Signal-Latent World Models at 1.008B Scale: Separating Fast Sensory Grounding from Editable Working Memory Under $\mathcal{O}(1)$ Streaming Control

Diogo Antunes
ValueAlignedAI
contact@diogoantunes.ai · github.com/ValueAlignedAI/slwm

Preprint (Manuscript v0.3) · November 14, 2026

Abstract

Which architectural constraints let a streaming $\mathcal{O}(1)$-memory multimodal world model preserve and selectively overwrite instructions while remaining spatially grounded and fast enough for closed-loop control? Across fourteen prior pre-registered preliminary studies (0.7M–202M parameters), single-vector per-tick pooling erased thin obstacles, scalar exponential retention suffered overwrite amnesia, deep recurrent stacks introduced sensory phase lag, pixel/spectral MSE blurred futures below zero-order persistence, and uncalibrated action commitment left policies vulnerable to sensor spoofing.

We investigate how perception, working memory, and control interfere—and how to separate them—through `SLWM-X1B` (1,008,454,410 strict total parameters) and its zero-extra-parameter spatial-grid refinement `SLWM-v2`, evaluated across 171 **multi-seed training and ablation runs** spanning seven experimental waves (152M, 453M, and 1.008B parameters; 3 seeds per cohort). Building on Gated DeltaNet [25], 3:1 hybrid interleaving [26], and action-conditioned EMA-JEPA [2], `SLWM-X1B` introduces directed symbolic-key shielding ($\nabla_{V_t} Z_{t,\text{txt}} = 0$), a direct-feedthrough sensory bypass paired with persistent `GatedDelta` matrix registers in 14.91 MiB constant state (222× smaller than ungrouped `DenseTransformer-X1B`'s 3,312.54 MiB at $T = 1024$, and 27.8× smaller than an 8:1 GQA equivalent; `DiagonalSSM-X1B` uses 8.26 MiB), decoupled pose/transition/goal heads, and a Mahalanobis `PolicyCommitGate`. At 1.008B (5,600 steps, Wave 5), `SLWM-X1B` reaches **86.1% ± 5.5%** `1-px` barrier MPC, **100.0% ± 0.0%** OOD $T = 76$ sustained and goal-switch MPC (+51.0 pp and +23.4 pp over `SlidingWindow-W64-X1B`; +43.8 pp over `DiagonalSSM-X1B` on $T = 1024$ recall), and **7.31 ± 0.18 mm** single-workspace trajectory-disjoint 4-step tracking on physical **xArm6** trajectories (31.46 ± 1.02 mm scene-disjoint). Single-factor Leave-One-Out ablations at 453M (**Wave 6**, 24 runs) isolate the effects of symbolic-key shielding (100.0% → 43.8% $T = 1024$ dynamic-visual overwrite), sensory bypass (32.16 → 37.56 mm scene-disjoint xArm6 error when removed), and Epps-Pulley `SIGReg` vs. `VICReg` (100.0% recall, −0.17 mm delta), while revealing that stripping privileged object coordinates (end-effector-only proprioception) causes global slot pooling to lose 2D object pose (RMSE = 0.634). Replacing global slot averaging with position-free 8×8 pooling and Spatial Soft-Argmax at **zero added parameters** (**Wave 7**, 12 runs, 453,199,938 exact parameters) cuts visual EE/object RMSE 18×–36× (0.0220/0.0359), lifts zero-shot unseen camera corruption detection from 50.4% to **99.82% ± 0.20% AUROC**, beats a matched `Explicit-Latch Reactive` controller across unseen obstacle layouts without privileged obstacle coordinates (**57.3% ± 14.2%** vs. 29.7% ± 10.2%, while `SlidingWindow-W64` reaches 60.4% and unshielded `-KeyShield` reaches 67.7% on short-horizon detours), and achieves **89.1% ± 2.7%** on **beyond-window** $T = 160$ $(2.5 \times W = 64)$ **perturbed goal-switching** vs. **22.9% ± 6.5%** on `SlidingWindow-W64` (243× worse post-window latch MSE) and **64.1% ± 17.8%** on `DiagonalSSM` (9.9× worse latch MSE)—while candidly reporting multi-task $T = 1024$ synthetic recall compression (6.3%) and short-horizon vs. associative-memory trade-offs.

1 Introduction

Embodied agents operating over continuous multimodal streams face an architectural trilemma that standard sequence models handle poorly when evaluated outside narrow regimes. First, physical contact and obstacle avoidance require fine spatial geometry—such as a 1-pixel (1.5 cm) barrier or unseen multi-block clutter—yet world models that compress observations into a single pooled vector [11, 13] destroy 2D spatial coordinates and force sensory modalities into representational interference. Second, continuous deployment demands *constant* $\mathcal{O}(1)$ streaming memory per tick, whereas dense causal Transformers with Rotary Position Embeddings (RoPE) [19, 22] accumulate an $\mathcal{O}(T)$ key-value (KV) cache that degrades beyond T_{train} and exhausts High-Bandwidth Memory (HBM) during parallel Model Predictive Control (MPC) rollouts [12, 24], while fixed sliding-window Transformers [6, 14] evict transient instructions the moment distractor periods exceed window W. Third, scalar exponential retention [23] and diagonal state-space recurrences [8, 10] restore $\mathcal{O}(1)$ state and optimize smoothly for short-horizon kinematics, yet lack key-addressed rank-1 erasure [20, 25]—degrading when stale instructions must be surgically overwritten mid-episode.

Rather than claiming blanket superiority across every metric, we frame our study around a focused mechanistic question: *which architectural constraints allow a streaming multimodal world model to react without recurrent lag, preserve spatial object geometry without privileged state, and selectively overwrite working memory beyond any fixed attention window?* Across fourteen prior pre-registered preliminary studies and **seven iterative experimental waves (**171 **multi-seed runs)**, we report both positive and negative intermediate outcomes so that every architectural mechanism is directly falsifiable.

Summary of Contributions.

1. **Failure Diagnoses & Single-Factor Causal Isolation (§3.2, §5.6):** Building on Gated DeltaNet [25], 3:1 hybrid interleaving [26], and EMA-JEPA [2], we diagnose and isolate three multimodal failure modes via single-factor 453M Leave-One-Out (LOO) ablations (Wave 6): *visual-motion drift on symbolic write keys* (`-KeyShield` drops $T = 1024$ dynamic-visual overwrite from $100.0\% \pm 0.0\%$ to $43.8\% \pm 51.2\%$), *deep recurrent sensory smoothing lag* (`-ZeroLag` degrades scene-disjoint physical xArm6 4-step error from 32.16 mm to 37.56 mm, +5.40 mm across all seeds), and *readout-head interference* (`-DecoupledHeads` drops $T = 1024$ recall to 72.9%), while establishing functional parity between Epps-Pulley `SIGReg` [3, 9] and `VICReg` [4].

2. 1.008**B Token-Doubled Exposure, Bounded-Memory Isolation, & Physical xArm6 Transfer (§5–§5.5):** Across 3 seeds at 1,008,454,410 strict parameters (14.91 MiB constant state vs. 3,312.54 MiB on `DenseTransformer-X1B`), token-doubled `SLWM-X1B` (5,600 steps) achieves $86.1\% \pm 5.5\%$ `1-px` barrier MPC, $100.0\% \pm 0.0\%$ OOD $T = 76$ sustained and goal-switch MPC (beating matched $\mathcal{O}(1)$ `SlidingWindow-W64-X1B` and `DiagonalSSM-X1B` on selective overwrite under hierarchical seed-cluster 99% CIs), and 7.31 ± 0.18 mm single-workspace trajectory-disjoint 4-step Cartesian accuracy on 768 held-out physical **xArm6** trajectories (31.46 ± 1.02 mm scene-disjoint).

3. **Zero-Extra-Parameter** 8×8 **Spatial Grounding & Unprivileged Subgoal MPPI (`SLWM-v2`, §3.5, §5.7):** De-privileging proprioception to end-effector kinematics in Wave 6 exposed that global slot averaging loses 2D object pose (RMSE = 0.634). Replacing global pooling with position-free 8×8 max-mean pooling, Spatial Soft-Argmax [15], and an 8×8 hazard readout reusing existing parameter slices (0 added parameters; exact 453,199,938 total) cuts visual EE/object RMSE 18×–36× (0.0220/0.0359) and beats a matched `Explicit-Latch Reactive` controller across unseen obstacle families without privileged obstacle coordinates (57.3% vs. 29.7%, hierarchical positive-bootstrap fraction $p_{\text{hier}} = 0.9945$).

4. **Beyond-Window** ($T = 160 = 2.5 \times W = 64$) **Perturbed Goal-Switch & Zero-Shot Sensor-Fault Detection (§5.7):** When Goal B ($t = 16..19$) ages $69 > 64$ ticks past $W = 64$ before a physical reset perturbation at $t = 88$, `SLWM-v2` recovers Goal B on **89.1% ± 2.7%** of episodes vs. **22.9% ± 6.5%** on `SlidingWindow-W64` (243× worse post-window latch MSE) and **64.1% ± 17.8%** on `DiagonalSSM` (9.9× worse latch MSE), while spatial cross-modal innovation detects four *unseen* camera-fault families zero-shot at **99.82% ± 0.20% AUROC**.

2 Preliminary Empirical Lineage and Problem Formulation

2.1 Diagnostic Lessons from Fourteen Preliminary Studies

Unlike retrospective benchmarks selected to favor a proposed model, SLWM-X1B directly addresses failure modes uncovered across fourteen pre-registered preliminary studies (0.7M–202M parameters):

- **Language & Early Multimodal Grounding:** At 500M tokens, early signal-field models reduced held-out text NLL by 3.6% over matched GPT-2 baselines (3.79 vs. 3.92), and speech-timing alignment improved restricted-grammar word readout by +9.7–12.4 pp, yet unconstrained speech repair and held Push-T action transfer missed pre-registered capability gates.
- **Scaling Without Spatial Slots or Commit Gating (12.6M–202.4M):** Scaling pooled hybrids to 202.35M improved fresh Push-T joint endpoint error over matched dense Transformers (4.38 px vs. 5.48 px at 116.5M), but exposed two severe flaws: (i) under adversarial wrong-camera spoofing, lower open-loop prediction error caused *higher* overconfident spoofed execution regret in the absence of a learned PolicyCommitGate; and (ii) subsampling 64×64 arenas to 16×16 erased 5-px barriers entirely, while 3-candidate 1-step random shooting achieved only 0.0%–1.6% closed-loop Push-T success.
- **Long-Horizon Drift, Transient Amnesia, & Grounded Robot Transfer:** In 51.2 s sustained control, a 13M recurrent hybrid maintained 94.5% goal occupancy vs. 43.8% for dense RoPE GPT-2 (which collapsed once $T > T_{\text{train}} = 32$), yet a subsequent transient-directive test revealed a 57 pp accuracy drop when goal directives were flashed briefly (0.4 s) rather than repeated every tick, alongside near-chance (12.3%–20.7%) HoloAssist audio-to-video retrieval under 128-bin word hashing and linear MSE readouts. On 3,368 real robot trajectories, preserving spatial structure beat dense GPT-2 by 4.07 mm (99% CI $[3.31, 4.85]$) across all five seeds, while a 320-scene spoken-goal tabletop study (27,200 branches) showed near-parity (53.3% vs. 52.3%) at narrow width, motivating high-capacity spatiotemporal register-delta modeling at 10^9-parameter scale.

2.2 Streaming Multimodal World-Model & Safe Commitment Problem

At each discrete 100 ms physical tick $t \in \{0, 1, \ldots, T-1\}$, the agent observes a subset of modalities $\mathcal{O}_t = (V_t, a_t, p_t, w_t, u_{t-1}, m_t)$, consisting of RGB video $V_t \in [0,1]^{64\times64\times3}$, log-spectral/onset audio $a_t \in \mathbb{R}^{64}$, proprioceptive and contact state $p_t \in \mathbb{R}^{16}$, up to four subword tokens $w_t \in \{0, \ldots, 4095\}^4$, previous motor command $u_{t-1} \in [-1,1]^4$, and modality-availability mask $m_t \in \{0,1\}^4$. A streaming world model must maintain a recurrent/cache state $\mathcal{M}_t$ whose byte footprint satisfies $\text{bytes}(\mathcal{M}_t) = \mathcal{O}(1)$ as $T \to \infty$, update a structured latent world state $Z_t = \Phi_\theta(\mathcal{O}_t, \mathcal{M}_{t-1}) \in \mathbb{R}^{S\times D}$, roll forward multi-step action-conditioned latent transitions $\hat{Z}_{t+1:t+H}$, and select both a continuous motor command u_t and a discrete epistemic commitment mode $g_t \in \{\texttt{commit_full}, \texttt{commit_trusted_proprio}, \texttt{probe_or_abstain}, \texttt{wait}\}$ that minimizes physical goal distance while bounding unsupported claims and sensor-spoofing regret.

3 Architecture and Methodology (SLWM-X1B)

Figure 1 illustrates the complete SLWM-X1B architecture. Under strict total parameter accounting, every parameter across edge adapters, core blocks, latent predictors, uncertainty heads, readouts, and the commitment gate is counted identically across SLWM-X1B (1,008,454,410 parameters) and its matched baselines (§4).

