.071}$	1.034 ± 0.115

Among fitted conditions, only GRU and bounded attention train their sequence encoders with object labels; all other object supervision is confined to the reader. Bold marks the lowest mean in each column; † indicates SD < 0.001.

parameters, and object reader remain fixed during policy training. Real observations continue to update the episode state, while proposed future poses only query the existing field and never write imagined observations.

We compare three policy access pathways. For each of six candidate actions, a motion predictor proposes three successive future poses from the current real-observation prefix. **PPO** receives no VMF readout. **Local** queries the field at the proposed poses and supplies recalled visual content together with ridge leverage, indicating how well each query is supported by accumulated observations. **Global** instead re-expresses the object reader's mixture belief for the cued object in each candidate pose's body-centered frame. The resulting descriptors combine candidate-relative displacement with characteristic-function features at the scaffold's pure-spatial frequencies, retaining uncertainty beyond a point estimate. These candidate-conditioned features enter the recurrent policy's action and value predictions. Thus Local asks "what might I see there?", whereas Global asks "where is the target relative to where this action may take me?"

This is a comparison of complete access pathways rather than isolated readout features. Global also changes the information used during recursive motion proposals, so the Global–Local difference includes both the scene-level readout and its effect on proposal geometry. Candidate queries remain read-only in all conditions; only actual observations update VMF. The candidate-frame transform is given in Appendix A.6. All three pathways use the same recurrent PPO backbone and the same executed-action reward-prediction auxiliary objective. A 128-unit SiLU head predicts the external reward of the executed action from the recurrent state and action encoding,

$$\mathcal{L}_{\text{aux}} = 0.1 \ \text{SmoothL1}\big(\operatorname{asinh} \hat{r}_t, \operatorname{asinh} r_t^{\text{ext}}\big) .$$

This auxiliary loss trains the recurrent representation; its prediction is not used as an imagined reward, return, or action-selection target.

To separate memory access from exploration shaping, we additionally cross each policy pathway with an optional VMF-based visual-surprise bonus. The bonus measures the change induced by a real observation relative to the pre-write field and is annealed to zero during training. PPO receives no memory readout even in this condition; it consults VMF only to compute the bonus. Reported

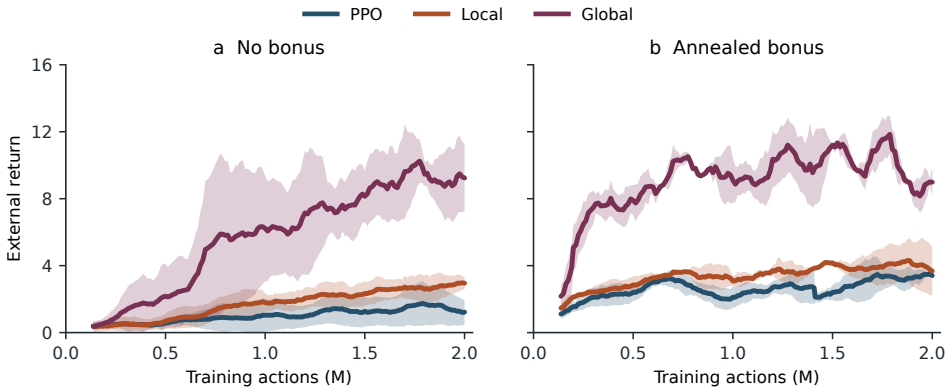

Figure 4: **Control with a frozen VMF.** Colors denote PPO, Local and Global; panels switch the KL-surprise bonus off/on. Curves show mean ± sample SD across three seeds of trailing-128-episode external return, excluding bonus. Training and evaluation protocols are in Appendix F.

returns and the reward-prediction target always use external reward. The exact KL expression, annealing schedule, and optimization details are given in Appendix F.

Figure 4 and Table 3 compare the three access pathways with and without the visual-surprise bonus. Without the bonus, Global obtains substantially higher held-out return than Local and PPO, while Local has a higher mean than PPO. The Global advantage is especially pronounced over actions 501–1000, indicating that the scene-centric access pathway remains useful after substantial memory accumulation. The surprise bonus changes the weaker pathways more than Global: PPO's mean return increases markedly, whereas Local and Global change less and show greater seed variation. We therefore treat the bonus as an auxiliary exploration condition rather than the source of Global's advantage. The Global–Local comparison still concerns the complete access pathway, including Global's effect on motion proposals.

Table 3: **Held-out control return.** Mean ± SD across three test seeds.

	Full episode	Actions 501–1000
PPO	1.620 ± 0.832	0.726 ± 0.439
Local	2.984 ± 0.524	1.457 ± 0.327
Global	$\mathbf{9.646 \pm 1.095}$	$\mathbf{6.113 \pm 0.657}$
PPO + bonus	3.677 ± 0.804	1.858 ± 0.348
Local + bonus	3.524 ± 1.295	1.653 ± 0.618
Global + bonus	$\mathbf{10.215 \pm 2.711}$	$\mathbf{6.655 \pm 1.664}$

5 Discussion

VMF shows that a bounded episode state can preserve visual experience for questions specified only after accumulation. Long-history and withheld-view probes test access to early and unwritten views, while a separately supervised object reader extracts scene-centric beliefs from the same accumulated memory. Restricted-history and geometry-only controls indicate that this localization relies on retained visual evidence beyond recent observations or visitation structure. The scaffold kernel determines how evidence is shared across poses, while sparse Fourier support allocates this sharing pattern under a fixed persistent-state budget. Frequency-difference moments compactly represent the regression geometry, and the learned spectrum is best interpreted through the pose interactions induced by its kernel. Fixed capacity still imposes a limit: observations mapped to nearly identical addresses must interfere (Appendix C.1). For control, Global re-expresses object-location belief in candidate-relative coordinates, whereas Local queries recalled visual content at candidate poses. Global attains higher external return in the reported setting, but this comparison evaluates the complete access pathway, including its effect on motion proposals, rather than the coordinate transform alone.

AI USE STATEMENT

OpenAI Codex was used for manuscript drafting and editing, literature retrieval, implementation and debugging of experiment orchestration and analysis code, numerical checks, and assembly of this manuscript. This includes substantive assistance with the presentation of derivations and interpretation of results. The authors take responsibility for the technical claims, references, code and this disclosure. No generative image model was used for the reported experiment figures.

REPRODUCIBILITY STATEMENT

The appendices specify the writer, support search, supervision, evaluation masks, metrics and uncertainty calculations. The accompanying project provides the figure assets, table sources and provenance information needed to trace reported results to their data and model checkpoints.

REFERENCES

Caswell Barry, Robin Hayman, Neil Burgess, and Kathryn J. Jeffery. Experience-dependent rescaling of entorhinal grids. *Nature Neuroscience*, 10:682–684, 2007. doi: 10.1038/nn1905.

Andrej Bicanski and Neil Burgess. A neural-level model of spatial memory and imagery. *eLife*, 7: e33752, 2018. doi: 10.7554/eLife.33752.

Patrick Byrne, Suzanna Becker, and Neil Burgess. Remembering the past and imagining the future: A neural model of spatial memory and imagery. *Psychological Review*, 114(2):340–375, 2007. doi: 10.1037/0033-295X.114.2.340.

Sarthak Chandra, Sugandha Sharma, Rishidev Chaudhuri, and Ila Fiete. Episodic and associative memory from spatial scaffolds in the hippocampus. *Nature*, 638:739–751, 2025. doi: 10.1038/s41586-024-08392-y.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1734, 2014. URL https://aclanthology.org/D14-1179/.

Fei Deng, Junyeong Park, and Sungjin Ahn. Facing off world model backbones: RNNs, transformers, and S4. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/e6c65eb9b56719c1aa45ff73874de317-Paper-Conference.pdf.

Kuan Fang, Alexander Toshev, Li Fei-Fei, and Silvio Savarese. Scene memory transformer for embodied agents in long-horizon tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 538–547, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Fang_Scene_Memory_Transformer_for_Embodied_Agents_in_Long-Horizon_Tasks_CVPR_2019_paper.html.

Marco Fraccaro, Danilo Jimenez Rezende, Yori Zwols, Alexander Pritzel, S. M. Ali Eslami, and Fabio Viola. Generative temporal models with spatial memory for partially observed environments. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 2018. URL https://proceedings.mlr.press/v80/fraccaro18a.html.