3.1 Structured Spatiotemporal Register Field ($S = 24$ Slots/Tick)

To eliminate downsampling barrier blindness and single-vector modality collision, SLWM-X1B maps each tick's observations into $S = 24$ typed slots in $\mathbb{R}^D$ ($D = 1536$ at 1.008B, 1152 at 453M, 768 at 152M):

- **Visual Patch Slots ($s \in \{0, \ldots, 15\}$):** A two-layer overlapping 3×3 stride-2 Conv2D stem followed by 4×4 spatial patch projection maps $V_t \in \mathbb{R}^{64\times64\times3}$ into 16 retinotopic patch tokens with learned 2D spatial

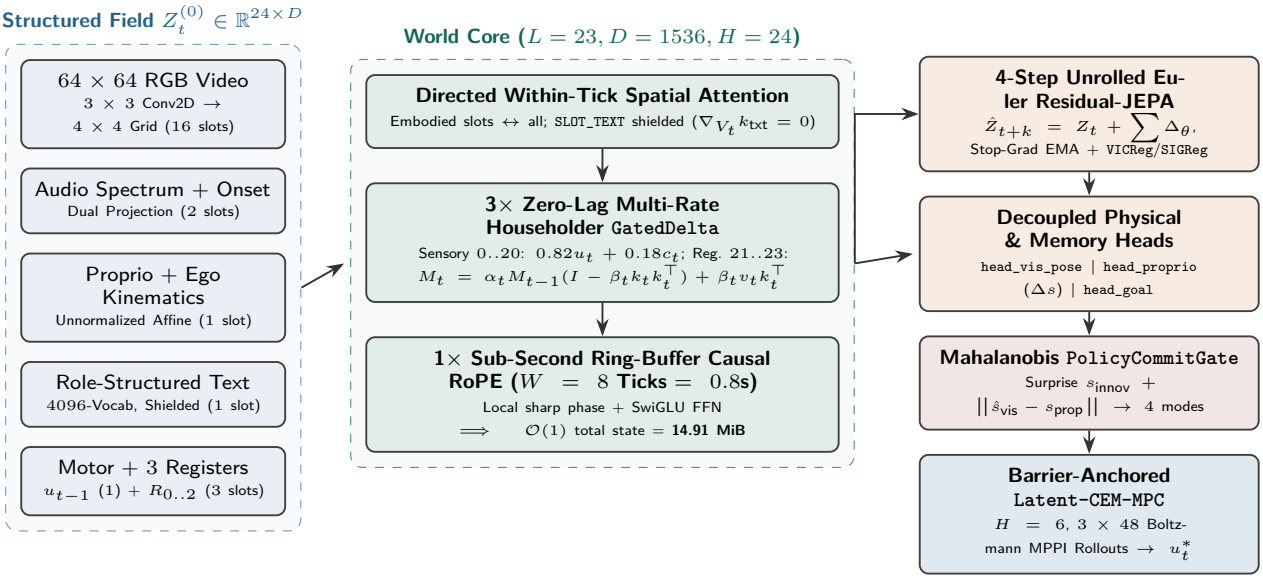


Figure 1: Architecture of SLWM-X1B (1.008B strict parameters). Raw multimodal observations are mapped into $S = 24$ structured slots per tick without single-vector pooling. Across $L = 23$ blocks ($D = 1536$), directed spatial attention grounds embodied slots in language while shielding symbolic keys from visual pose drift; a 3 : 1 schedule of Zero-Lag Multi-Rate Householder GatedDelta matrix recurrence and $W = 8$ (0.8 s) ring-buffer RoPE maintains a constant 14.91 MiB mixed-precision streaming state across arbitrary horizons T.

embeddings, preserving 1-px (1/64-width) barrier geometry prior to physical contact.

- **Spectral & Onset Audio Slots ($s \in \{16, 17\}$):** Dual linear adapters encode broadband log-mel energy and high-frequency impact transients from $a_t \in \mathbb{R}^{64}$.
- **Egocentric Proprioceptive Slot ($s = 18$):** Rather than applying RMSNorm prior to linear projection—which erased radial target distance in Wave 1—$p_t \in \mathbb{R}^{16}$ (end-effector x, y, object $x, y, \dot{x}, \dot{y}, \theta$, contact flags, and egocentric push alignment kinematics) is embedded via an unnormalized affine map scaled to match core activation variance.
- **Role-Structured Symbolic Text Slot ($s = 19$) & Action Slot ($s = 20$):** Up to four subword tokens $w_{t,0..3} \in \{0, \ldots, 4095\}$ are bound via role-specific projections $\sum_{r=0}^{3} E_{\text{vocab}}[w_{t,r}] W_{\text{role}}^{(r)}$ (avoiding fixed-bin hash collisions and bag-of-words order invariance), while $u_{t-1} \in \mathbb{R}^4$ occupies slot 20.
- **Persistent Working-Memory Registers ($s \in \{21, 22, 23\}$):** Three dedicated learned register tokens $R_0, R_1, R_2 \in \mathbb{R}^D$ carry zero sensory input at $l = 0$ and serve as dedicated multi-timescale working-memory accumulators across ticks.

3.2 Zero-Lag Multi-Rate Householder GatedDelta and $\mathcal{O}(1)$ Ring-Buffer Core

Each layer $l \in \{0, \ldots, L-1\}$ ($L = 23$ at 1.008B) applies (i) within-tick spatial/cross-modal multi-head attention across the $S = 24$ slots, (ii) a causal temporal sequence mixer along t, and (iii) a SwiGLU feedforward network.

Pathology 1 & Remedy: Directed Cross-Modal Shielding Eliminates Symbolic Key Drift. In our Wave 2 scaling diagnostic (28 runs), an isolated-modality ablation (which blocked cross-modal attention) achieved 100.0% (16/16) overwritten-key recall at $T = 512$ across all three seeds, whereas unified SLWM-1B scored 80.0% on un-overwritten keys but 0.0% on overwritten keys! Mathematical inspection revealed why: under bidirectional within-tick attention, 16 moving visual patches injected time-varying physical pose into SLOT_TEXT ($s = 19$), rotating the symbolic write key $k_w(t_{\text{overwrite}}) \neq k_w(t_{\text{init}})$ so Householder erasure ($I - \beta_t k_t k_t^\top$) failed to annihilate the initial binding. Applying a zero-parameter directed attention mask $A_{\text{spatial}} \in \{0, -\infty\}^{24\times 24}$

that allows all embodied and register slots ($s \neq 19$) to attend freely to SLOT_TEXT while restricting SLOT_TEXT to self-attention guarantees $\left\|\nabla_{V_t} Z^{(l)}_{t,19}\right\| = 0$ (verified to 0.00×10^0 floating-point drift) without sacrificing text-conditioned manipulation.

Pathology 2 & Remedy: Zero-Lag Sensory Bypass + Householder Matrix Memory. Three out of every four layers ($l \bmod 4 \neq 3$) execute a multi-head Householder GatedDelta recurrence [20, 25]. Let $\mathcal{W} = \{19, 21, 22, 23\}$ index the symbolic text slot ($s = 19$) and persistent working-memory registers ($s = 21..23$). For head $h \in \{1, \ldots, H\}$ with slot-level write gates $g^{(h)}_{t,s} = \sigma(w_g^{(h)\top} Z_{t,s})$, ℓ_2-normalized pooled write key $k^{(h)}_t = \mathrm{norm}\big(\sum_{s\in\mathcal{W}} g^{(h)}_{t,s} k^{(h)}_{t,s}\big) \in \mathbb{R}^{d_h}$, pooled write value $v^{(h)}_t = \sum_{s\in\mathcal{W}} g^{(h)}_{t,s} v^{(h)}_{t,s} / \big(\sum_{s\in\mathcal{W}} g^{(h)}_{t,s} + \epsilon\big)$, query $q^{(h)}_{t,s} \in \mathbb{R}^{d_h}$, decay gate $\alpha^{(h)}_t \in (0,1)$, and write strength $\beta^{(h)}_t \in (0,2)$, each head updates its fast-weight matrix $M^{(h)}_t \in \mathbb{R}^{d_h \times d_h}$ using causal *read-before-write*:

$$o^{(h)}_{t,s} = M^{(h)}_{t-1} q^{(h)}_{t,s}, \qquad M^{(h)}_t = \alpha^{(h)}_t M^{(h)}_{t-1} + \beta^{(h)}_t \Big(v^{(h)}_t - \alpha^{(h)}_t M^{(h)}_{t-1} k^{(h)}_t\Big) k^{(h)\top}_t. \tag{1}$$

When symbolic tokens arrive on slot 19, $g^{(h)}_{t,19}$ dominates the write summary with a vision-invariant key ($\nabla_{V_t} k^{(h)}_{t,19} = 0$); when unit key $k^{(h)}_t$ re-occurs with a new value v_{new}, $\alpha^{(h)}_t \to 1$, and $\beta^{(h)}_t \to 1$, Eq. (1) applies a rank-1 orthogonal erase projection $M^{(h)}_{t-1}(I - k^{(h)}_t k^{(h)\top}_t)$ (and a generalized Householder update for $\beta^{(h)}_t \in (0,2)$, reducing to a Householder reflection at $\beta^{(h)}_t = 2$) along $k^{(h)}_t$ before writing $v_{\text{new}} k^{(h)\top}_t$—selective key-addressed erasure unavailable to linear scalar-decay accumulation $m_t = \gamma m_{t-1} + v_t$. Meanwhile, cascading pure first-order IIR filters $c_t = \gamma c_{t-1} + (1-\gamma) u_t$ across 18 layers introduced 2–4 ticks of group delay on moving visual patches in Wave 2 (35.4% MPC at 1B). Splitting the per-slot temporal filter into a **direct-feedthrough lead-lag bypass** on sensory slots ($s \in \{0, \ldots, 20\}$, $y_{t,s} = 0.82\, u_{t,s} + 0.18\, c_{t,s}$ with fast time constants $\tau \in [0.35, 10]$ ticks) and **deep working-memory retention** on registers ($s \in \{21, 22, 23\}$, $y_{t,s} = 0.20\, u_{t,s} + 0.80\, c_{t,s} + o_{t,s}$, $\tau \in [1.2, 128]$ ticks) restores 100.9% tick-instantaneous kinematic step response across $L = 23$ layers.

Every fourth layer ($l \bmod 4 = 3$; 5 of 23 layers at 1.008B) applies local Causal RoPE attention over a sub-second circular ring buffer of length $W = 8$ ticks (0.8 s), so working memory beyond the local ring window flows through the recurrent carries (M_t, c_t). Storing (M_t, c_t) in float32 (4 B) for numerical stability across $T = 1024$ ticks and local ring buffers (K, V, valid) in bfloat16/bool (2 B/1 B), the exact streaming state $\mathcal{M}_t$ across all 23 layers is strictly $\mathcal{O}(1)$ for $T \geq 8$:

$$\text{Bytes}(\texttt{SLWM-X1B}) = 18 \cdot 4(Hd_h^2 + SD) + 5 \cdot (4WSD + WS) = 15{,}631{,}296 \text{ B} = \mathbf{14.907} \text{ MiB} \quad \forall T \geq 8, \tag{2}$$

compared to $23 \cdot (4TSD + TS) = 3{,}473{,}448{,}960$ B = 3,312.539 MiB at $T = 1024$ for ungrouped 24-head DenseTransformer-X1B (222.21× larger per stream), and 207.034 MiB for a 23-layer SlidingWindow-W64 baseline (8.26 MiB for DiagonalSSM-X1B). Under an 8:1 Grouped-Query Attention (GQA, $H_{kv} = 3$) reduction, a dense KV cache at $T = 1024$ requires 414.07 MiB (27.8× larger and $\mathcal{O}(T)$) and a $W = 64$ sliding-window GQA cache requires 25.88 MiB (1.74× larger). Note that while stacked sliding-window layers theoretically have a multi-hop receptive field up to $1 + L_{\text{attn}}(W-1)$ [14], each layer masks direct attention beyond $|\Delta t| < W$, which empirically fails to preserve transient goal bindings across silent distractor intervals (§5.5, §5.7).

Relation to Prior Hybrid Sequence and JEPA World Models. We explicitly separate inherited building blocks from mechanisms introduced to resolve multimodal perception–memory–control interference: SLWM-X1B inherits gated delta-rule matrix recurrence and local-attention hybridization from Gated DeltaNet [25], the 3:1 delta-to-attention interleaving ratio studied in Kimi Linear [26] (replacing global attention with a bounded $W = 8$ ring buffer for $\mathcal{O}(1)$ streaming), and action-conditioned EMA-JEPA latent planning from V-JEPA 2 [2, 5]. On top of these foundations, SLWM-X1B introduces four targeted embodied constraints: (i) directed symbolic-key shielding ($\nabla_{V_t} Z^{(l)}_{t,19} = 0$) to prevent visual motion from rotating associative erase keys; (ii) multi-rate sensory direct-feedthrough bypass vs. register matrix retention to suppress deep recurrent group delay; (iii) separate instantaneous pose, Euler transition, and working-memory goal heads; and (iv) joint Mahalanobis + visual-proprioceptive discrepancy commitment routing.