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

Zizhan He, Maxime Daigle, and Pouya Bashivan. Building spatial world models from sparse transitional episodic memories. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2505.13696.

Naoki Kiyohara, Harrison Bo Hua Zhu, Riccardo El Hassanin, Zhuo Sun, Wenlong Chen, Samir Bhatt, and Yingzhen Li. Interdomain attention: Beyond token-level key-value memory, 2026. URL https://arxiv.org/abs/2605.24330.

Julija Krupic, Marius Bauza, Stephen Burton, Caswell Barry, and John O'Keefe. Grid cell symmetry is shaped by environmental geometry. *Nature*, 518:232–235, 2015. doi: 10.1038/nature14153.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e4cd50120b6d7e8daff1749d6bbaa889-Abstract-Conference.html.

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*, 2020.

Emilio Parisotto and Ruslan Salakhutdinov. Neural map: Structured memory for deep reinforcement learning. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Bk9zbyZCZ.

Jurgis Pasukonis, Timothy Lillicrap, and Danijar Hafner. Evaluating long-term memory in 3D mazes. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=yHLvIlE9RGN.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007. URL https://proceedings.neurips.cc/paper/2007/hash/013a006f03dbc5392effeb8f18fda755-Abstract.html.

Mohammad Reza Samsami, Artem Zholus, Janarthanan Rajendran, and Sarath Chandar. Mastering memory tasks with world models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=1vDArHJ68h.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Aman Sinha and John C. Duchi. Learning kernels with random features. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/hash/e70611883d2760c8bbafb4acb29e3446-Abstract.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Johannes von Oswald, Nino Scherrer, Seijin Kobayashi, Luca Versari, Songlin Yang, Sarthak Mittal, Maximilian Schlegel, Kaitlin Maile, Yanick Schimpf, Oliver Sieberling, Alexander Meulemans, Rif A. Saurous, Guillaume Lajoie, Charlotte Frenkel, Razvan Pascanu, Blaise Agüera y Arcas, and João Sacramento. MesaNet: Sequence modeling by locally optimal test-time training. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2506.05233.

Greg Wayne, Chia-Chun Hung, David Amos, Mehdi Mirza, Arun Ahuja, Agnieszka Grabska-Barwinska, Jack Rae, Piotr Mirowski, Joel Z. Leibo, Adam Santoro, Mevlana Gemici, Malcolm Reynolds, Tim Harley, Josh Abramson, Shakir Mohamed, Danilo Rezende, David Saxton, Adam Cain, Chloe Hillier, David Silver, Koray Kavukcuoglu, Matt Botvinick, Demis Hassabis, and Timothy Lillicrap. Unsupervised predictive memory in a goal-directed agent, 2018. URL https://arxiv.org/abs/1803.10760.

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.

Gerald Woo, Chenghao Liu, Doyen Sahoo, Akshat Kumar, and Steven Hoi. Learning deep time-index models for time series forecasting. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 37217–37237, 2023. URL https://proceedings.mlr.press/v202/woo23b.html.

A COMPLETE MEMORY AND READOUT SPECIFICATION

This appendix specifies the shared address, episode writer, and two readouts; Algorithms 1–2 summarize construction and use.

A.1 A REUSABLE SPATIAL SCAFFOLD

Observations and coordinates. At time t, an agent receives an image o_t and a measured planar pose. We express poses as $q_t = (p_t, \theta_t) \in \mathbb{R}^2 \times S^1$ in the coordinate frame of the first observation: p_t is displacement along the initial forward/right axes, in maze cells, and θ_t is heading in this chart. The chart origin is retained throughout the episode. Pose estimation is assumed available. A pretrained visual codec supplies a normalized latent $z_t = E(o_t) \in \mathbb{R}^c$, with $c = 256$; its decoder D maps reconstructed latents back to images. We absorb the fixed latent normalization into E and its inverse into D. Both are frozen during memory learning and subsequent task learning.

Fourier addresses. Let $\Lambda = \{n_j = (u_j, \ell_j)\}_{j=1}^{J}$ contain one representative of each selected nonzero conjugate frequency pair, where $u_j \in \mathbb{Z}^2$ and $\ell_j \in \mathbb{Z}$. A learned matrix $P \in \mathbb{R}^{2\times 2}$ and a configured scale $L > 0$ define phases

$$\varphi_j(q) = u_j^\top P \frac{p}{L} + \ell_j \theta. \tag{8}$$

The real address vector is

$$\phi(q) = \left[1, \left\{\sqrt{2}\, a_j \cos \varphi_j(q),\ \sqrt{2}\, a_j \sin \varphi_j(q)\right\}_{j=1}^{J}\right]^\top \in \mathbb{R}^d, \qquad d = 1 + 2J, \tag{9}$$

with positive learned amplitudes $a_j = \exp(s_j)$. The constant component has fixed amplitude one. All visual channels use the same addresses. Integer angular orders make the representation exactly periodic in heading. At a fixed heading, each pair with $u_j \neq 0$ describes a spatial plane wave with wavevector $P^\top u_j / L$; heading changes its phase. The selected 9×9 instantiation has $d = 165$: one constant and 82 sine/cosine pairs, rather than 165 independent spatial waves.

We call the shared map $q \mapsto \phi(q)$ the *spatial scaffold*. It supplies both coordinates for storing content and a learned similarity between pose addresses:

$$k(q, q') = \phi(q)^\top \phi(q') = 1 + 2 \sum_{j=1}^{J} a_j^2 \cos\left(u_j^\top P \frac{p - p'}{L} + \ell_j(\theta - \theta')\right). \tag{10}$$

This kernel is positive semidefinite and depends only on pose differences within the chart. Its frequencies and amplitudes determine spatial scales, directions, and angular selectivity. Once learned, k is fixed across episodes; accumulated observations change the field content, not this address similarity. Sharing the scaffold means reusing its parameters, without assuming that reconstruction quality is invariant to changes in environment size. In particular, $P^\top u_j / L$ makes the physical effect of changing L explicit.

A.2 BINDING SENSORY CONTENT AND RETRIEVING VIEWS

A heading-dependent visual frame. Images observed at different headings are expressed in different visual coordinates. Before writing, we align their latents using a learned periodic orthogonal map

$$\begin{aligned} Q(\theta) &= I + V\big[\mathcal{R}(\theta) - I\big]V^\top, & V^\top V &= I, \\ \mathcal{R}(\theta) &= \operatorname{blockdiag}_{r=1}^{R} \operatorname{Rot}(m_r \theta), & m_r &\in \mathbb{Z}. \end{aligned} \tag{11}$$

Here $V \in \mathbb{R}^{c\times 2R}$ specifies orthogonal latent planes, and Rot is the two-dimensional rotation matrix. The complement of these planes is unchanged. Consequently, $Q(\theta)^\top = Q(-\theta)$ and $Q(\theta)Q(\theta') = Q(\theta + \theta')$. We write the aligned value $y_t = Q(\theta_t)^\top z_t$. Position determines where content is addressed; heading enters both the address and the visual frame. This learned alignment approximates relations between view latents; it does not assume that occlusion or newly visible content is an exact orthogonal transformation of an image.

Episode-specific field. Given real observations through time t, content is represented by the ridge-regression solution

$$W_t = \arg\min_{W\in\mathbb{R}^{d\times c}} \sum_{i=1}^{t} \|W^\top \phi(q_i) - y_i\|_2^2 + \lambda\|W\|_F^2, \qquad \lambda > 0, \tag{12}$$

$$f_t(q) = Q(\theta) W_t^\top \phi(q), \qquad \widehat{o}_t(q) = D\big(f_t(q)\big). \tag{13}$$

Orthogonality of Q makes the fitting objective equivalent to latent reconstruction in the original visual coordinates. Writing is online regression with fixed shared parameters, rather than gradient updates to the codec or scaffold at deployment.

The retrieval rule makes the roles of addressing and alignment explicit. Let X_t have rows $\phi(q_i)^\top$ and set $A_t = \lambda I + X_t^\top X_t$. Then

$$f_t(q) = \sum_{i=1}^{t} \alpha_{t,i}(q) Q(\theta - \theta_i) z_i, \qquad \alpha_t(q) = X_t A_t^{-1} \phi(q). \tag{14}$$

The coefficients depend on the joint geometry of all observed addresses, while relative-heading transforms align their sensory contributions. The coefficients can be signed. This is a regression read, with no requirement to retrieve a stored image or select a nearest observation. Equation equation 14 is an interpretation of the read; the implementation does not retain X_t or the latent history.