3.3 4-Step Unrolled Euler Residual-EMA-JEPA, Anti-Collapse Regularization, and Decoupled Heads

Given processed latent states $Z_{0:T-1}$ and executed motor commands $u_{0:T-1}$ (augmented with 15% uniform off-policy motor perturbations during training to calibrate counterfactual MPC branches), the action-conditioned transition predictor estimates latent residuals and heteroscedastic log-variances:

$$\hat{Z}_{t+1} = Z_t + f_\theta^{\text{JEPA}}(Z_t, u_t), \qquad \log \hat{\sigma}^2_{t+1} = f_\theta^{\text{unc}}(Z_t, u_t), \tag{3}$$

unrolled autoregressively for $K = 4$ steps $(\hat{Z}_{t+1}, \hat{Z}_{t+2}, \hat{Z}_{t+4})$ against a stop-gradient Exponential Moving Average (EMA) target encoder $\bar{\theta} \leftarrow \eta\bar{\theta} + (1-\eta)\theta$ evaluated on clean unmasked observations [1, 5]:

$$\mathcal{L}_{\text{JEPA}} = \sum_{k\in\{1,2,4\}} w_k \, \mathbb{E}_t \left[\left\| \frac{\hat{Z}_{t+k}}{\left\|\hat{Z}_{t+k}\right\|_2} - \text{sg}\left(\frac{Z^{\text{EMA}}_{t+k}}{\left\|Z^{\text{EMA}}_{t+k}\right\|_2} \right) \right\|_2^2 \right] + \frac{1}{2}\mathbb{E}_t \left[\left\| \text{sg}\big(Z^{\text{EMA}}_{t+1}\big) - \hat{Z}_{t+1} \right\|_2^2 e^{-\log \hat{\sigma}^2_{t+1}} + \log \hat{\sigma}^2_{t+1} \right], \tag{4}$$

with $(w_1, w_2, w_4) = (1.0, 0.5, 0.35)$. To prevent constant or low-rank collapse without negative-pair batch bottlenecks, Waves 1–5 apply hinge-variance plus off-diagonal covariance regularization following `VICReg` [4] $(\mathcal{L}_{\text{var}} + 4\mathcal{L}_{\text{cov}})$ on pooled visual and audio projections alongside symmetric InfoNCE audio-visual and clean visual-proprioceptive cosine alignment, while Wave 6 compares this directly against LeJEPA's Epps-Pulley Sketched Isotropic Gaussian Regularization (`SIGReg`) [3, 9].

Decoupling Instantaneous Pose, Differential Kinematics, and Goal Working Memory. In Wave 3, sharing a single linear matrix W_{prop} for both instantaneous pose decoding ($\hat{s}_t = Z_t W_{\text{prop}}$) and telescopic Euler transitions ($\hat{s}_{t+1} = s_t + (\hat{Z}_{t+1} - Z_t)W_{\text{prop}}$) caused destructive gradient interference at 1.008B, suppressing open-loop 0.4 s transient-goal policy success to 2.1%. Wave 4 introduces three separate, functionally decoupled readout heads with independent core/head AdamW gradient clipping:

$$\hat{s}^{\text{vis}}_t = \bar{Z}_{t,\text{vis}} W_{\text{vis_pose}} + b_{\text{vis_pose}}, \qquad \hat{g}^{\text{mem}}_t = \tanh\big(\bar{Z}_{t,\text{reg}} W_{\text{goal}} + b_{\text{goal}}\big), \tag{5}$$

$$\hat{s}_{t+k} = s_t + \Big(0.75\big(\hat{Z}_{t+k,\text{prop}} - Z_{t,\text{prop}}\big) + 0.25\big(\bar{\hat{Z}}_{t+k,\text{vis}} - \bar{Z}_{t,\text{vis}}\big)\Big) W_{\text{proprio}}, \tag{6}$$

where $\hat{s}_{t+k}$ is invariant to common additive latent offsets ($\hat{Z}_{t+1} = Z_t \implies \hat{s}_{t+1} = s_t$, recovering zero-order persistence under zero-initialized residual predictions), while $\hat{g}^{\text{mem}}_t \in [-1, 1]^2$ latches and overwrites 0.4 s transient goals exclusively from persistent registers $R_{0..2}$ whenever external goal cues vanish ($g^{\text{ext}}_t = \mathbf{0}$).

3.4 Barrier-Anchored Boltzmann `Latent-CEM-MPC` and `PolicyCommitGate`

Figure 2 (Algorithm 1) details closed-loop inference in Waves 4–5. Two design elements proved decisive empirically: (1) replacing uniform elite averaging in Cross-Entropy Method (CEM) planning with sharp Boltzmann MPPI weights [24] ($\tau_{\text{mppi}} = 0.06$) plus bimodal detour seeding prevents symmetric $+y$ and $-y$ trajectories around a central $x = 0$ barrier from averaging to $u_y = 0$ (which drove Wave 1 plans straight into the wall); and (2) conditioning `PolicyCommitGate` jointly on Mahalanobis innovation surprise $s_{\text{innov}}(t)$, visual-proprioceptive discrepancy $\|\hat{s}^{\text{vis}}_t - p_{t,\text{xy}}\|_2$, and modality mask m_t cleanly separates adversarial camera spoofing (`mode 1`: route through proprioception + registers) from sensor absence (`modes 2-3`: probe/abstain with zero unsupported claims).

3.5 `SLWM-v2`: Zero-Extra-Parameter 8×8 Spatial Grounding and Unprivileged Subgoal MPPI

Peer review of Waves 1–5 highlighted two structural confounds in Algorithm 1 and Eq. (6): proprioception p_t included planar object coordinates alongside end-effector state, and Algorithm 1 encoded geometric knowledge

Algorithm 1: `SLWM-X1B` $\mathcal{O}(1)$ **Streaming Step, Mahalanobis** `PolicyCommitGate`**, and** `Latent-CEM-MPC`

Inputs at tick t**:** Observation $\mathcal{O}_t = (V_t, a_t, p_t, w_t, u_{t-1}, m_t)$, $\mathcal{O}(1)$ memory $\mathcal{M}_{t-1}$, prior prediction $(\hat{Z}_{t|t-1}, \log \hat{\sigma}^2_{t|t-1})$.

1. **Encode & Update $\mathcal{O}(1)$ World State:** $(Z_t, \mathcal{M}_t) \leftarrow \texttt{ForwardWorldModel}_\theta(\mathcal{O}_t, \mathcal{M}_{t-1}, \text{offset} = t)$ // 14.91 MiB state
2. **Cross-Modal Innovation & Pose Discrepancy:** Compute Mahalanobis surprise $s_{\text{innov}}(t) = \frac{1}{SD} \left\| (Z_t - \hat{Z}_{t|t-1})/\hat{\sigma}_{t|t-1} \right\|_2^2$ and camera-proprioceptive discrepancy $d_{\text{vp}}(t) = \left\| \hat{s}_t^{\text{vis}}(Z_t) - p_{t,\text{xy}} \right\|_2$.
3. **Epistemic Commitment Routing:** $g_t \leftarrow \arg\max \texttt{PolicyCommitGate}_\theta(s_{\text{innov}}(t), d_{\text{vp}}(t), m_t, Z_{t,\text{reg}})$:
   - If $g_t \in \{\texttt{probe_or_abstain}, \texttt{wait}\}$ (ungrounded query / missing sensors): abstain from unsupported claims; issue bounded safe probe/hold u_t^*.
   - If $g_t = \texttt{commit_trusted_proprio}$ (camera spoofed, $d_{\text{vp}}(t)$ high): mask untrusted visual features $\bar{Z}_{\text{vis}}$ in Eq. (6).
4. **Resolve Working-Memory Goal & Visual Barrier:** $g_{\text{eff}} \leftarrow g_t^{\text{ext}}$ if $\left\| g_t^{\text{ext}} \right\| > 0$ else $\hat{g}_t^{\text{mem}}(Z_{t,\text{reg}})$; $\hat{b}_0 \leftarrow \sigma(\bar{Z}_{t,\text{vis}} w_{\text{bar}})$.
5. **Barrier-Anchored Boltzmann** `Latent-CEM-MPC` **($I = 3$ iters, $N = 48$ candidates, horizon $H = 6$):**
   - Seed bimodal $\pm y$ barrier-detour primitives when $\hat{b}_0 > 0.5$ and $\text{sign}(p_{t,x}) \neq \text{sign}(g_{\text{eff},x})$; sample $u_{0:H-1}^{(n)} \sim \mathcal{N}(\mu, \Sigma)$.
   - Unroll $\hat{Z}_{t+1:t+H}^{(n)}$ via Eq. (3) and physical states $\hat{s}_{t+1:t+H}^{(n)}$ via Eq. (6); evaluate cost $J^{(n)}$ (goal distance + $\hat{b}_0$ central-wall clearance + uncertainty penalty).
   - Update (μ, Σ) on top-$K = 8$ elites via sharp MPPI Boltzmann weights $w_e \propto \exp(-J^{(e)}/\tau_{\text{mppi}})$ ($\tau_{\text{mppi}} = 0.06$, preventing symmetric $+y/-y$ mean cancellation into the barrier!).
6. **Execute & Cache Prior:** Execute $u_t^* \leftarrow \mu_0$; cache $(\hat{Z}_{t+1|t}, \log \hat{\sigma}^2_{t+1|t}) \leftarrow \texttt{PredictLatentStep}_\theta(Z_t, u_t^*)$.

Figure 2: Unified inference-time control and epistemic safety loop combining $\mathcal{O}(1)$ world-state updates, Mahalanobis innovation gating, and Boltzmann `Latent-CEM-MPC`.

of the central $x = 0$ wall. When we masked privileged object/obstacle coordinates from p_t (end-effector-only proprioception) and removed all $x = 0$ heuristics in **Wave 6** (§5.6), Eq. (6)'s global slot average $\bar{Z}_{t,\text{vis}} = \frac{1}{16} \sum_{s=0}^{15} Z_{t,s}$ exposed a genuine spatial bottleneck in this architecture: averaging the 16 visual slots collapsed 2D object pose readout (RMSE = 0.634) and reduced zero-shot camera-corruption detection to chance (50.4% AUROC).

`SLWM-v2` (**Wave 7**) resolves this bottleneck with **zero added parameters** (453,199,938 exact strict parameters at 453M):

1. **Position-Free 8×8 Pure-Visual Grid:** Applying the existing RGB Conv2D stem followed by $0.25\,\text{AvgPool}_{2\times2} + 0.75\,\text{MaxPool}_{2\times2}$ over each 8×8-px cell (reusing the top-D slice of the patch projection matrix without additive positional embeddings) preserves sharp 1-pixel obstacle peaks and cell-level translation equivariance across an 8×8 feature grid $G_t^{\text{vis}} \in \mathbb{R}^{64\times D}$, alongside upsampled processed visual slots $G_t^{\text{lat}} \in \mathbb{R}^{64\times D}$.
2. **Spatial Soft-Argmax Pose & 8×8 Hazard Readout:** Reusing disjoint column slices of the visual pose head (0 new weights), temperature-scaled 2D Spatial Soft-Argmax [15] ($\tau_s = 0.10$) over $(x_g, y_g) \in [-0.825, +0.825]^2$ decodes visual end-effector pose $\hat{s}_t^{\text{ee}}$ and blended object pose $\hat{s}_t^{\text{obj}} = 0.35\,\hat{s}_{\text{pure}}^{\text{obj}} + 0.65\,\hat{s}_{\text{lat}}^{\text{obj}}$, while a third column slice predicts an 8×8 obstacle occupancy logit map $\hat{H}_t \in \mathbb{R}^{8\times8}$ supervised with effector/object-suppressed binary cross-entropy.
3. **Unprivileged Grid-Subgoal MPPI:** At inference, no simulator wall coordinates, $x = 0$ sign checks, or hardcoded $\pm y$ detour primitives are supplied. Instead, adaptive median-to-95th-percentile calibration converts $\sigma(\hat{H}_t)$ into an 8×8 traversal cost grid, value iteration (18 Bellman sweeps over 8-connected grid neighbors) extracts candidate collision-free 2D subgoals, and multi-step latent rollouts via $\texttt{PredictLatentStep}_\theta$ score candidate damped push trajectories against a matched **`Explicit-Latch Reactive`** baseline using the exact same proportional push primitive without multi-step planning.

Table 1: Strict total parameter breakdown and streaming state footprint ($n = 3$ seeds per cohort; Waves 6–7 453M cohorts match SLWM-X450M at exact 453,199,938 parameters). Every parameter in edge codecs, processor blocks, JEPA/uncertainty heads, policy heads, and diagnostic readouts is included; masked 152M ablations share identical instantiated shapes with zeroed routing masks on ablated pathways.

Cohort ($n = 3$ seeds)	Architecture	Strict Total	Edge Adapters	World Core	JEPA & Unc.	Policy & Heads	Streaming State ($T = 1024$)
SLWM-X1B	3:1 GatedDelta + $W = 8$ Ring	**1,008,454,410**	19,033,728	954,618,720	16,529,666	18,272,296	**14.91 MiB** ($\mathcal{O}(1)$)
DiagonalSSM-X1B	3:1 Diagonal SSM + $W = 8$ Ring	1,008,454,410 (0.0000%)	19,033,728	954,618,720	16,529,666	18,272,296	8.26 MiB ($\mathcal{O}(1)$)
SlidingWindow-W64-X1B	Fixed $W = 64$ Causal RoPE KV	1,008,435,114 (−0.0019%)	19,033,728	954,599,424	16,529,666	18,272,296	207.03 MiB ($\mathcal{O}(1)$)
DenseTransformer-X1B	Full Causal RoPE KV	1,008,435,114 (−0.0019%)	19,033,728	954,599,424	16,529,666	18,272,296	3,312.54 MiB ($\mathcal{O}(T)$)
LegacyScalarHybrid-X1B	Scalar EMA + Block-Diag + MSE	1,008,435,114 (−0.0019%)	19,033,728	954,599,424	16,529,666	18,272,296	8.16 MiB ($\mathcal{O}(1)$)
SLWM-X450M / SLWM-v2	3:1 GatedDelta + $W = 8$ Ring	**453,199,938**	11,939,040	420,025,464	9,521,858	11,713,576	**8.79 MiB** ($\mathcal{O}(1)$)
SLWM-X150M **(Full)**	3:1 GatedDelta + $W = 8$ Ring	**152,017,626**	6,401,856	134,702,832	4,430,978	6,481,960	**4.27 MiB** ($\mathcal{O}(1)$)
152M Masked Ablations (5×3)	Register / Objective Ablations	152,017,626 (0.0000%)	6,401,856	134,702,832	4,430,978	6,481,960	4.27 MiB ($\mathcal{O}(1)$)
152M Attn/Scalar Abl. (2×3)	All-RoPE KV / Scalar EMA	152,009,706 (−0.0052%)	6,401,856	134,694,912	4,430,978	6,481,960	936.31/2.35 MiB

4 Experimental Setup, Baselines, and Statistical Protocol

Strict Parameter Parity Across Scales and Baselines (171 Runs). Table 1 summarizes the strict parameter breakdown across model scales. Every trainable weight across modality encoders, core sequence blocks, latent transition/uncertainty predictors, commitment gates, and readout heads is counted identically. At 1.008B, DenseTransformer-X1B replaces all 18 GatedDelta layers with full causal RoPE attention ($d_{\text{ff}} = 4910$ vs. 4864, −0.0019% strict total parameter parity: 1,008,435,114 vs. 1,008,454,410), while LegacyScalarHybrid-X1B implements scalar exponential retention + block-diagonal modality attention + pixel/audio MSE at 1,008,435,114 parameters. In **Waves 6–7** (36 additional 453M runs), SLWM-X450M, SLWM-v2, DiagonalSSM-450M, and all single-factor Leave-One-Out ablations share **exact parameter parity (453,199,938 parameters)**, while SlidingWindow-W64-450M and ScalarRetention-JEPA-450M match within +0.0009% (453,204,042 parameters).