A.3 Fixed-storage statistics and exact-arithmetic equivalence

Ordinary sufficient statistics for Eq. equation 12 are the $d \times d$ Gram matrix and a $d \times c$ content matrix. Fourier closure allows the Gram matrix to be represented by shared moments. Define the signed support and its difference set as

$$\Omega = \{0\} \cup \Lambda \cup (-\Lambda), \qquad \Gamma = \Omega - \Omega, \tag{15}$$

and let $\chi_n(q) = \exp(\mathrm{i}\varphi_n(q))$. Since $\chi_n\overline{\chi_m} = \chi_{n-m}$, every product of two real address components is a linear combination of characters in Γ. It suffices to update

$$h_t(\delta) = h_{t-1}(\delta) + \chi_\delta(q_t), \qquad \delta \in \Gamma, \tag{16}$$

$$B_t = B_{t-1} + \phi(q_t) y_t^\top. \tag{17}$$

Both statistics are initialized to zero. The zero-frequency moment records the observation count; conjugate moments share a real/imaginary representation. Fixed trigonometric identities recover $G_t = \sum_{i\le t} \phi(q_i)\phi(q_i)^\top$ from h_t, including the learned amplitudes. Retrieval compiles

$$W_t = (G_t + \lambda I)^{-1} B_t \tag{18}$$

using a Cholesky solve. The writer stores content statistics before shared spectral scaling, which is algebraically equivalent to Eq. equation 17.

Storage and exactness. For the single shared spectrum, the persistent writer stores dc content entries, $|\Gamma|$ real moment entries, and three origin coordinates. With FP32 content and origin and FP64 moments, its byte cost is

$$\mathcal{B}(\Lambda) = 4dc + 8|\Gamma| + 12. \tag{19}$$

This count is independent of episode length. The selected 9×9 scaffold has $d = 165$, $c = 256$, and $|\Gamma| = 1023$, giving 177,156 bytes per episode. Shared model parameters, temporary factorization workspaces, compiled read coefficients, and downstream policy state are accounted for separately.

In exact arithmetic, the moments recover the complete Gram matrix and B_t its cross-statistic, giving the same ridge solution as fitting all observations in the selected feature space. Additivity makes the final state independent of write grouping and order for a fixed chart. Finite precision and feature span still limit recall and numerical validity.

The episode state is reset at episode boundaries. Every real observation is written, with no online support-set selection or forgetting factor. Queries leave the writer unchanged. Where needed, the same Gram statistics provide the ridge leverage

$$u_t(q) = \phi(q)^\top (G_t + \lambda I)^{-1} \phi(q). \tag{20}$$

Under an isotropic Gaussian regression interpretation this is proportional to latent mean uncertainty. It measures address coverage under that model, rather than calibrated visual correctness or physical reachability.

A.4 LEARNING THE SCAFFOLD UNDER THE PERSISTENT-STATE BUDGET

Learning has two stages after codec pretraining: calibrating the visual frame, then searching the Fourier scaffold from DC-only support with that frame fixed. The codec and frame are frozen during support search; neither stage uses object-location labels or downstream rewards.

Visual-frame calibration. From training trajectories, we collect pairs of observations at the same quantized position with a small nonzero heading difference $\Delta\theta_{ij} = \theta_j - \theta_i$. We fit

$$\mathcal{L}_Q = \frac{\sum_{(i,j)\in\mathcal{P}} \|Q(\Delta\theta_{ij})z_i - z_j\|_2^2}{\frac{1}{2}\sum_{(i,j)\in\mathcal{P}}(\|z_i\|_2^2 + \|z_j\|_2^2)}. \tag{21}$$

A skew-symmetric least-squares initializer supplies an orthogonal basis and an integer-frequency envelope. Finite-angle calibration then alternates exact integer-order selection within that envelope with orthogonal-basis optimization. The latter uses the parameterization $O = O_0 \exp(H - H^\top)$ and L-BFGS. Zero-order planes are folded into the identity when compiling Eq. equation 11. The resulting Q is frozen for the subsequent scaffold search and throughout deployment.

Recall and withheld-view objectives. An offline training case consists of an episode, a write set $\mathcal{S}$, and a query set $\mathcal{U}$. These are masks over observations, distinct from the Fourier support Λ. Let $f_\mathcal{S}$ denote the field obtained by writing only observations in $\mathcal{S}$. For a collection of cases $\mathcal{C}$, define

$$\mathcal{E}(\mathcal{C}) = \frac{\sum_{(\mathcal{S},\mathcal{U})\in\mathcal{C}} \sum_{i\in\mathcal{U}} \|f_\mathcal{S}(q_i) - z_i\|_2^2}{\sum_{(\mathcal{S},\mathcal{U})\in\mathcal{C}} \sum_{i\in\mathcal{U}} \|z_i\|_2^2}. \tag{22}$$

Full-recall cases write the complete training trajectory and query the first occurrence of each quantized pose. Withheld-view cases select a target using poses alone and remove nearby observations inside an angular gap; retained views bracket the target heading. The target image never enters that case's write statistics. We use requested gaps of 10°, 20°, and 30° and optimize

$$\begin{aligned} \min_{\Lambda,P,s} \quad & \mathcal{L}_{\text{mem}} = \tfrac{1}{2}\mathcal{E}(\mathcal{C}_{\text{recall}}) + \tfrac{1}{2}\mathcal{E}(\mathcal{C}_{\text{heldout}}), \\ \text{subject to} \quad & \mathcal{B}(\Lambda) \le \mathcal{B}_{\max}, \qquad \log(10^{-6}) \le s_j \le \log(16). \end{aligned} \tag{23}$$

The first term trains retention of written content; the second constrains interpolation across missing views. Both are pooled normalized errors over the 500 TRAIN trajectories and 1998 registered cases for the 9×9 search. Randomly withheld views provide a separate evaluation of reuse beyond the specific angular-gap training rule.

Discrete support and continuous geometry. We initialize with the DC component alone ($\Lambda = \emptyset$), then search conjugate pairs in the declared finite integer domain $[-2, 2]^2 \times [-8, 8]$. The 9×9 persistent-state ceiling is $\mathcal{B}_{\max} = 177{,}156$ bytes. For fixed Λ, we differentiate through the ridge solves and refine P and the active log-amplitudes using bound-constrained L-BFGS-B. For fixed continuous parameters, we score the complete feasible neighborhood of one-pair additions, deletions, and swaps under Eq. equation 19, accepting improving moves. An added pair increases both the content dimension and, potentially, the difference-moment bank. Thus the budget constrains the actual sufficient statistics, rather than only the number of retained frequencies. Deletions and swaps allow earlier choices to be revised.

We repeat the blocks until the continuous projected-gradient infinity norm is at most 10^{-5} and no feasible neighborhood move improves the objective by more than 10^{-8} at the same iterate. This is a joint local stopping criterion over the declared search domain, not a global-optimality guarantee. The raw-carrier ridge is kept fixed as support size changes ($\lambda = 10^{-4} \times 153$ in the reported configuration), so changing d does not silently change regularization. The implementation's normalized features $\phi/\sqrt{d}$ are equivalent when paired with ridge λ/d. The TRAIN objective selects the final scaffold; validation and test metrics do not enter checkpoint selection. The selected scaffold is frozen before downstream reader training; object labels never update it.

Amplitude conditioning. We use $a_j = \exp(s_j)$ with the bounds in Eq. equation 23. Positivity alone does not bound conditioning: reparameterizing a weighted carrier coefficient as $b_j = a_j w_j$

Algorithm 1 Construct the shared address scaffold (offline)

Require: TRAIN trajectories with poses; frozen codec E; byte ceiling $\mathcal{B}_{\max}$
1: Calibrate heading alignment Q on same-position view pairs; freeze E, Q.
2: Set $\Lambda \leftarrow \emptyset$ (DC only); initialize P and candidate log-amplitudes.
3: **repeat**
4: Fit episode ridge fields; score full recall and withheld-heading queries.
5: Refine P, s by L-BFGS-B with Λ fixed and $s_j \in [\log(10^{-6}), \log(16)]$.
6: Score feasible support additions, deletions and swaps under $\mathcal{B}(\Lambda) \leq \mathcal{B}_{\max}$.
7: Accept an improving move, if one exists.
8: **until** projected gradient and support moves meet the declared stopping rules
9: Freeze the final iterate.
10: **return** frozen Q, Λ, P, a shared across episodes.