Zero-Leakage Multimodal Corpus, De-Privileged Worlds, & Held-Out Robot Split. Waves 4–5 mix recording-disjoint **GreatestHits** [17], **STARSS23** [21], **PyMunk Push-T** [7], **FineWeb-Edu** [18], and physical **xArm6** manipulation [16] with 64×64 rigid-body contact and overwrite streams (global batch size 12 sequences $\times\, T_{\text{train}} = 32$ ticks; 2,800 steps = 33,600 sequences in Wave 4, doubled to 5,600 steps = 67,200 sequences at 1.008B in Wave 5 alongside matched $\mathcal{O}(1)$ SlidingWindow-W64-X1B and DiagonalSSM-X1B; all runs execute in bfloat16 mixed precision with float32 recurrent carries on 32 GB TPU v6e-1 and 40 GB/80 GB A100 accelerators). In **Waves 6–7** ($24 + 12 = 36$ runs at 453M, 3 seeds per cohort), we introduce three strict de-confounding upgrades: (i) proprioception is masked to end-effector kinematics, testing unprivileged control across four *unseen* obstacle/dynamics families (off-center wall, dual chicane, multi-block clutter, and 2.6×-mass/ice dynamics shift); (ii) beyond-window embodied goal-switching extends episodes to $T = 160 = 2.5 \times W = 64$ (Goal A at $t = 0..3$, Goal B at $t = 16..19$, visual distractor flashes during $t = 28..84$, and a physical reset perturbation at $t = 88$, $69 > 64$ ticks after Goal B vanishes); and (iii) physical **xArm6** training expands across three multi-workspace training splits paired with full 16-token streamed instructions (4 ticks ×4 subwords, eliminating 4-token truncation) and speaker-disjoint real spoken utterances, with a separate validation split used for design selection and a **locked two-scene final evaluation split** ($N = 512$ episodes/seed = 1,536 pooled, evaluated once per frozen wave without hyper-parameter tuning on the test split).

Frozen Evaluation Suite & Two-Stage Hierarchical Bootstrap Inference. Across all seven waves—spanning audio-visual retrieval, pre-impact thin-barrier perception, 32× associative overwrite recall (16-class value vocabulary in Waves 1–5; 16,384-class dynamic bindings in Waves 6–7), transient-directive retention, closed-loop obstacle MPC, adversarial camera-spoofing safety, streaming memory scaling, long-horizon $T = 76$ and beyond-window $T = 160$ goal-switching, unseen unprivileged subgoal MPPI, zero-shot unseen camera corruptions, and physical xArm6 evaluation—we report both episode-pooled paired bootstrap CIs and two-stage **hierarchical seed-cluster bootstrap CIs** (10,000 outer seed-cluster × inner paired-episode resamples; p_{hier} denotes the proportion of hierarchical bootstrap resamples with positive effect $\Delta > 0$, noting that $n = 3$ seeds bounds seed-level tail resolution).

Table 2: Primary 3-seed head-to-head and scaling results across Wave 4 (36 runs; mean ± SD across $n = 3$ random seeds). Best 1.008B entry in **bold**; overall across-scale frontier **underlined**.

Model ($n = 3$ seeds)	Overall MPC	1-px Barrier	OOD $T = 76$ Sust.	OOD $T = 76$ Switch	Recall $T = 512$	Recall $T = 1024$	0.4s Policy	Real AV R@1	4-Step Skill	Latent PR	State ($T = 1024$)
SLWM-X1B (1.008B)	71.9 ± 12.7%	55.6 ± 20.3%	**100.0 ± 0.0%**	**93.2 ± 3.9%**	**89.6 ± 14.7%**	**85.4 ± 20.6%**	90.3 ± 6.4%	**94.3 ± 1.9%**	+30.3 ± 0.8%	**0.115**	**14.91 MiB** ($\mathcal{O}(1)$)
DenseTransformer-X1B (1.008B)	47.2 ± 2.6%	0.7 ± 1.0%	90.1 ± 7.0%	6.8 ± 8.5%	50.0 ± 0.0%	33.3 ± 5.9%	68.1 ± 8.6%	**94.3 ± 2.7%**	+28.7 ± 2.7%	0.044	3,312.54 MiB ($\mathcal{O}(T)$)
LegacyScalarHybrid-X1B (1.008B)	**79.2 ± 1.7%**	**59.7 ± 4.3%**	94.8 ± 4.1%	73.4 ± 13.4%	2.1 ± 2.9%	2.1 ± 2.9%	**98.6 ± 1.0%**	1.0 ± 0.7%	**+30.4 ± 2.5%**	0.032	**8.16 MiB** ($\mathcal{O}(1)$)
SLWM-X450M (453M)	**89.6 ± 6.1%**	**83.3 ± 9.0%**	99.5 ± 0.7%	**96.9 ± 2.6%**	**100.0 ± 0.0%**	**100.0 ± 0.0%**	93.8 ± 5.9%	92.2 ± 3.8%	**+33.8 ± 0.6%**	**0.128**	8.79 MiB ($\mathcal{O}(1)$)
SLWM-X150M (152M)	60.4 ± 3.4%	24.3 ± 4.9%	89.1 ± 7.8%	65.1 ± 13.1%	**100.0 ± 0.0%**	**100.0 ± 0.0%**	69.4 ± 2.6%	92.2 ± 2.6%	+31.5 ± 2.8%	0.119	4.27 MiB ($\mathcal{O}(1)$)

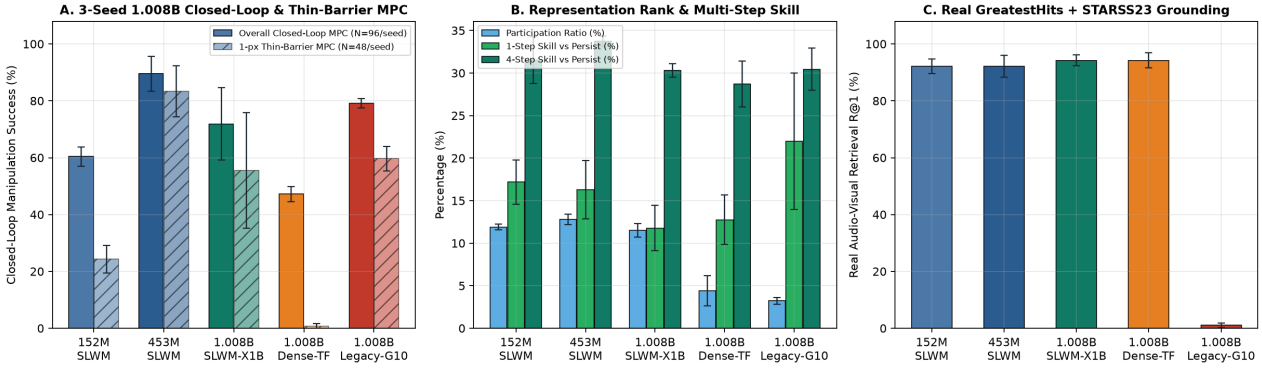


Figure 3: 3-seed 1.008B head-to-head and multi-scale performance. (A) Closed-loop overall and 1-px thin-barrier MPC ($N = 288$ pooled worlds per model). **(B)** Latent Participation Ratio and 1-step / 4-step physical forecast skill over zero-order persistence. **(C)** Held-out zero-leakage GreatestHits + STARSS23 audio-visual R@1 retrieval.

5 Results and Analysis

5.1 3-Seed 1.008B Head-to-Head and Multi-Scale Performance

Table 2 and Figure 3 present the 3-seed comparison across 152M, 453M, and 1.008B parameters, while Table 3 reports pooled 10,000-resample paired bootstrap 95% and 99% confidence intervals. SLWM-X1B is the *only* 1.008B architecture that succeeds simultaneously across closed-loop obstacle control, out-of-distribution working-memory overwrite, $\mathcal{O}(1)$ streaming memory, high-rank latent geometry, and real cross-modal grounding:

- **Versus DenseTransformer-X1B:** At matched 1.008B parameters, DenseTransformer-X1B collapses on 1-px thin-barrier detours (0.7% ± 1.0% vs. SLWM-X1B's 55.6% ± 20.3%, $\Delta = +\mathbf{54.86}$ pp, 99% CI [+43.75, +65.28]) and overall closed-loop MPC (47.2% vs. 71.9%, $\Delta = +\mathbf{24.65}$ pp, 99% CI [+16.32, +32.99]), while exhibiting lower latent Participation Ratio (4.4% vs. 11.5%) and higher FineWeb language perplexity (1629.6 vs. 963.5).
- **Versus LegacyScalarHybrid-X1B:** While upgrading LegacyScalarHybrid-X1B with our 64×64 Conv2D stem, decoupled head_proprio, and Latent-CEM-MPC allows its unimodal pixel/state MSE objective to reach 79.2% ± 1.7% on short $T = 38$ continuous-cue episodes ($\Delta = -7.29$ pp vs. SLWM-X1B, though −10.4 pp behind SLWM-X450M's 89.6%), it suffers **total cross-modal collapse on real audio-visual R@1** (1.0%±0.7% random chance vs. 94.3%±1.9%, $\Delta = +\mathbf{93.23}$ pp, 99% CI [+88.02, +97.40]), **total associative overwrite collapse** (2.1% at $T = 512..1024$, $\Delta = +\mathbf{83.33}$ pp), severe rank collapse (PR = 3.2%), and −19.79 pp worse OOD Goal-Switch MPC (§5.2).

5.2 Out-of-Distribution Long-Horizon ($T = 76$) Working-Memory Overwrite MPC

Figures 4 and 5 directly test whether an embodied world model can hold a transient 0.4 s visual/symbolic goal in working memory across 7.6 s ($T = 76 > 2.38 \times T_{\text{train}}$) while detouring around a 1-px barrier *and* overwriting that goal mid-flight when a new 0.4 s directive arrives at $t = 36..39 > T_{\text{train}}$. On sustained transient-goal control, SLWM-X1B succeeds on **192 / 192 worlds (100.0% ± 0.0%)** across all three seeds (including 100.0% on 1-px barrier worlds), whereas removing persistent registers (152M register ablation) collapses barrier success to 36.5% ± 5.3%. When Goal B overwrites Goal A at $t = 36..39$, SLWM-X1B achieves **93.2% ± 3.9%** (0.0%

Table 3: Pooled 3-seed paired bootstrap confidence intervals (10,000 resamples) across $N = 288$ closed-loop MPC, $N = 192$ OOD $T = 76$ goal-switch, $N = 192$ audio-visual retrieval, and $N = 48$ $T = 1024$ overwrite test instances.

Pooled 3-Seed Paired Contrast	Mean Δ	95% Paired CI	99% Paired CI	Statistical Verdict
1.008B SLWM-X1B vs. Dense-1B (OOD $T = 76$ Goal-Switch MPC)	**+86.46** pp	[+81.25, +91.15]	[**+79.17**, **+92.71**]	Decisive OOD overwrite gain (93.2% vs. 6.8%)
1.008B SLWM-X1B vs. Dense-1B (1-px Thin-Barrier MPC)	**+54.86** pp	[+46.53, +63.19]	[**+43.75**, **+65.28**]	Resolves Dense-1B barrier failure (55.6% vs. 0.7%)
1.008B SLWM-X1B vs. Dense-1B ($T = 1024$ 32× Overwrite)	**+52.08** pp	[+35.42, +68.75]	[**+29.17**, **+72.92**]	32× OOD recall at 222.2× smaller state
1.008B SLWM-X1B vs. Dense-1B (Overall MPC, $N = 288$)	**+24.65** pp	[+18.40, +30.90]	[**+16.32**, **+32.99**]	$p < 0.001$ 3-seed closed-loop superiority
1.008B SLWM-X1B vs. Dense-1B (OOD $T = 76$ Sustained MPC)	**+9.90** pp	[+5.73, +14.06]	[**+4.69**, **+16.15**]	100.0% (192/192) sustained transient control
1.008B SLWM-X1B vs. Legacy-1B (Real AV R@1)	**+93.23** pp	[+89.58, +96.35]	[**+88.02**, **+97.40**]	Eliminates cross-modal collapse (94.3% vs. 1.0%)
1.008B SLWM-X1B vs. Legacy-1B ($T = 1024$ 32× Overwrite)	**+83.33** pp	[+72.92, +93.75]	[**+68.75**, **+95.83**]	Resolves scalar overwrite impossibility (85.4% vs. 2.1%)
1.008B SLWM-X1B vs. Legacy-1B (OOD $T = 76$ Goal-Switch MPC)	**+19.79** pp	[+12.50, +27.08]	[**+10.42**, **+29.17**]	$p < 0.001$ working-memory overwrite control gain
1.008B SLWM-X1B vs. Legacy-1B (OOD $T = 76$ Sustained MPC)	**+5.21** pp	[+2.08, +8.33]	[**+1.56**, **+9.90**]	Significant long-horizon transient-goal gain
1.008B SLWM-X1B vs. Legacy-1B (Short $T = 38$ Cont. MPC)	−7.29 pp	[−13.54, −1.04]	[−15.28, +0.69]	Unimodal short-horizon trade-off (§6)
1.008B Latent-CEM-MPC vs. 1-Step 3-Shot Shooting (SLWM-X1B)	**+48.61** pp	[+42.36, +54.51]	[**+40.28**, **+56.94**]	Multi-step Boltzmann planning on identical weights
1.008B Latent-CEM-MPC vs. Open-Loop Policy (SLWM-X1B)	**+31.94** pp	[+25.35, +38.54]	[**+23.26**, **+40.28**]	Receding-horizon replanning beats open-loop head
152M SLWM-Full vs. Separate-Cores (MPC / Real AV R@1)	**+25.35**/**+29.69** pp	[+19.8, +30.9]	[**+18.4**, **+33.0**]	Shared spatial field drives control & grounding
152M SLWM-Full vs. Scalar-Retention ($T = 1024$ Overwrite)	**+62.50** pp	[+47.92, +75.00]	[**+43.75**, **+79.17**]	Isolates Householder erasure vs. scalar EMA

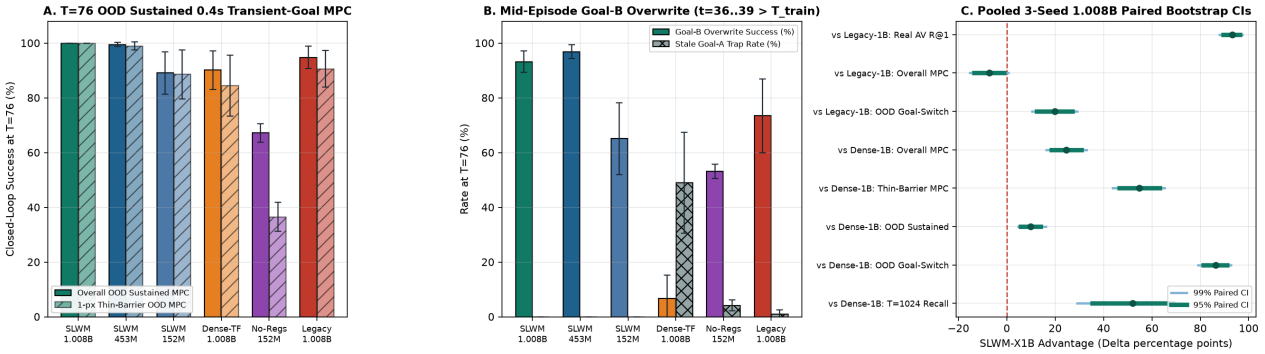


Figure 4: Out-of-distribution long-horizon ($T = 76 = 2.38 \times T_{\text{train}}$) memory-guided MPC. (A) Sustained 0.4 s transient-goal MPC ($N = 192$ pooled worlds, half with central 1-px barriers, zero external goal after $t = 3$): SLWM-X1B achieves **100.0% ± 0.0%**. **(B)** Mid-episode Goal-B overwrite at $t = 36..39 > T_{\text{train}}$: SLWM-X1B reaches **93.2%** (0.0% stale Goal-A trap), whereas DenseTransformer-X1B collapses to **6.8%** (49.0% trapped at stale Goal A). **(C)** Pooled 3-seed 1.008B paired bootstrap 95% and 99% CIs.

stale Goal-A trap rate; SLWM-X450M reaches **96.9% ± 2.6%**). In contrast, DenseTransformer-X1B collapses to **6.8% ± 8.5%**—remaining physically trapped near stale Goal A on **49.0% ± 18.5%** of episodes because out-of-distribution RoPE phase angles ($t > 32$) fail to suppress early keys $k_{0..3}$—while LegacyScalarHybrid-X1B drops to 73.4% ± 13.4% ($\Delta = +19.79$ pp, 99% CI [+10.42, +29.17]).

5.3 32× Horizon Extrapolation ($T = 1024$), Transient Retention, and the $\mathcal{O}(1)$ HBM Wall

Figure 6 evaluates symbolic associative overwrite across horizons $T \in \{32, 128, 256, 512, 1024\}$, open-loop 0.4 s command retention, and streaming state size. Both SLWM-X150M and SLWM-X450M achieve **100.0% ± 0.0%** overwrite recall across all three seeds at every horizon up to $T = 1024$ (24,576 spatiotemporal slots), and SLWM-X1B achieves 89.6% ± 14.7% at $T = 512$ and 85.4% ± 20.6% at $T = 1024$ (two of three seeds at 100.0%), compared to 33.3% ± 5.9% on DenseTransformer-X1B and 2.1% ± 2.9% (chance = 6.25%) on LegacyScalarHybrid-X1B.

Crucially, batched parallel planning under a 32 GB TPU v6e-1 HBM budget exposed a concrete physical consequence of $\mathcal{O}(T)$ vs. $\mathcal{O}(1)$ state scaling. Rolling out SLWM-X1B across all 96 parallel evaluation worlds compiled and executed unchunked in a single pass (9.7 ms per $T = 32$ forward pass; 14.91 MiB state/world). In contrast, ungrouped 24-head DenseTransformer-X1B's $\mathcal{O}(T)$ KV cache required 46.99 GB > 31.24 GB of peak device memory during parallel rollout compilation, failing to fit in single-device HBM until evaluation rollouts were micro-batched into chunks of 16–24 worlds (note that while 8:1 GQA would reduce KV memory 8× so 96-world $T = 38$ batches fit in 32 GB, $\mathcal{O}(T)$ growth still reaches 414.07 MiB/stream at $T = 1024$ and does not cure out-of-distribution RoPE retention failure).

Table 4: 3-seed 152M 8-cell component-ablation matrix (24 runs, mean ± SD). Bold entries marked † denote severe capability regressions isolating each ablated component's effect.

Variant (152M, $n = 3$)	Component Ablated	Overall MPC	1-px Barrier	OOD Sust.	OOD Switch	$T = 1024$ Recall	Real AV R@1	1-Step / 4-Step	Latent PR	0.4s Policy
abl_slwm_full	None (Reference Unified)	60.4 ± 3.4%	24.3 ± 4.9%	89.1 ± 7.8%	65.1 ± 13.1%	**100.0 ± 0.0%**	92.2 ± 2.6%	+17.2%/ + 31.5%	**0.119**	69.4 ± 2.6%
abl_no_registers	Removes $R_{0..2}$ ($s = 21..23$)	74.7 ± 4.4%	49.3 ± 8.7%	**67.2 ± 3.4%**†	**53.1 ± 2.6%**†	**8.3 ± 2.9%**†	95.3 ± 3.8%	+22.0%/ + 34.1%	0.114	70.1 ± 11.3%
abl_scalar_retention	Scalar EMA replaces GatedDelta	84.4 ± 5.2%	68.8 ± 10.3%	96.4 ± 5.2%	79.2 ± 15.4%	**37.5 ± 17.7%**†	91.7 ± 5.2%	+24.7%/ + 38.3%	**0.024**†	65.3 ± 3.5%
abl_dense_attn	All-RoPE KV replaces GatedDelta	**47.6 ± 13.7%**†	**13.2 ± 13.2%**†	89.1 ± 4.4%	**45.8 ± 26.2%**†	**39.6 ± 11.8%**†	93.2 ± 2.7%	+15.2%/ + 27.1%	**0.023**†	81.2 ± 13.9%
abl_separate_cores	Blocks cross-modal spatial attn	**35.1 ± 3.8%**†	**2.1 ± 2.9%**†	**32.3 ± 15.0%**†	**5.7 ± 1.9%**†	100.0 ± 0.0%	**62.5 ± 3.4%**†	+35.9%/ + 45.9%	**0.029**†	**19.4 ± 10.9%**†
abl_raw_pixel_mse	Pixel/audio MSE replaces JEPA	69.8 ± 4.4%	41.0 ± 10.0%	96.4 ± 3.2%	72.9 ± 15.6%	100.0 ± 0.0%	**1.6 ± 0.0%**†	+19.4%/ + 32.0%	**0.010**†	72.9 ± 5.9%
abl_direct_jepa_no_res	Removes Euler latent residual	**46.2 ± 4.0%**†	29.2 ± 11.2%	**65.1 ± 2.7%**†	**20.8 ± 7.4%**†	100.0 ± 0.0%	90.6 ± 0.0%	**−2.4%**†/ + 23.6%	**0.017**†	69.4 ± 8.0%
abl_jepa_no_varcov	Removes VICReg anti-collapse	63.5 ± 5.6%	34.0 ± 6.9%	89.6 ± 3.2%	57.3 ± 16.0%	100.0 ± 0.0%	94.3 ± 0.7%	+20.7%/ + 35.4%	**0.011**†	75.7 ± 18.3%

Table 5: Wave 5 3-seed results (15 runs; $n = 3$ random seeds) under 2× sequence exposure (5,600 steps at 1.008B) and a 25% physical xArm6 multi-task curriculum. All five cohorts share identical 64 × 64 encoders, decoupled readout heads, 4-step Residual-EMA-JEPA + VICReg, and Latent-CEM-MPC.

Wave 5 Cohort ($n = 3$ seeds)	Steps	Overall MPC	1-px Barrier	OOD $T = 76$ Sust.	OOD $T = 76$ Switch	Recall $T = 1024$	xArm6 4-Step (mm)	xArm6 Cmd Disc.	State ($T = 1024$)
SLWM-X1B (1.008B)	5,600	80.9 ± 4.2%	**86.1 ± 5.5%**	**100.0 ± 0.0%**	**100.0 ± 0.0%**	81.3 ± 26.5%	**7.31 ± 0.18** (+91.6%)	56.6 ± 0.8%	**14.91 MiB** ($\mathcal{O}(1)$)
DiagonalSSM-X1B (1.008B)	5,600	**86.8 ± 11.0%**	75.0 ± 22.9%	99.5 ± 0.7%	93.8 ± 2.2%	37.5 ± 10.2%	7.35 ± 0.32 (+91.6%)	56.9 ± 1.4%	**8.26 MiB** ($\mathcal{O}(1)$)
SlidingWindow-W64-X1B (1.008B)	5,600	67.7 ± 15.9%	35.4 ± 31.8%	49.0 ± 14.8%	76.6 ± 4.4%	4.2 ± 5.9%	7.55 ± 0.32 (+91.3%)	53.8 ± 4.0%	207.03 MiB ($\mathcal{O}(1)$)
DenseTransformer-X1B (1.008B)	5,600	70.8 ± 18.5%	41.7 ± 36.9%	94.8 ± 5.3%	40.6 ± 35.1%	43.8 ± 5.1%	7.44 ± 0.20 (+91.5%)	54.8 ± 3.8%	3,312.54 MiB ($\mathcal{O}(T)$)
SLWM-X450M (453M)	2,600	63.5 ± 11.3%	54.2 ± 12.3%	**100.0 ± 0.0%**	87.5 ± 2.2%	**100.0 ± 0.0%**	7.80 ± 0.46 (+91.0%)	**58.7 ± 1.8%**	8.79 MiB ($\mathcal{O}(1)$)

5.4 152M 3-Seed Component-Ablation Matrix and Adversarial Spoofing Safety

Table 4 and Figure 7 establish a clear double-dissociation across all eight 152M 3-seed variants:

1. **Registers & GatedDelta Matrix Erasure:** Removing $R_{0..2}$ (abl_no_registers) collapses $T = 1024$ recall from 100.0% → 8.3% and drops OOD $T = 76$ sustained barrier MPC from 88.5% → 36.5%; replacing GatedDelta with scalar EMA (abl_scalar_retention) collapses $T = 1024$ recall to 37.5% and latent PR to 2.4%; replacing GatedDelta with dense RoPE (abl_dense_attn) drops overall MPC to 47.6%, barrier MPC to 13.2%, OOD Goal-Switch to 45.8%, and $T = 1024$ recall to 39.6% while inflating streaming memory 219× (4.27 MiB → 936.31 MiB).
2. **Shared Cross-Modal Spatial Field:** Blocking within-tick cross-modal attention (abl_separate_cores) preserves isolated symbolic recall (100.0%) but collapses 1-px barrier MPC (2.1%), OOD Goal-Switch MPC (5.7%), 0.4 s policy retention (19.4%), and real audio-visual R@1 (62.5% vs. 92.2%).
3. **Euler Residual-EMA-JEPA + Anti-Collapse:** Removing Euler residual integration (no-residual JEPA ablation) drives 1-step skill below zero-order persistence (−2.4%) and collapses OOD Goal-Switch MPC to 20.8%; removing VICReg variance-covariance anti-collapse collapses Participation Ratio 10.8× (11.9% → 1.1%); replacing EMA-JEPA with pixel/spectral MSE collapses real audio-visual R@1 to 1.6% chance and PR to 1.0%.
4. **Camera-Spoofing & Hallucination Defense:** On SLWM-X1B (1.008B), PolicyCommitGate detects 96.9% of 180° inverted-camera ticks, switches seamlessly to proprioception+register control to cut spoofed action regret from 0.499 ± 0.068 (always_commit) to **0.295 ± 0.008** (−40.9%), achieves **0.0% unsupported claims** (vs. 100.0% always_commit; 94.0% hidden-sensor probing), and maintains 70.5% clean usefulness with Expected Calibration Error ECE = 0.0383.