Algorithm 2 Write per observation; read the frozen episode state

Require: Frozen E, Q, ϕ, decoder D, optional trained object reader; ridge λ
1: Initialize difference moments $h_0 = 0$, content statistic $B_0 = 0$ and episode chart.
2: **for** each real observation (o_t, q_t) **do**
3: $y_t \leftarrow Q(\theta_t)^\top E(o_t)$; $h_t(\delta) \leftarrow h_{t-1}(\delta) + \chi_\delta(q_t)$ for $\delta \in \Gamma$.
4: $B_t \leftarrow B_{t-1} + \phi(q_t) y_t^\top$ (equivalent unweighted-carrier implementation).
5: **end for**
6: Recover G_t from h_t; solve $W_t \leftarrow (G_t + \lambda I)^{-1} B_t$.
7: **View:** return $D(Q(\theta) W_t^\top \phi(q))$.
8: **Scene:** aggregate field queries and ρ_t with the separately trained reader.
9: **Action (optional):** express the belief in a candidate frame via Eq. 26.

changes its ridge penalty to $\lambda b_j^2 / a_j^2$, so unbounded amplitude growth can effectively remove regularization. The upper bound prevents this mechanism, while the small positive lower bound allows negligible carriers. Codec and frame calibration precede Fourier support search. The learned scaffold is frozen before visual and object readouts are fitted.

A.5 Reading scene content from a frozen field

After memory learning, a task reader can be trained on the field while keeping E, Q, Λ, P, and s frozen. This separates the sensory memory objective from the supervision used to interpret its contents. For object localization, we evaluate the visual field at spatial Sobol nodes and evenly spaced headings. The nodes cover the disclosed arena domain and are transformed into the episode chart. Fourier synthesis, including $Q(\theta)$, precedes the nonlinear reader; the angular evaluation is exact at these nodes for the finite visual spectrum. Aggregation after the nonlinear head is a finite quadrature approximation.

The stored moments also define a nonnegative visitation measure. For $t > 0$, let

$$\rho_t(q) = \frac{1}{t} \sum_{i=1}^{t} \left| \frac{1}{|\Omega|} \sum_{n \in \Omega} \chi_n(q) \overline{\chi_n(q_i)} \right|^2 . \tag{24}$$

Expanding the square involves only $\Omega - \Omega$, so ρ_t is computed from the existing moment bank without stored addresses or extra persistent state. It measures visitation under the selected Fourier support; it is neither an object-visibility label nor the learned kernel's posterior variance.

At query q_j, a shared head applied to $f_t(q_j)$ predicts, for each object identity o, a local mean b_{jo}, positive-definite covariance C_{jo}, and mixture logit v_{jo}. Let R_{θ_j} map local forward/right offsets into the episode chart. Define

$$\mu_{jo} = p_j + R_{\theta_j} b_{jo}, \qquad \Sigma_{jo} = R_{\theta_j} C_{jo} R_{\theta_j}^\top,$$
$$\pi_{jo} = \frac{\rho_t(q_j) \exp(v_{jo})}{\sum_k \rho_t(q_k) \exp(v_{ko})}, \qquad p_\psi(g_o \mid M_t) = \sum_j \pi_{jo} \mathcal{N}(g_o; \mu_{jo}, \Sigma_{jo}). \tag{25}$$

The equal quadrature weights cancel in the mixture normalization. Here M_t denotes the persistent memory state; the mixture is a derived readout on the chart plane. The reader is trained by negative log-likelihood of labeled object coordinates at causal trajectory prefixes. Its analytic mixture mean supplies a point prediction for localization evaluation. Object labels update only ψ, and source validation selects the reader. No object-conditioned writes are added to the memory.

A.6 Re-expressing scene beliefs for downstream control

The policy study freezes its VMF and object reader throughout training, as documented in Appendix F. The transform below applies to a frozen VMF belief.

A scene belief can be expressed relative to any candidate agent pose without rewriting or re-encoding the episode. For the currently cued object, let g_o follow the mixture in Eq. equation 25. At a proposed pose $\widetilde{q} = (\widetilde{p}, \widetilde{\theta})$, its body-relative displacement is $Y = R_{\widetilde{\theta}}^{\top}(g_o - \widetilde{p})$. Alongside the unwrapped mean of Y, we use its characteristic function at the scaffold's existing pure-spatial wavevectors $\omega = P^{\top}u/L$, with $(u, 0) \in \Lambda$:

$$\begin{aligned} v_\omega &= R_{\widetilde{\theta}}\omega, \\ \Phi_t^{\mathrm{obj}}(\omega; \widetilde{q}) &= \mathbb{E}\big[\exp(\mathrm{i}\omega^{\top}Y)\big] \\ &= \sum_j \pi_j \exp\big(\mathrm{i}v_\omega^{\top}(\mu_j - \widetilde{p}) - \tfrac{1}{2}v_\omega^{\top}\Sigma_j v_\omega\big)\,. \end{aligned} \tag{26}$$

This is an analytic transform of the complete mixture: its real and imaginary parts retain frequency-dependent information about dispersion and multimodality. The finite frequency set is not a complete encoding of every possible distribution, and the unwrapped mean supplements its periodic features.

In the control instantiation, a motion predictor proposes three successive poses for each of six candidate actions from the same real-observation prefix. Global access supplies transformed object-belief descriptors; Local access supplies recalled visual content and ridge leverage. The candidate-conditioned features contribute to action and value corrections in recurrent PPO. Global holds the current visual latent during recursive proposals, so the ablation also changes proposal geometry. Reward-supervised conditions predict the executed action's external reward from PPO recurrent state ξ_t and action encoding $e(a_t)$. A 128-unit SiLU hidden layer produces the scalar prediction $\hat{r}_t = g_\eta(\xi_t, e(a_t))$. The training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{PPO}} + 0.1\ \mathrm{SmoothL1}(\operatorname{asinh}\hat{r}_t, \operatorname{asinh} r_t^{\mathrm{ext}}), \tag{27}$$

where $\mathcal{L}_{\mathrm{PPO}}$ includes the usual policy, critic and entropy terms. The auxiliary loss updates the recurrent representation; the reward predictor does not supply imagined returns or select actions. There are no visual, cue or continuation prediction losses, separate task-value head, or imagined-policy teacher. Codec, memory parameters and object reader remain frozen. Only actual observations update M_t; candidate poses query the existing field without writing predictions.

A.7 Deployment cost

Persistent-state cost and storage comparison The accounting comparison in Table 4 holds the selected 9×9 scaffold fixed and changes only the Gram representation. Native Fourier moments use $8|\Gamma| = 8184$ bytes; packed symmetric and dense FP64 Grams use $8d(d+1)/2 = 109560$ and $8d^2 = 217800$ bytes. All three add the same 168960 content bytes and 12 origin bytes. These exact counts isolate persistent representation cost, not runtime.

VMF stores a fixed-dimensional episode state rather than an observation archive. For a selected support Λ, the persistent writer state consists of the Fourier difference moments h_t, the visual cross-statistic B_t, and the episode origin. With FP32 content and origin and FP64 difference moments, its storage cost is

$$\mathcal{B}(\Lambda) = 4dc + 8|\Gamma| + 12 \quad \text{bytes}, \tag{28}$$

where d is the real address dimension, c is the visual latent dimension, and

$$\Gamma = \Omega - \Omega := \{n - n' : n, n' \in \Omega\}, \qquad \Omega = \{\mathbf{0}\} \cup \Lambda \cup (-\Lambda).$$

Table 4: **Writer-state accounting.** Selected DC-initialized 9×9 scaffold with $d = 165, c = 256$; bytes exclude shared weights and temporary workspace.

Quantity	Bytes
Native writer, B	177,156
Same field with packed symmetric FP64 Gram, B	278,532
Same field with dense FP64 Gram, B	386,772

For the illustrated 9×9 scaffold, $d = 165$, $c = 256$, and $|\Gamma| = 1023$, giving a persistent writer state of

$$177{,}156 \text{ bytes}$$

per episode. This allocation is independent of trajectory length: additional real observations update the existing statistics but do not increase the number of retained entries.

For comparison, an uncompressed FP32 archive storing one 256-dimensional visual latent and a three-coordinate pose per observation grows by

$$4(256 + 3) = 1036 \text{ bytes}$$

for every additional observation. VMF therefore trades exact retention of individual observations for a fixed-dimensional sufficient-statistic representation of the ridge field.

The Fourier difference moments also reduce the persistent cost of the regression geometry relative to storing the Gram matrix explicitly. For the same $d = 165$ field, the native moment representation stores

$$8|\Gamma| = 8{,}184 \text{ bytes}$$

for the geometric statistics. A packed symmetric FP64 Gram matrix would require

$$8\, d(d + 1)/2 = 109{,}560 \text{ bytes},$$

while a dense FP64 Gram would require

$$8d^2 = 217{,}800 \text{ bytes}.$$

Including the same FP32 content cross-statistic and episode origin in all cases gives total persistent-state costs of $177{,}156$, $278{,}532$, and $386{,}772$ bytes, respectively. Thus, for this selected field, Fourier difference moments use 36.4% fewer persistent writer bytes than the packed-Gram representation.

These counts refer only to persistent episode state. Shared model parameters, temporary Gram reconstruction and factorization workspace, compiled query coefficients, and downstream reader or policy state are accounted for separately.

B Related work in detail

Reusable structure and episode content. The Tolman–Eichenbaum Machine separates relational structure from sensory associations (Whittington et al., 2020), and Vector-HaSH binds content to a grid-cell scaffold (Chandra et al., 2025). VMF instead uses supplied poses to index an explicit Fourier kernel and binds content by episode-wise regression. Reference-frame models relate spatial codes to viewpoint-dependent imagery (Byrne et al., 2007; Bicanski & Burgess, 2018); grid periodicity and geometric sensitivity motivate the address analysis, without implying cellular equivalence (Hafting et al., 2005; Barry et al., 2007; Krupic et al., 2015).

Embodied memory and reusable readouts. Neural Map writes a spatial tensor (Parisotto & Salakhutdinov, 2018); Scene Memory Transformer retrieves observation embeddings (Fang et al., 2019); GTM-SM combines spatial memory with dynamics (Fraccaro et al., 2018). MERLIN and R2I use learned predictive or recurrent states for control (Wayne et al., 2018; Samsami et al., 2024); episodic spatial world models use transition memories (He et al., 2026). VMF tests whether a bounded visual field can supply a later supervised reader without changing its writer.

Compressed queryable fields. Learned Fourier features and DeepTime relate to function approximation and episode-wise regression (Sinha & Duchi, 2016; Woo et al., 2023); MesaNet also performs test-time regression (von Oswald et al., 2026), and Interdomain Attention reads fixed-size basis coefficients (Kiyohara et al., 2026). VMF's specific choice is to store Fourier difference moments for the regression Gram matrix and select pose frequencies within a writer-byte budget. Its field predicts observation latents rather than rays or volume density as in NeRF (Mildenhall et al., 2020).

C SCAFFOLD ANALYSIS: SELECTED SUPPORT AND CAPACITY

What the learned spatial scaffold contributes The VMF scaffold determines which poses are represented distinctly and how evidence is shared between writes and later queries. For the selected 9×9 scaffold, the real address dimension is $d = 165$, corresponding to one constant component and 82 sine–cosine pairs. Among these pairs, 67 jointly couple position and heading and account for 54.7% of the nonconstant spectral weight, while 12 are purely spatial and 3 are purely angular. Aggregating conjugate spatial components across heading leaves 12 distinct nonzero spatial frequency pairs.

These mixed modes make the address at a fixed location depend on heading, while the scaffold kernel determines how strongly evidence written at one pose influences another. For two poses q and q', this interaction is governed by

$$k(q, q') = \phi(q)^\top \phi(q') = 1 + 2 \sum_j a_j^2 \cos\big(u_j^\top P(p - p')/L + \ell_j(\theta - \theta')\big) . \tag{29}$$

Pure spatial modes distinguish locations, pure angular modes distinguish headings, and mixed modes make the resulting addresses view-selective.

The same geometry also governs how new observations change the fitted field. Let

$$A_t = G_t + \lambda I,$$

and suppose a new aligned address–content pair (x, y) is added to the current regression state. Applying the Sherman–Morrison identity gives

$$\Delta W = \frac{A_t^{-1} x}{1 + x^\top A_t^{-1} x} \left(y - W_t^\top x\right)^\top . \tag{30}$$

The update therefore factors into a coverage-dependent gain and a content innovation. The term

$$x^\top A_t^{-1} x$$

reflects how well the new address is already supported by accumulated observations, while

$$y - W_t^\top x$$

measures the residual visual content not explained by the current field. Consequently, prior coverage determines where and how strongly a new observation can alter later queries, even though the scaffold kernel itself remains fixed after training.

This analysis concerns the functional role of the selected spectrum rather than the visual appearance of any individual Fourier feature. Different feature-coordinate representations can preserve the same kernel and ridge predictions, so the kernel and its induced pose interactions are the relevant objects for interpreting the learned scaffold.

We distinguish the shared pose-to-feature map $\phi(q)$, its kernel $k(q, q') = \phi(q)^\top \phi(q')$, and the episode-specific coefficients W_t. The first two form the spatial scaffold; observations change W_t and its sufficient statistics without changing that scaffold. We use synthetic probes to examine capacity and interference, and spectral diagnostics to examine the selected scaffold.

The spectral counts below describe the illustrated VMF scaffold; the capacity and update statements follow from finite regression features.

Joint position–heading structure. The selected scaffold has $d = 165$ real coordinates: one constant and 82 sine–cosine pairs. Of these, 67 couple position and heading (carrying 54.7% of nonconstant spectral weight), 12 are purely spatial, and 3 are purely angular. View addresses can therefore depend jointly on location and orientation. Their elementary spatial components are plane waves, so hexagonal firing is not implied by the parameterization.

Aggregating spatial conjugates across heading yields 12 distinct nonzero spatial frequency pairs in this selected fit. Their superposition determines the kernel's spatial cross-section. Orthogonal changes of feature coordinates can alter individual activity maps while preserving that kernel and every ridge prediction; neither an individual wave nor a visually regular pattern is evidence of biological grid-cell firing.

Bounded storage and finite representational capacity. For N observations, write $X_N = [\phi(q_i)^\top]_{i=1}^N$, $G_N = X_N^\top X_N$, and $H_N = X_N(G_N + \lambda I)^{-1} X_N^\top$. The number of fitted degrees of freedom is

$$d_{\text{eff}}(N) = \operatorname{tr}(H_N) = \sum_j \frac{\sigma_j^2}{\sigma_j^2 + \lambda} \le d, \tag{31}$$

where σ_j are the singular values of X_N. Fixed storage does not imply lossless retention of arbitrarily many independent observations. For independent zero-mean unit-variance scalar targets, the conditional training error obeys

$$\mathbb{E}[\|(I - H_N)y\|^2/N \mid X_N] = 1 - 2\operatorname{tr} H_N/N + \operatorname{tr}(H_N^2)/N \ge (N-d)_+/N. \tag{32}$$

The lower bound follows because at least $N - d$ eigenvalues of $I - H_N$ equal one. Additional observations can instead improve estimation of a function already represented by the scaffold. Sufficient statistics preserve the ridge objective, not every observation separately.

Write dynamics and interference. Let $A_t = \lambda I + \sum_{i \le t} \phi_i \phi_i^\top$ and let a new observation have feature x and aligned target y. The coefficient update is given by Eq. equation **??**. At a query q, the prediction changes by the innovation multiplied by $\phi(q)^\top A_t^{-1} x/(1 + x^\top A_t^{-1} x)$. Its sign and magnitude depend on prior coverage, and periodic aliases can couple distant poses. Repeated writes reinforce representable content while reducing the gain of each subsequent innovation; exact aliases force distinct targets to compete for the same prediction (Proposition 3 below).

Scale sensitivity. For fixed support and phase matrix, changing L to sL gives $\phi_{sL}(sp, \theta) = \phi_L(p, \theta)$. It rescales spatial wavelengths without adding capacity. On actual 15×15 trajectories, however, adding maze cells changes visual content and topology as well as extent, so this identity alone cannot predict recall.

Environmental rescaling of spatial codes (Barry et al., 2007; Krupic et al., 2015) motivates this computational comparison. VMF receives poses as inputs, so its phase-transport identity describes the address map rather than pose estimation.