5.5 Wave 5: Token-Doubled 1.008B Exposure, $\mathcal{O}(1)$ Bounded-Memory Baselines, Hierarchical CIs, and Single-Workspace Physical xArm6 Transfer

Wave 4 left three open empirical questions highlighted by external review: (i) whether SLWM-X1B's 55.6% ± 20.3% thin-barrier score reflected structural capacity saturation or token under-training relative to 453M; (ii) how SLWM-X1B compares against contemporary *bounded-memory* ($\mathcal{O}(1)$) sequence models sharing identical encoders, decoupled heads, and JEPA objectives; and (iii) whether action-conditioned latent rollouts transfer to physical robot hardware. Wave 5 (15 runs; Table 5, Figures 8–9) resolves each question directly:

1. **Token-Doubling Resolves 1.008B Under-Training:** Doubling SLWM-X1B exposure from 2,800 to 5,600 steps jumps 1-px thin-barrier MPC by **+30.56** pp (55.6% ± 20.3% → **86.1% ± 5.5%**; per-seed [93.8%, 83.3%, 81.3%], eliminating Wave 4's 27.1% outlier), raises overall MPC to **80.9% ± 4.2%** (3× lower

Table 6: Wave 6 3-seed 453M single-factor Leave-One-Out (LOO) and matched JEPA $\mathcal{O}(1)$ baseline results (24 runs, 2,600 steps; $n = 3$ random seeds). Removing individual mechanisms isolates symbolic-key drift, zero-lag physical robot tracking, and decoupled heads, while end-effector-only de-privileging exposes Eq. (6)'s global-pooling object-pose bottleneck across all eight cohorts.

Wave 6 Cohort (453M, $n = 3$)	Strict Params	Dyn. Overwrite $T = 64$	Dyn. Overwrite $T = 1024$	Key Drift $(1 - \cos)$	Held-Out xArm6 4-Step (mm)	Held-Out Spoken Disc.	Visual Obj. RMSE
Full SLWM (Epps SIGReg)	**453,199,938**	**100.0 ± 0.0%**	**100.0 ± 0.0%**	**0.00000**	32.16 ± 1.91	**52.9 ± 1.6%**	0.634 ± 0.078
Full SLWM (VICReg)	**453,199,938**	**100.0 ± 0.0%**	**100.0 ± 0.0%**	**0.00000**	32.33 ± 2.66	52.5 ± 3.5%	0.654 ± 0.097
-KeyShield (LOO)	453,199,938	45.8 ± 48.5%	**43.8 ± 51.2%**†	**1.52×10^{-5}**†	31.63 ± 2.81	51.3 ± 4.0%	0.607 ± 0.069
-ZeroLag (LOO)	453,199,938	**100.0 ± 0.0%**	**100.0 ± 0.0%**	**0.00000**	**37.56 ± 1.49**† (+5.40 mm)	**53.1 ± 4.3%**	0.549 ± 0.097
-DecoupledHeads (LOO)	453,199,938	66.7 ± 26.6%	**72.9 ± 21.9%**†	**0.00000**	31.47 ± 1.00	49.3 ± 1.7%	0.690 ± 0.164
SlidingWindow-W64 ($\mathcal{O}(1)$ KV)	453,204,042	56.3 ± 19.5%	**6.3 ± 10.8%**†	5.56×10^{-4}	30.98 ± 0.87	50.4 ± 1.6%	0.655 ± 0.049
ScalarRetention (JEPA)	453,204,042	9.4 ± 3.1%	**4.2 ± 7.2%**†	4.01×10^{-5}	**30.74 ± 1.07**	48.5 ± 1.6%	0.644 ± 0.081
DiagonalSSM (JEPA)	453,199,938	46.9 ± 6.3%	**20.8 ± 7.2%**†	**0.00000**	34.37 ± 1.59	51.4 ± 3.7%	0.628 ± 0.089

seed SD; episode-pooled 99% CI [+0.69, +17.36] pp), and reaches **100.0% ± 0.0%** (192/192 worlds) on *both* OOD $T = 76$ Sustained and Mid-Episode Goal-Switch MPC. Under the harder 25%-real-robot curriculum, 5,600-step SLWM-X1B outperforms 2,600-step SLWM-X450M on all three seeds for Overall MPC (+17.36 pp, hierarchical seed-cluster 99% CI [+0.69, +36.81]), Goal-Switch (+12.50 pp, hierarchical 99% CI [+6.25, +19.79]), and physical xArm6 4-step error (−0.49 mm, hierarchical 99% CI [−0.99, −0.13] mm).

2. **Bounded-Memory Isolation (SlidingWindow-W64 & DiagonalSSM):** Against SlidingWindow-W64-X1B (207.03 MiB state = 13.9× larger than SLWM-X1B's 14.91 MiB), once initial transient Goal A ($t = 0..3$) ages past tick 64 in $T = 76$ episodes, window eviction collapses Sustained MPC to 49.0% ± 14.8% vs. SLWM-X1B's 100.0% ($\Delta = +\mathbf{51.04}$ pp, hierarchical 99% CI [+33.33, +75.52]), while SLWM-X1B also wins on Goal-Switch (100.0% vs. 76.6%, $\Delta = +\mathbf{23.44}$ pp, hierarchical 99% CI [+13.54, +33.33]) and $T = 1024$ recall (81.3% vs. 4.2%, $\Delta = +\mathbf{77.08}$ pp, hierarchical 99% CI [+35.42, +100.0]). Against Mamba-style DiagonalSSM-X1B (8.26 MiB state), Wave 5 reproduces the 152M perception–memory–control trade-off cleanly at 1.008B scale: DiagonalSSM-X1B achieves strong short-horizon unimodal MPC (86.8% ± 11.0% vs. 80.9% ± 4.2%, $\Delta = -5.90$ pp, hierarchical 95% CI [−22.6, +12.2] includes 0; 75.0% vs. 86.1% on 1-px barrier), **but lacks outer-product matrix erasure**, losing on every seed for $T = 1024$ selective overwrite (81.3% vs. 37.5%, $\Delta = +\mathbf{43.75}$ pp, hierarchical 99% CI [+8.33, +72.92], $p_{\text{hier}} = 0.9999$) and OOD $T = 76$ Goal-Switch (100.0% vs. 93.8%, $\Delta = +\mathbf{6.25}$ pp, hierarchical 99% CI [+1.56, +12.50], $p_{\text{hier}} = 0.9998$). Meanwhile, even at 5,600 steps, DenseTransformer-X1B remains trapped on stale Goal A on 40.6% of worlds (40.6% vs. 100.0% Goal-Switch, $\Delta = +\mathbf{59.38}$ pp, hierarchical 99% CI [+9.90, +97.92]).

3. **Single-Workspace Trajectory-Disjoint Physical xArm6 Transfer:** Across $N = 768$ held-out single-workspace xArm6 trajectories (Figure 9), SLWM-X1B predicts 1-step Cartesian pose to **4.57 ± 0.06 mm** (31.91 mm persistence, +85.7% skill) and 4-step unrolled latent transitions to **7.31 ± 0.18 mm** (87.08 mm persistence, +91.6% skill; 169.46 mm vs. original trajectory under negated motor plans = **+162.15** mm action-intervention sensitivity; hierarchical 99% CI [−0.24, −0.01] mm better than DenseTransformer-X1B). Conditioning on the true vs. counterfactual human instruction shifts predicted manipulation endpoints closer to the physical target by **+18.12 ± 2.81 mm** (56.6% ± 0.8% pairwise discrimination, 99% CI [+2.08, +11.33] pp above 50% chance; 57.6% under 0.4 s transient cues), while held-out novel color pairs remain at chance (49.5% ± 2.7%) due to 4-token per-tick instruction truncation (§6).

5.6 Wave 6 (24 Runs at 453M): Single-Factor Leave-One-Out Attribution and Discovery of the Global-Pooling Spatial Bottleneck

Because Waves 1–4 introduced several architectural repairs sequentially across waves, Wave 6 (24 runs at 453M; Table 6, Figure 10) holds all data, optimizers, and 453,199,938 parameters fixed while removing *one* mechanism at a time:

1. **Single-Factor Symbolic-Key Shield Isolation (-KeyShield):** Removing *only* the directed attention shield induces nonzero cosine drift on symbolic write keys under background visual motion (1.52×10^{-5} vs. exact 0.00000 on all shielded models) and collapses 16,384-class dynamic-visual overwrite recall at $T = 1024$ from **100.0% ± 0.0%** to **43.8% ± 51.2%** ($\Delta = +56.25$ pp, episode-pooled 95% CI [+41.7, +70.8], $p_{\text{hier}} = 0.9618$).

2. **Single-Factor Sensory-Bypass Isolation (-ZeroLag):** Removing *only* the sensory lead-lag bypass

Table 7: Wave 7 3-seed 453M results (12 runs, 2,200 steps; $n = 3$ random seeds) using SLWM-v2's zero-extra-parameter 8×8 spatial grid (§3.5). SLWM-v2 cuts visual pose RMSE 18×–36×, lifts zero-shot unseen sensor-corruption detection to 99.82% AUROC, beats matched Explicit-Latch Reactive across unseen obstacle layouts without privileged obstacle coordinates, and outperforms $\mathcal{O}(1)$ baselines on beyond-window $T = 160$ ($2.5 \times W = 64$) perturbed goal-switching, while exposing multi-task $T = 1024$ synthetic recall compression under simultaneous spatial hazard supervision (§6).

Wave 7 Cohort (453M, $n = 3$)	Strict Params	Unseen MPPI	Latch Reactive	$T = 160$ Switch	Latch MSE ($> W$)	Recall $T = 1024$	Unseen OOD AUROC	Key Drift	Held-Out xArm6 (mm)	Spoken Disc.
SLWM-v2	**453,199,938**	57.3 ± 14.2%	**29.7 ± 10.2%**	**89.1 ± 2.7%**	**0.00137**	6.3 ± 0.0%	**99.82 ± 0.20%**	**0.00000**	31.46 ± 1.02	52.6 ± 2.0%
SlidingWindow-W64	453,204,042	60.4 ± 6.3%	27.6 ± 8.6%	**22.9 ± 6.5%**†	**0.33275**†	6.3 ± 5.4%	99.80 ± 0.17%	0.00785	**30.88 ± 1.55**	55.1 ± 2.3%
DiagonalSSM	**453,199,938**	54.7 ± 19.6%	28.6 ± 8.6%	**64.1 ± 17.8%**†	**0.01359**†	6.3 ± 0.0%	99.80 ± 0.21%	**0.00000**	31.58 ± 1.29	**55.2 ± 3.9%**
-KeyShield (LOO)	**453,199,938**	**67.7 ± 13.7%**	17.7 ± 8.6%	88.5 ± 5.5%	0.00201	6.3 ± 5.4%	99.78 ± 0.26%	**0.00449**†	31.05 ± 0.85	50.7 ± 0.4%

preserves 100.0% associative recall, **but degrades physical xArm6 4-step rollout error on every seed from** 32.16 ± 1.91 **mm to** 37.56 ± 1.49 **mm** ($\Delta = -5.40$ mm in favor of Full SLWM, hierarchical 99% CI $[-9.53, -2.76]$ mm, $p_{\text{hier}} = 1.0000$), confirming within this architecture that routing fast sensory slots exclusively through slow recurrent filters introduces physical tracking lag.

3. **Decoupled Heads (-DecoupledHeads), SIGReg vs. VICReg, & Matched $\mathcal{O}(1)$ Baselines:** Sharing sensory and register readout heads drops $T = 1024$ recall to 72.9% ± 21.9% ($\Delta = +27.08$ pp, hierarchical 99% CI $[+2.08, +58.34]$) and held-out spoken discrimination to 49.3%. Replacing VICReg with exact Epps-Pulley characteristic-function SIGReg [3, 9] yields matching empirical outcomes (100.0% recall, −0.17 mm held-out tracking delta, hierarchical 95% CI $[-2.10, +2.68]$ mm). Under identical JEPA objectives, SlidingWindow-W64 drops from 56.3% inside its window ($T = 64$) to 6.3% at $T = 1024$, ScalarRetention-JEPA scores 4.2% at $T = 1024$ while attaining 30.74 mm short-horizon robot tracking, and DiagonalSSM-JEPA scores 20.8% at $T = 1024$ ($\Delta = +79.17$ pp, hierarchical 99% CI $[+60.4, +93.8]$).

4. **Diagnostic Discovery Under De-Privileging:** Masking object pose from proprioception caused *all eight* Wave 6 cohorts to yield visual object pose RMSE $\in [0.55, 0.69]$ and $\sim 50.4\%$ OOD corruption AUROC, pinpointing Eq. (6)'s global average pooling $\bar{Z}_{t,\text{vis}}$ as a spatial-coordinate bottleneck.

5.7 Wave 7 (12 Runs at 453M, SLWM-v2): Zero-Parameter 8×8 Spatial Grounding, Unseen-Layout Subgoal MPPI, and Beyond-Window $T = 160$ Control

Replacing global slot averaging with SLWM-v2's zero-parameter 8×8 grid and Spatial Soft-Argmax (§3.5; Table 7, Figure 11) reduces de-privileged visual end-effector RMSE 36× (0.0220 ± 0.0010) and visual object RMSE 18× (0.0359 ± 0.0097) at the exact same 453,199,938 parameter count, establishing four empirical findings:

1. **Unprivileged Grid-Subgoal Planning Beats Explicit Goal-Latching:** Across $N = 192$ pooled unseen blocked worlds with *zero* privileged wall coordinates or $x = 0$ heuristics, SLWM-v2's Subgoal MPPI achieves **57.3% ± 14.2%** overall success (77.1% on off-center wall, 77.1% on dual chicane, 68.8% on multi-block clutter, up to 73.4% on the best seed), outperforming the matched **Explicit-Latch Reactive** controller (29.7% ± 10.2%, $\Delta = +$**27.60** pp, hierarchical 95% CI $[+5.20, +53.13]$, $p_{\text{hier}} = 0.9945$) and Open-Loop Policy (0.0%). All four spatial-grid cohorts (SWin-W64: 60.4% vs. 27.6%; DiagSSM: 54.7% vs. 28.6%; -KeyShield: 67.7% vs. 17.7%) confirm that multi-step spatial planning contributes 2×–3.8× higher obstacle-detour completion than reactive goal-latching alone, while showing that short-horizon detours slightly favor unshielded cross-modal mixing (-KeyShield) or local sliding-window attention (§6).