Scaffold mechanism controls. Table 5 compares fitted components on the same 9×9 TRAIN500 trajectories, with a frozen codec, visual-only objective, and 177,156-byte writer ceiling. Fixed support chooses frequency pairs by a predeclared, image-independent radial order, then fits the phase matrix and amplitudes. The $Q = I$ arm instead removes latent heading alignment and reruns the entire support and continuous-parameter fit from the DC-only initialization; it is not a post-hoc deletion of Q. The packed-Gram-capacity arm reruns support search with a 117-feature cap under the packed-FP64-Gram accounting in Table 4. Its predictor is evaluated through equivalent difference moments. All arms use the same VAL100 maps without an object reader. In these fits, the selected scaffold has the lowest latent NMSE in every column. Fixed support and removal of heading alignment increase error; the packed-Gram cap admits fewer features under the same byte budget.

C.1 Algebraic properties and limitations of the scaffold

Use the unnormalized real features $\phi = [1, \sqrt{2}a_j \cos\varphi_j, \sqrt{2}a_j \sin\varphi_j]_j$, $\varphi_j = u_j^\top Pp/L + \ell_j\theta$, with raw-carrier ridge $\lambda = 0.0153$. The native implementation divides these features by $\sqrt{d}$ and uses λ/d, giving identical predictions. Let $y_i = Q(\theta_i)^\top z_i$. The following statements concern exact arithmetic and a fixed episode pose gauge.

Table 5: 9×9 scaffold controls: paired VAL100 latent NMSE (lower is better). Angle columns query held-out intermediate headings from the corresponding 10°/20°/30° gap case. KiB is the writer state charged to the common cap; the packed-Gram row is a capacity simulation. One scaffold fit per row.

Scaffold	d	KiB	Written	10°	20°	30°
Selected VMF	165	173.0	**.304**	**.284**	**.296**	**.310**
Fixed support	163	171.1	.410	.387	.434	.545
$Q = I$, refitted	163	172.9	.388	.486	.556	.616
Packed-Gram cap	117	170.9	.360	.301	.339	.330

Proposition 1 (scaffold, scale and transport). The kernel is

$$k(q, q') = 1 + 2 \sum_j a_j^2 \cos\left(u_j^\top P(p - p')/L + \ell_j(\theta - \theta')\right).$$

It is independent of episode contents. If P is invertible, the columns of $\Pi_{\mathrm{per}} = 2\pi L P^{-1}$ are common spatial periods at equal heading. They need not be shortest period vectors. Scaling L, p together preserves ϕ. Each supplied pose increment rotates every sine–cosine block by its corresponding phase increment, preserving feature norm. *Proof.* Apply $\cos a \cos b + \sin a \sin b = \cos(a - b)$, integer periodicity, and the angle-addition identities.

Proposition 2 (sufficiency of difference moments). After N writes, the ridge solution depends only on $G_N = \sum_{i=1}^N \phi_i \phi_i^\top$ and $B_N = \sum_{i=1}^N \phi_i y_i^\top$. Products of Fourier features use frequencies in $\Gamma = \Omega - \Omega$, with $\Omega = \{0\} \cup \Lambda \cup (-\Lambda)$; trigonometric moments indexed by this symmetric set reconstruct G_N. The stored cross-statistics may use unweighted carriers because the fixed diagonal amplitudes can be applied at solve time. Shared learned parameters, codec and temporary solve memory are accounted for separately.

The byte formula and complete accounting appear in Eq. equation 6 and Table 4; Eqs. equation 31–equation 32 give the separate capacity bound. For noiseless in-span targets $Y = X_N W_*$, the coefficient error is $-\lambda(G_N + \lambda I)^{-1} W_*$, so increased coverage can improve estimation in represented directions. Orthogonality of Q preserves observation-wise Euclidean errors.

Difference moments versus a packed Gram matrix. Difference moments and a packed Gram represent the same ridge field in exact arithmetic. Table 4 gives their persistent-byte costs for the selected support. Under the same writer budget, a packed FP64 Gram admits at most 117 real feature coordinates; Table 5 compares the separately fitted capacity-matched scaffold.

Amplitudes and effective regularization. Let $\phi^{(1)}$ be the same real Fourier features with unit amplitudes and let $S_a = \mathrm{diag}(1, a_1, a_1, \ldots, a_J, a_J)$, so $\phi = S_a \phi^{(1)}$. With $\widetilde{W} = S_a W$, the fitted predictions are unchanged and the objective becomes

$$\sum_i \|\widetilde{W}^\top \phi^{(1)}(q_i) - y_i\|^2 + \lambda \|S_a^{-1} \widetilde{W}\|_F^2.$$

Thus an amplitude a_j sets a coefficient penalty proportional to λ / a_j^2. Spectral weight is not a measure of information retained in that direction: amplitude growth can weaken regularization without introducing new frequencies and can impair conditioning.

Proposition 3 (update dynamics and exact aliases). The Sherman–Morrison identity applied to $A + xx^\top$ yields Eq. equation **??**. Writing the initial leverage as $\kappa_0 = x^\top A^{-1} x$, the leverage before the nth write at x is $\kappa_0/(1 + (n-1)\kappa_0)$; that write changes the prediction at x by the innovation times $\kappa_0/(1 + n\kappa_0)$. Two same-heading poses with identical features necessarily receive identical predictions and identical updates. For different scalar targets y_1, y_2 at such poses, the minimum average squared error is $(y_1 - y_2)^2/4$. A distant common period therefore creates a genuine representational ambiguity even in exact arithmetic. High but nonunit normalized kernel similarity alone does not establish this ambiguity: the full inverse Gram matrix matters.

Proposition 4 (coordinate non-identifiability and order). For any orthogonal matrix R, replacing ϕ with $R\phi$ and W with RW preserves the ridge penalty, kernel and predictions. Coordinate-wise spatial patterns can therefore change without a functional change. Additive sufficient statistics are permutation invariant in exact arithmetic provided the same origin and heading gauge are retained.

D VISUAL RECALL: PROTOCOLS AND SUPPLEMENTARY RESULTS

D.1 PROTOCOL AND METRIC DETAILS

The final 64 9×9 maps exclude TRAIN wall geometries. Test outcomes did not select models, maps or masks.

Figure 5 displays the complete Paired-64 map cohort used for held-out evaluation. Each recorded wall layout is shown through the Memory Maze overhead camera, in evaluation order.



Figure 5: **Paired-64 evaluation maps.** All 64 held-out 9×9 wall layouts, indexed in evaluation order and rendered with the official Memory Maze overhead camera. Wall geometry matches the recorded episodes; decorative textures are renderer-generated.

An N-action prefix contains $N + 1$ observations. Canonical queries use the first occurrence of each rounded world pose. Early targets come from the first 101 frames; only later writes can be removed by the no-refresh control. One uniform random value per frame creates nested holdout masks, using seed 74119 plus map index. Frame zero stays in support. The nonduplicate subset uses the same 0.001 pose quantization.

Yaw tests remove route writes within 0.5 cells of the anchor and strictly inside the angular gap. Only true-rendered endpoints are inserted, at the anchor and before future observations. Only interior angles are scored. Scene MSE excludes a four-pixel border. Angular-difference MSE compares adjacent 1° image differences within the gap, without dividing by angular step size; it is not a derivative expressed per radian.

Latent NMSE is the ratio of pooled squared error to pooled target energy across maps, not an average of per-map NMSEs. Figure 2b,c shows pointwise sample SD across three fitted memories for each method; panel (d) shows the same fit-to-fit SD for VMF depth, with single-fit comparison bars. Main Table 1 and Table 9 also use three-fit sample SD. The supplementary RGB, holdout and angle-sweep tables use three-fit SD only for VMF; their comparison rows are single fits. PSNR is $-10 \log_{10}(\mathrm{MSE})$ for RGB in $[0, 1]$.

Comparison scores come from local implementations evaluated on shared data, not from the cited papers. Trained ordinary baselines use 500 training and 100 validation episodes, the frozen codec and 40 epochs. DeepTime (Woo et al., 2023), Scene Memory Transformer (SMT) (Fang et al., 2019), and Fourier ridge (Rahimi & Recht, 2007) supply architectural precedents; RBF ridge, nearest stored view and pose kernel are local controls. Matched-supervision variants also receive VMF's fixed Q and half-full/half-angle objective. The supplementary RGB, holdout and angle-sweep tables report one fitted comparison memory per row; their VMF rows summarize three independently fitted scaffolds. Table 9 and main Table 1 report three-fit comparison scores.

Table 6: **RGB fidelity after recall.** Random 20% holdout at 2048 actions on 64 maps; the last column restricts to nonduplicate targets. RGB MSE is in 10^{-3} units and PSNR in dB. VMF reports mean and sample SD over three fits; comparison memories use one fit. Codec-only reconstruction is a privileged reference.

Memory	RGB MSE ↓	PSNR ↑	SSIM ↑	Nondup. NMSE ↓
VMF	$\mathbf{7.760 \pm 0.045}$	$\mathbf{21.101 \pm 0.025}$	$\mathbf{0.7356 \pm 0.0004}$	$\mathbf{0.497 \pm 0.001}$
DeepTime	10.604	19.750	0.686	0.634
RBF ridge	11.617	19.350	0.673	0.669
Fourier ridge	10.707	19.700	0.686	0.640
Nearest stored	19.815	17.030	0.610	1.223
Pose kernel	14.205	18.480	0.646	0.831
SMT	12.026	19.200	0.676	0.654
VMF, frozen $Q = I$	9.381	20.280	0.702	0.610
DeepTime matched Q	10.901	19.630	0.682	0.640
Fourier matched Q	9.897	20.050	0.700	0.588
SMT matched Q	19.610	17.080	0.580	0.977
Codec reconstruction	0.120	39.23	0.983	—

Table 7: **Random-holdout sensitivity.** Columns give withholding probability and observed-action length. VMF reports mean and sample SD over three fits; comparison memories use one fit. Masks are nested and paired, and held-out targets never enter the writer.

1000 actions			
Memory	10%	20%	40%
VMF	$\mathbf{0.458 \pm 0.003}$	$\mathbf{0.482 \pm 0.004}$	$\mathbf{0.586 \pm 0.015}$
DeepTime	0.633	0.646	0.696
RBF ridge	0.663	0.670	0.694
Fourier ridge	0.636	0.648	0.692
Nearest stored	1.153	1.155	1.158
Pose kernel	0.778	0.780	0.781
SMT	0.608	0.608	0.609
VMF, frozen $Q = I$	0.619	0.650	0.793
DeepTime matched Q	0.617	0.632	0.676
Fourier matched Q	0.558	0.571	0.608
SMT matched Q	0.960	0.960	0.961
2048 actions			
Memory	10%	20%	40%
VMF	$\mathbf{0.491 \pm 0.002}$	$\mathbf{0.4972 \pm 0.0015}$	$\mathbf{0.530 \pm 0.002}$
DeepTime	0.631	0.634	0.655
RBF ridge	0.668	0.669	0.685
Fourier ridge	0.638	0.640	0.660
Nearest stored	1.231	1.223	1.222
Pose kernel	0.834	0.831	0.827
SMT	0.656	0.654	0.651
VMF, frozen $Q = I$	0.601	0.610	0.650
DeepTime matched Q	0.636	0.640	0.660
Fourier matched Q	0.586	0.588	0.607
SMT matched Q	0.976	0.977	0.972

D.2 Delayed heading interpolation

Table 8 measures image error and angular variation after a long delay.

Table 8: **Delayed 40° heading gap.** This early-anchor probe writes endpoints at action 100 and queries interior views at 2048; Figure 2c instead anchors at action 500. VMF reports mean and sample SD over three fits; comparison memories use one fit. Image-based MSE columns are multiplied by 10^3; endpoint blend is a privileged diagnostic with no registered scene or angular-difference score.

Memory	NMSE	RGB 10^3	Scene 10^3	Angular diff. 10^3
VMF	$\mathbf{0.588 \pm 0.002}$	$\mathbf{9.771 \pm 0.171}$	$\mathbf{6.560 \pm 0.318}$	0.525 ± 0.003
DeepTime	0.751	13.485	9.893	0.571
RBF ridge	0.763	14.321	10.336	0.525
Fourier ridge	0.766	13.542	10.087	0.530
Nearest stored	1.446	25.364	20.807	2.312
Pose kernel	0.975	18.163	13.971	0.539
SMT	0.775	15.402	11.043	0.538
VMF, frozen $Q = I$	0.742	12.264	9.084	0.523
DeepTime matched Q	0.803	15.376	10.773	0.930
Fourier matched Q	0.687	12.466	8.595	0.535
SMT matched Q	0.962	20.826	16.938	0.524
Endpoint blend (privileged)	0.535	5.191	—	—

RBF and frozen $Q = I$ have slightly lower angular-difference error than VMF despite larger image errors, illustrating that smoothness and correctness differ.

D.3 Across-fit recall

Table 9 reports fit-to-fit variation for the comparison memories; its matched-Q rows use the same scores and precision as main Table 1.

Table 9: **Three-fit baseline recall.** Written and random-20% withheld NMSE at 2048 actions on the same 64 maps; mean ± sample SD across fitted initializations.

Memory	Written NMSE	Withheld NMSE
DeepTime	0.548 ± 0.005	0.636 ± 0.006
RBF ridge	0.602 ± 0.000	0.669 ± 0.000
Fourier ridge	0.549 ± 0.004	0.639 ± 0.005
Pose kernel	0.780 ± 0.001	0.829 ± 0.003
SMT	0.639 ± 0.006	0.658 ± 0.005
Matched heading alignment and visual objectives		
DeepTime matched Q	0.557 ± 0.017	0.642 ± 0.020
Fourier ridge matched Q	0.519 ± 0.012	0.598 ± 0.015
SMT matched Q	0.974 ± 0.010	0.975 ± 0.008

D.4 Depth readout of recalled images

Visual recall can preserve appearance while distorting recoverable geometry. We probe this distinction with a frozen Depth Anything 3 (DA3) model (Lin et al., 2026). The protocol accumulates observations 0–1000 of one episode, freezes each memory, and recalls the same 32 recorded poses in four shared eight-view groups. DA3-LARGE receives the same cameras, intrinsics, crop, resolution and inference settings for every image source. Original RGB and codec-only reconstructions are privileged references.

Table 10: **Single-episode depth readout.** DA3 AbsRel on 32 shared views after 1000 actions. VMF reports mean ± sample SD over three fitted memories; comparison memories use one fit.

Image source	AbsRel ↓
Original RGB (reference)	0.125
Codec only (reference)	0.139
VMF	$\mathbf{0.252 \pm 0.039}$
Fourier + matched Q/angle	0.436
DeepTime + matched Q/angle	0.833
SMT + matched Q/angle	3.355
Pose kernel	0.488
Joint pose RBF	1.207

True simulator depth is used for scoring, not DA3 inference or scale fitting. All methods use the same valid-ray mask: finite ground-truth optical-axis depth between 0.18 and 4 cells. Writing predicted depth as $\hat{\zeta}$ and simulator depth as $\zeta^\star$, AbsRel averages $|\hat{\zeta} - \zeta^\star|/\zeta^\star$ within each view and then across the 32 views. The VMF mean and sample SD span three fitted memories, as in Figure 2d; comparison rows use one fit.

VMF has the lowest AbsRel among the tested memories (Table 10); the original-image and codec-only references score lower. The eight-view 3D illustration in Figure 2g is separate from the 32-view metric cohort in panel (d).

D.5 QUALITATIVE PATH AND VISUAL-READ PROTOCOL

Map 0 supplies the qualitative path: writing stops at action 1000, and queries follow recorded poses at actions 32–88. Figure 2a marks the trajectory; panels (e,f) display observed-path and unwritten-view queries from the same frozen states. Image memories share observations, poses and codec; readout does not modify their states.

DA3-LARGE receives the same eight RGB views at actions $32, 40, \ldots, 88$ for each source. Inputs are native 64×64 images with a common four-pixel task-border crop, internally processed at resolution 504 and returned to 56×56. Supplied camera poses set scale through the official pose-alignment path; simulator depth is never an input to DA3 or fitted to its output.

Point clouds use a common display mask of finite simulator-depth rays in $(0.18, 4)$ cells; this mask is not given to the memories. The display box is $[-1, 10] \times [-1, 10] \times [-2, 2]$ cells, with out-of-box points counted explicitly. Predicted depth and RGB receive no method-specific geometric adjustment. All panels share a camera and raster crop, so the geometry reflects recall followed by the frozen depth estimator.

E OBJECT LOCALIZATION: EVALUATION DETAILS

E.1 TWO READOUTS OF ONE FROZEN STATE

The paired 64-map protocol accumulates observations through action 1000, freezes the VMF state, and applies written-view recall and three-object localization to that same state. Object queries use field features and visitation moments, not the observation archive; pose and the static coordinate frame remain supplied inputs. Reversing the readout order leaves the state and both outputs unchanged. Figure 3 visualizes the evolving object readout on one map; its entropy and error traces describe that episode, while Table 2 reports the cross-map evaluation.