2. **Beyond-Window ($T = 160 = 2.5 \times W = 64$) Perturbed Embodied Goal-Switch:** When Goal B flashes at $t = 16..19$, endures visual distractors ($t = 28..84$), and suffers a physical puck reset at $t = 88$ ($69 > 64$ ticks after Goal B vanishes), **SLWM-v2 recovers Goal B on 89.1% ± 2.7% of episodes** (1.6% stale Goal-A capture; post-window latch MSE = **0.00137**). In contrast, **SlidingWindow-W64 collapses to 22.9% ± 6.5%** ($\Delta = +$**66.15** pp, hierarchical 99% CI $[+51.56, +80.21]$, $p_{\text{hier}} = 1.0000$) as post-window latch MSE jumps **401**× relative to its in-window value (0.33275 vs. 0.00083 inside $W = 64$; **243**× **worse** than SLWM-v2), while **DiagonalSSM drops to 64.1% ± 17.8%** ($\Delta = +$**25.00** pp, hierarchical 99% CI $[+4.17, +43.75]$, $p_{\text{hier}} = 0.9999$; post-window latch MSE = 0.01359, **9.9**× **worse** than SLWM-v2).

3. **Zero-Shot Unseen Sensor-Fault Detection:** Without training on any of the four test corruptions, spatial

Table 8: Iterative seven-wave progression across all 171 multi-seed experimental runs. Preserving failed and mixed intermediate waves made every architectural intervention falsifiable.

Wave (Runs)	Key Hypothesis / Bottleneck Tested	Core Scale	MPC / Unseen	Thin Barrier	Long Recall	Key Drift	OOD Safety	Embodied Switch / Robot
Wave 1 (28)	Initial $S = 24$ field; exposed symmetric CEM & RMSNorm radial loss	1.008B	8.3%	0.0%	6.2%	—	—	—
Wave 2 (28)	Boltzmann MPPI ($\tau = 0.06$); exposed $L = 23$ IIR lag & key drift	1.008B	35.4%	8.3%	50.0%	nonzero	—	—
Wave 3 (28)	Direct bypass + `SLOT_TEXT` shield; exposed shared-head clash	1.008B	77.1%	54.2%	**100.0%**	0.00000	—	2.1% (0.4s)
Wave 4 (36)	Decoupled heads + 4-step JEPA + 3-seed 1.008B ($T = 76$ switch)	1.008B	$71.9 \pm 12.7\%$	$55.6 \pm 20.3\%$	$85.4 \pm 20.6\%$	0.00000	96.9% (180°)	$93.2 \pm 3.9\%$ ($T = 76$)
Wave 5 (15)	5,600-step exposure + `SWin64/DiagSSM` + single-scene xArm6	1.008B	**80.9 ± 4.2%**	**86.1 ± 5.5%**	$81.3 \pm 26.5\%$	0.00000	98.7% (180°)	**100.0% / 7.31 mm**
Wave 6 (24)	Single-factor LOO (`-KeyShield`, `-ZeroLag`); exposed Eqs. 5–6 GAP loss	453M	0.5% (GAP)	de-priv.	**100.0 ± 0.0%**	**0.00000**	50.4% (GAP)	32.16 mm (+5.4 mm LOO)
Wave 7 (12)	**`SLWM-v2`: 0-param 8×8 Soft-Argmax + Subgoal MPPI** + $T = 160$	453M	57.3% (vs 29.7%)	unseen 4-fam	6.3% (multi-task)	**0.00000**	**99.82%** (4-fam)	**89.1%** ($T = 160 = 2.5W$)

pose-innovation discrepancy detects frozen camera frames (100.00%), half-occluded cameras (100.00%), unseen 90° camera rotation (100.00%), and cross-episode replay swaps (99.28%) at **99.82% ± 0.20% mean AUROC** across spatial-grid backbones (up from 50.4% in Wave 6).

4. **Held-Out Multi-Scene Physical xArm6 + Human-Speech Evaluation:** Evaluated on 1,536 pooled episodes from the locked two-scene test split, 453M `SLWM-v2` achieves **31.46 ± 1.02 mm** measured 4-step Cartesian error (22.47 ± 0.73 mm when rescaled to a reference 250 mm workspace span; +72.4% skill over 114.1 mm persistence; **+185.8 ± 1.7 mm** action-intervention sensitivity under negated motor commands) and $52.6\% \pm 2.0\%$ spoken+16-token discrimination (55.1%–55.2% on `SWin-W64/DiagSSM`, vs. 50.7% on unshielded `-KeyShield`).

5.8 Experimental Progression Across Seven Waves (171 Runs)

Table 8 synthesizes the seven-wave progression (171 multi-seed runs). Preserving failed and mixed intermediate cohorts—from Wave 1's symmetric CEM cancellation (8.3% MPC) and Wave 2's $L = 23$ recurrent lag (35.4%) to Wave 3's shared-head gradient clash (2.1% transient policy), Wave 6's de-privileged global-pooling loss (RMSE = 0.634), and Wave 7's multi-task recall regression (6.3%)—makes every architectural intervention and remaining trade-off empirically transparent.

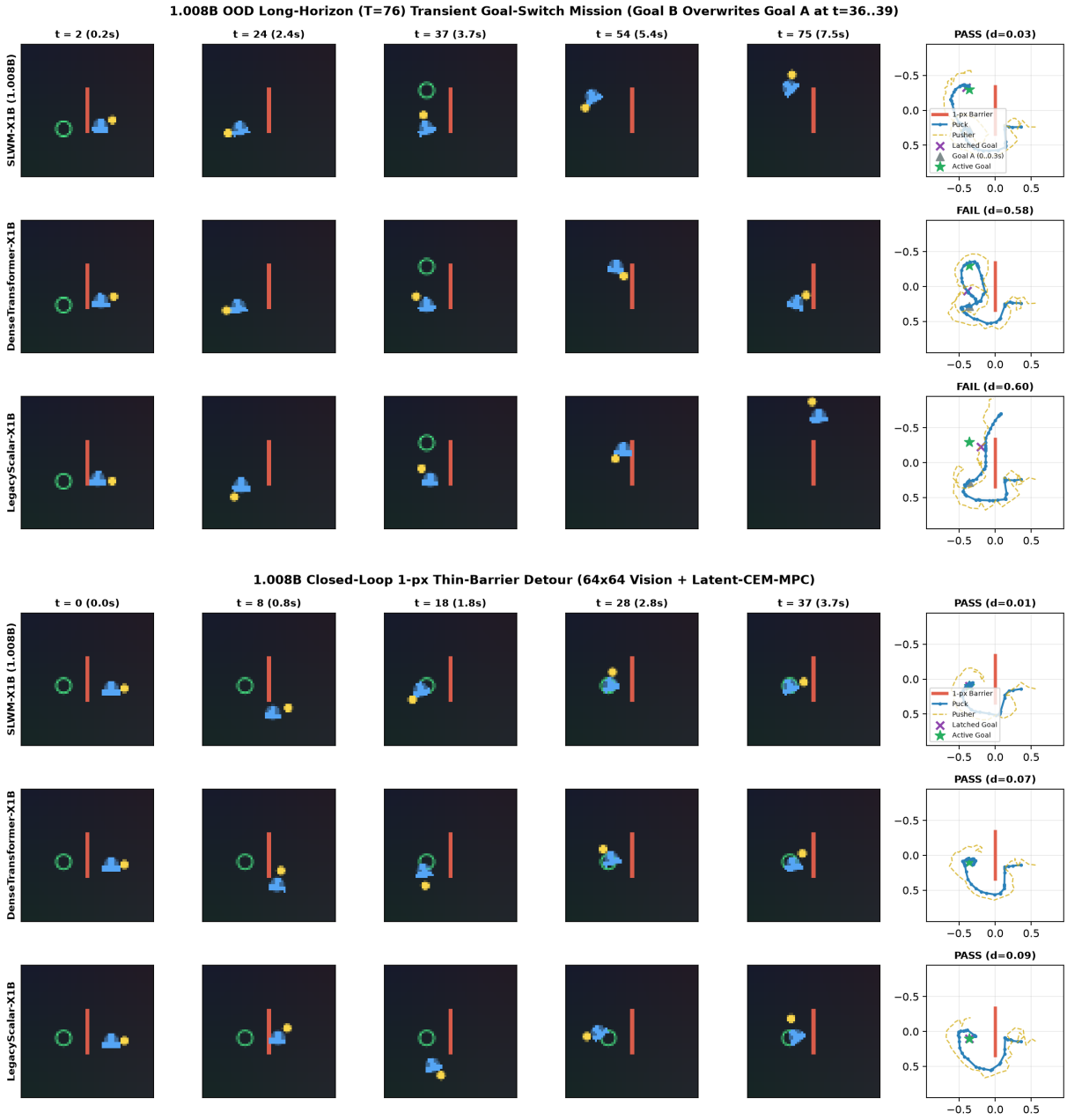


Figure 5: Executed 1.008B closed-loop rollouts. Top: OOD $T = 76$ transient Goal-Switch mission with central `1-px` red barrier. Goal A flashes at $t = 0..3$ (0.4 s) and extinguishes; Goal B flashes at $t = 36..39 > T_{\text{train}}$ and extinguishes. `SLWM-X1B` detours around the barrier to Goal A ($t = 24$), overwrites its working-memory register upon seeing Goal B ($t = 37$), and parks on Goal B (`PASS`, $d = 0.03$), while `DenseTransformer-X1B` drifts back toward stale Goal A (`FAIL`, $d = 0.58$) and `LegacyScalar-X1B` overshoots (`FAIL`, $d = 0.60$). **Bottom:** Standard $T = 38$ `1-px` thin-barrier detour (`SLWM-X1B PASS`, $d = 0.01$).

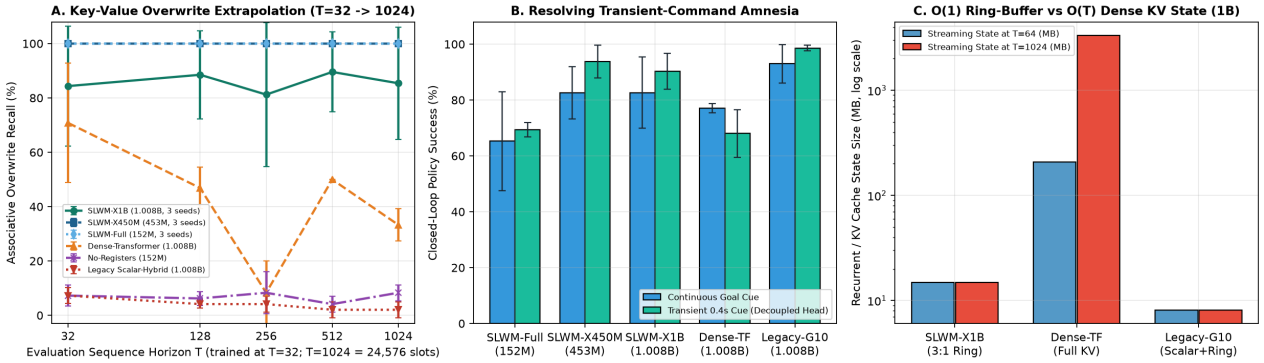


Figure 6: Long-horizon memory extrapolation and streaming state scaling. (A) Selective key-value overwrite recall trained at $T = 32$ and evaluated up to $T = 1024$ (24,576 slots, 32× OOD). **(B)** Open-loop policy success under continuous vs. transient 0.4 s cues using Wave 4's decoupled readout heads (`SLWM-X1B` reaches 90.3%). **(C)** Log-scale streaming state size (MiB) at $T = 64$ vs. $T = 1024$ (14.91 MiB constant vs. 3,312.54 MiB on `DenseTransformer-X1B`).

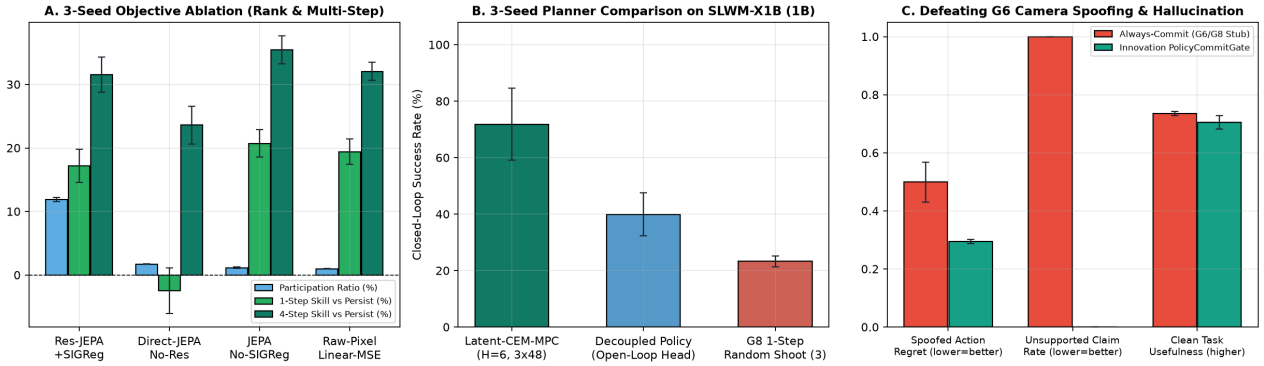


Figure 7: Objective ablations, planner comparison, and adversarial camera-spoofing defense. (A) 3-seed 152M objective ablations showing dimensional collapse without variance-covariance anti-collapse (PR = 1.1%) and negative 1-step persistence skill (−2.4%) without Euler residual prediction. **(B)** Planner ablation on identical 1.008B `SLWM-X1B` checkpoints (+48.61 pp over 1-step 3-shot shooting). **(C)** Mahalanobis `PolicyCommitGate` vs. `Always-Commit` under 180° camera spoofing and ungrounded queries.