E.2 SUPRISE BONUS IN POLICY LEARNING

We cross the three policy inputs with a visual-surprise bonus. For a real observation at pose q_t, let μ_t be the field's prediction before the write, $z_t \in \mathbb{R}^{256}$ its standardized codec latent, and $u_t = \phi(q_t)^\top (G_{t-1} + \lambda I)^{-1} \phi(q_t)$ the pre-write ridge leverage. Under unit observation variance, the mean posterior KL per latent channel is

$$b_t = \frac{1}{2} \left[\log(1 + u_t) - \frac{u_t}{1 + u_t} + \frac{u_t \|z_t - \mu_t\|_2^2}{256(1 + u_t)^2} \right]. \tag{33}$$

Bonus conditions train on $r_t^{\text{ext}} + \beta_t b_t$. The weight β_t is 0.01 through 500,000 real actions, follows a cosine decay to zero at 700,000, and is then zero; it is held fixed within each 128-step rollout. Even bonus-enabled PPO consults VMF to compute b_t, although it receives no memory readout. Reward prediction and reported returns use external reward alone.

E.3 READER PROTOCOL, SPLITS AND UNCERTAINTY

The registered frozen-memory comparison uses a frozen codec, measured poses, 500 training episodes, 64 causal training prefixes per episode and a fresh 100-map validation set. Each input condition receives separately fitted width-64 coordinate readers (AdamW, learning rate 10^{-4}, batch 16, at most 60 epochs, patience 10). Validation squared coordinate error selects checkpoints; test errors do not. Object labels never update a visual memory. Full-history, restricted-window, geometry and alternative-memory readers use 256 spatial Sobol nodes and 32 headings under a fixed evaluation query seed. The current-observation reader instead uses its observed pose and outputs one Gaussian per object. Window controls rebuild their fields from bounded latent–pose rings and use window-local visitation statistics; their temporary workspace is not retained state. Each reader is fitted to its own input, not transferred from VMF.

Wall layouts are compared modulo D4 rotations and reflections. Fresh validation excludes TRAIN geometry; Clean-865 removes 135 TRAIN-geometry-overlapping trajectories from the official 1000-trajectory test. Paired-64 has no TRAIN geometry overlap but shares 15 geometry classes

Table 11: **Reader inputs and persistent state.** The three full-history VMF fits retain 175,300–177,156 bytes. Bytes exclude shared reader weights, the common static frame input and temporary reconstruction workspace.

Reader input	Retained information	State bytes
VMF, full history	bounded visual field	175,300–177,156
Fourier, matched Q	bounded visual field	183,436
Pose kernel	bounded visual field	185,204
VMF, last 64 observations	recent latent–pose ring	66,332
Geometry / visitation	pose moments only	8,404
Current observation	no persistent episode state	0
GRU	recurrent hidden state	512
Bounded attention	retained latent–pose tokens	189,296

with the official archive. The populations and temporal windows are reported separately. The VMF readers use this split protocol.

The full-history VMF row reports means and sample SDs across three independently fitted memory–reader pairs. Comparison memories are fitted under their stated protocols. Pointwise 95% intervals resample whole maps, grouping objects and prefixes within maps (4000 draws). On Clean-865 actions 501–1000, VMF has a mean across fits of 1.2085 ± 0.0708 RMSE; pooling their squared errors instead gives 1.210 with whole-map interval $[1.140, 1.280]$.

The untrained arena-center reference predicts $(4.5, 4.5)$ in the supplied 9×9 world frame, transformed to current-agent coordinates; uniform guessing draws independently from $[0, 9]^2$. Its score uses expected squared error before the square root. Neither reference sees images, memory or walls.

E.4 Within-size 15×15 extension

Three VMF scaffolds were fitted on 15×15 training trajectories from DC-only support, with a six-object reader trained for each frozen scaffold. The codec and heading frame remained fixed, while the scaffold and reader used the size-15 coordinate chart. On the official 1000-maze evaluation, object localization over actions 501–1000 achieves 5.224 ± 0.246 cells RMSE (mean $\pm$ sample SD across the three memory–reader pairs). This measures the same construction at a larger maze size and object count.

For published context, S5WM reports Objects MSE 1.8 on 9×9 and 25.3 on 15×15 maps in its own protocol (Deng et al., 2023). Its splits, supervision, metric and state budget differ from this study's RMSE comparison, so these figures are not a matched ranking.

E.5 Task-supervised sequence controls

The GRU (Cho et al., 2014) and bounded-attention (Vaswani et al., 2017) controls use the same codec, training and fresh-validation maps, with three fitted seeds each. Their sequence encoders and coordinate heads receive object-label gradients directly, unlike frozen-memory readers; these are local adaptations, not published scores. Both use width 128, AdamW 10^{-4}, batch 16, at most 60 epochs and patience 10. GRU backpropagates through the full episode; attention retains 182 latent–pose tokens. Evaluation streams each declared bounded state, and causal streaming-equivalence checks pass. The main table reports both their terminal-prefix and common actions-501–1000 scores. Their different supervision, state budgets and optimization limit what their errors imply about memory architecture.

F Control implementation and evaluation

The control study uses a fixed VMF scaffold, Objects-supervised reader and visual codec. Policy training changes none of these parameters; only real observations update episode statistics, and proposed poses never write.

The six arms cross PPO, Local recall and Global object belief with bonus off/on. They share executed-action reward prediction, the frozen codec, supplied pose and cue, and matched base initialization. Local and Global use the same frozen VMF; Global also changes motion-proposal inputs. The comparison therefore concerns complete policy pathways.

For bonus arms, surprise is the *mean* per-channel $\mathrm{KL}(\text{new posterior} \,\|\, \text{old posterior})$ in standardized visual coordinates with unit likelihood variance (Eq. 33), evaluated on the real observation before its write. PPO trains on $r_t^{\mathrm{ext}} + \beta_t b_t$. With n_{act} the real-action count at rollout start, the bonus weight is 0.01 through $500{,}000$ actions, $0.005[1 + \cos(\pi(n_{\mathrm{act}} - 500{,}000)/200{,}000)]$ for $500{,}000 < n_{\mathrm{act}} < 700{,}000$, and zero thereafter. The weight is constant within a rollout. The prediction head and reported returns use external reward only. PPO receives no memory readout; its bonus arm uses VMF only to compute surprise.

Recurrent PPO (Schulman et al., 2017) uses a width-256 state, 64-step fragments, eight environments, 128-step rollouts, four epochs, four sequences per minibatch and Adam at 3×10^{-4}. A 128-SiLU head predicts executed-action external reward with loss $0.1\,\mathrm{SmoothL1}(\operatorname{asinh} \hat{r}_t, \operatorname{asinh} r_t^{\mathrm{ext}})$. Global uses the current visual latent rather than recalled future-pose latents for recursive proposals. Local and Global have equal allocated parameters, though disabled branches receive no gradients; ordinary PPO is not parameter-matched to them.

F.1 Training curves

Figure 4 reports trailing-128-episode external return as mean $\pm$ sample SD across three policies per condition over 2 million training actions.

F.2 Frozen-policy results on new maps

One no-bonus policy per condition was evaluated after 2 million training actions on 64 new, D4-distinct maps disjoint from training and validation. Three shared action-uniform streams per map started with fresh memory. Policy, codec, memory, reader and motion-model parameters were frozen throughout evaluation. Returns include external reward only.

Table 12: **Held-out control for one no-bonus policy per condition.** Mean external return and whole-map bootstrap 95% confidence interval ($10{,}000$ resamples, preserving paired action repetitions).

Frozen policy	Return, steps 1–1000	Return, steps 501–1000
PPO	1.010 $[0.849, 1.188]$	0.422 $[0.333, 0.516]$
Local	2.703 $[2.417, 3.005]$	1.281 $[1.109, 1.464]$
Global	10.146 $[8.625, 11.656]$	6.500 $[5.495, 7.500]$

Full-episode paired differences are 1.693 $[1.411, 1.984]$ for Local minus PPO, 7.443 $[5.995, 8.901]$ for Global minus Local, and 9.135 $[7.646, 10.630]$ for Global minus PPO. These intervals resample maps, not training initializations, and apply to the frozen VMF configuration described above.