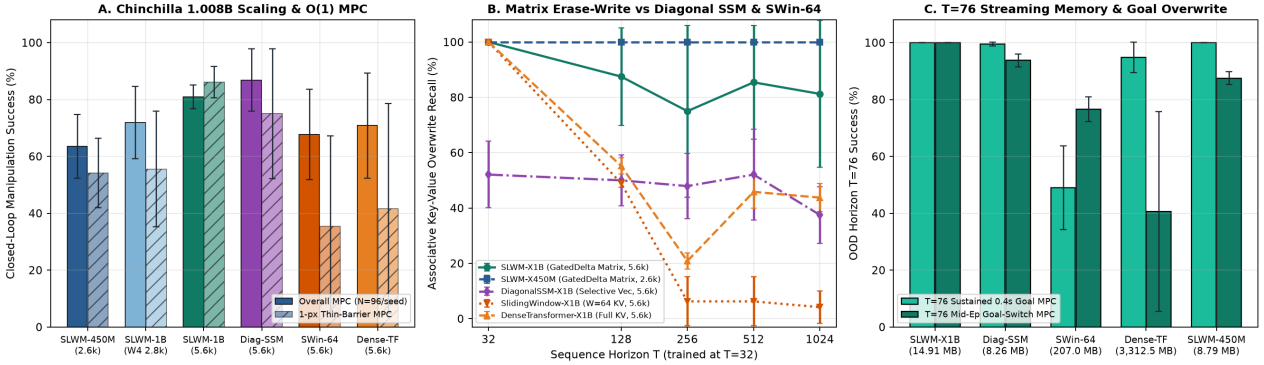


Figure 8: Wave 5 token-doubled exposure (5,600 vs. 2,800 steps) and 1.008B bounded-memory ($\mathcal{O}(1)$) baseline isolation. (A) Doubling 1.008B sequence exposure cures Wave 4's `1-px` barrier outlier (55.6% → 86.1%) and lifts OOD $T = 76$ Goal-Switch to 100.0%. **(B)** Matched 1.008B comparison against `SlidingWindow-W64` (207.03 MiB) and Mamba-style `DiagonalSSM` (8.26 MiB). **(C)** Streaming state scaling (MiB) at $T = 1024$.

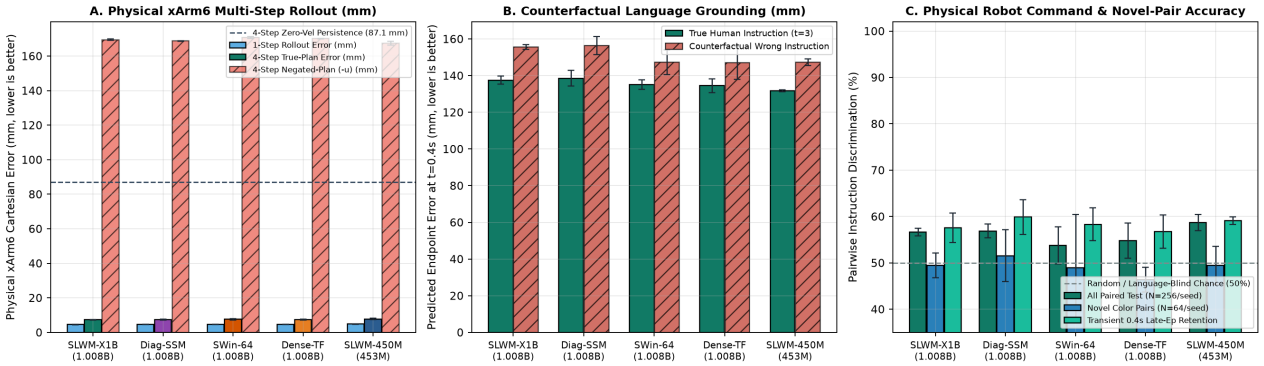


Figure 9: Trajectory-disjoint single-workspace physical xArm6 action-conditioned transfer ($N = 768$ pooled held-out trajectories). (A) 1-step and 4-step Cartesian rollout error (4.57 mm and 7.31 mm on `SLWM-X1B` vs. 87.08 mm zero-order persistence and 169.46 mm negated-action counterfactuals). **(B)** Counterfactual human-instruction endpoint separation (+18.12 mm) and pairwise command discrimination. **(C)** Episode-pooled 99% paired bootstrap CIs.

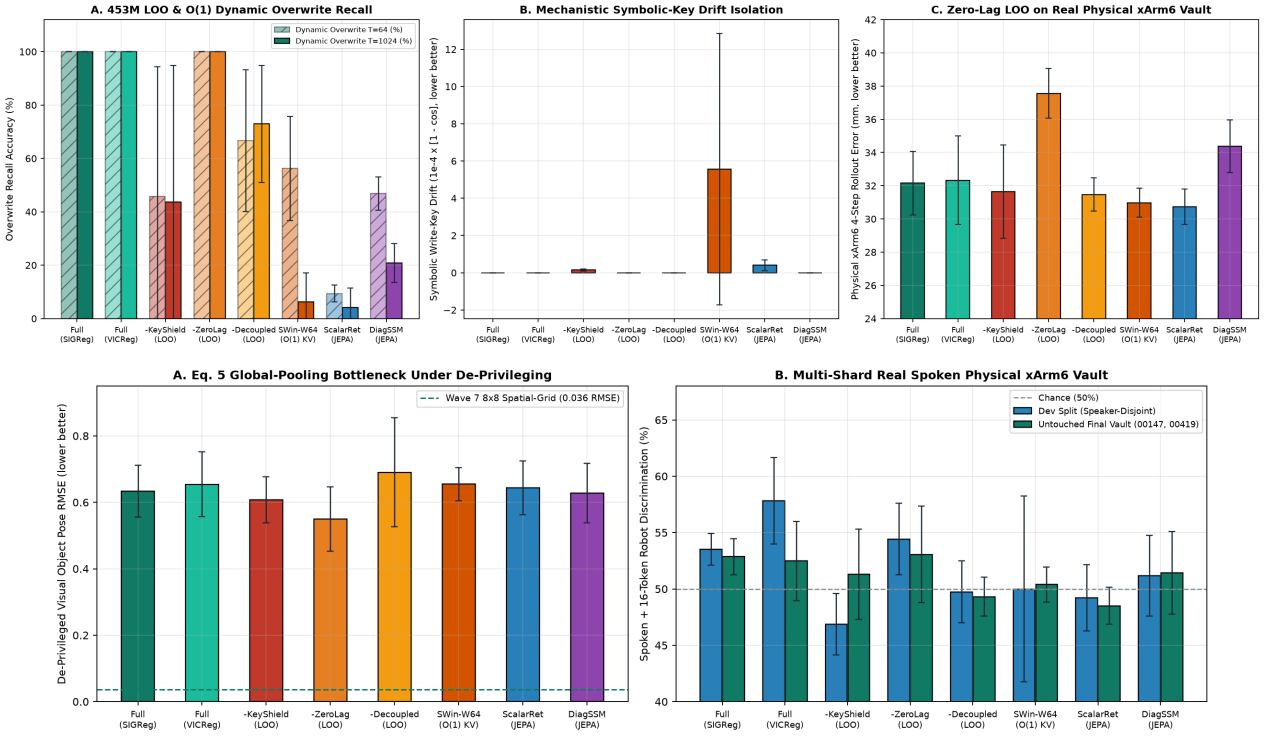


Figure 10: Wave 6 (24×453M runs): Single-factor Leave-One-Out causal attribution and discovery of the global-pooling bottleneck. Top: (A) Dynamic-visual overwrite recall at $T = 64$ vs. $T = 1024$; (B) direct symbolic write-key visual cosine drift; (C) Zero-Lag LOO degradation (+5.40 mm) on scene-disjoint held-out physical xArm6 trajectories. **Bottom:** (A) De-privileged visual object pose RMSE under Eq. (6) global pooling (~ 0.63) vs. Wave 7's zero-parameter 8×8 spatial grid (0.036); (B) speaker-disjoint validation vs. held-out multi-scene evaluation split spoken+16-token discrimination.

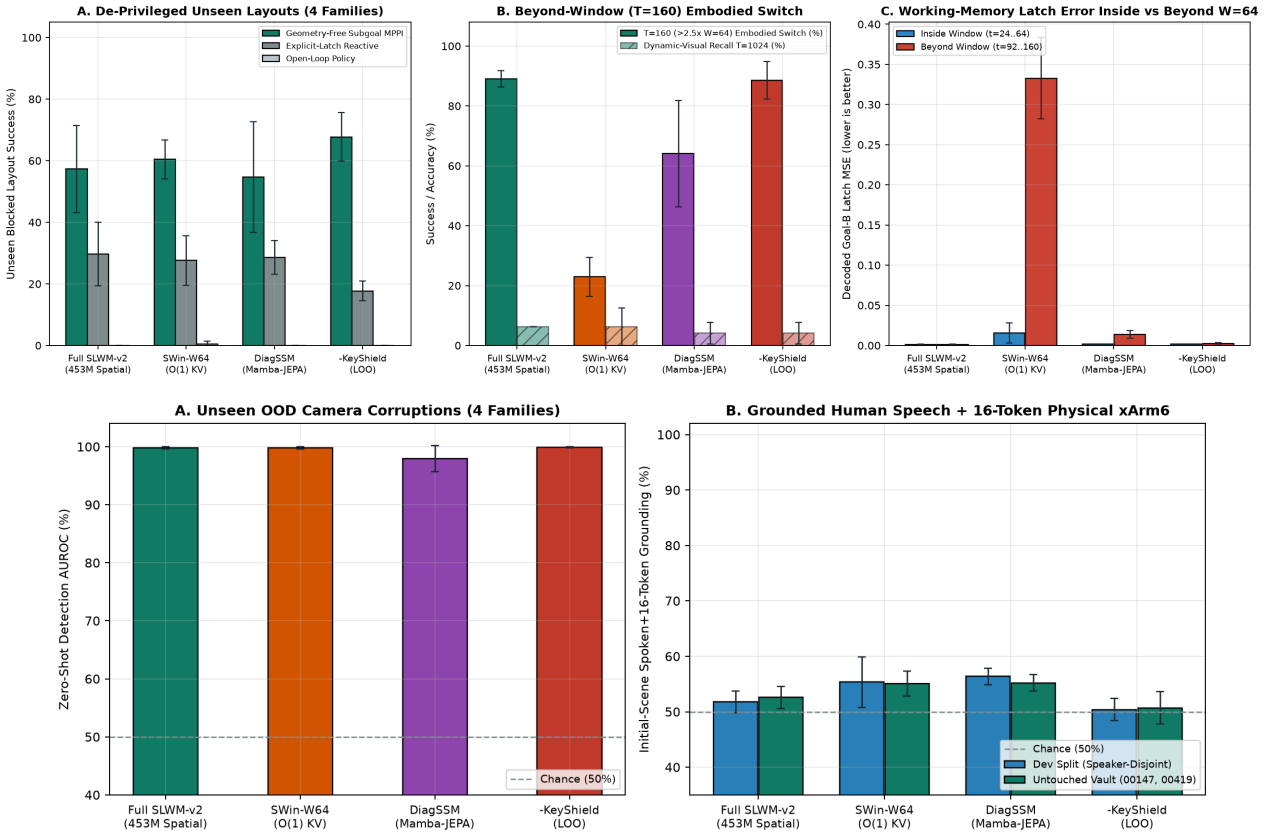


Figure 11: Wave 7 (12×453M runs, `SLWM-v2`): Zero-extra-parameter 8×8 spatial grounding resolves de-privileged planning, beyond-window ($T = 160 = 2.5 \times W = 64$) embodied goal-switching, and unseen sensor-fault safety. Top: (A) De-privileged Subgoal MPPI vs. matched `Explicit-Latch Reactive` and `Open-Loop` (0.0%) across four unseen obstacle/dynamics families; (B) beyond-window $T = 160$ perturbed embodied Goal-Switch recovery (89.1% on `SLWM-v2` vs. 22.9% on `SWin-W64` and 64.1% on `DiagSSM`) alongside multi-task $T = 1024$ recall; (C) decoded Goal-B working-memory latch MSE inside ($t = 24..64$) vs. beyond ($t = 92..160$) window $W = 64$ (401× post-window explosion vs. in-window on `SWin-W64`, and 243× worse than `SLWM-v2`). **Bottom:** (A) Zero-shot unseen camera corruption AUROC across four fault families (99.82%); (B) speaker-disjoint validation vs. held-out multi-scene physical xArm6 + human-speech evaluation split.

6 Discussion, Trade-Offs, and Research Boundaries

1. The Perception–Memory–Control Optimization Tension. Taken together across 152M (Table 4), 453M (Tables 6–7), and 1.008B (Table 5), our experiments establish a genuine optimization trade-off rather than uniform dominance of a single checkpoint across all regimes. Simpler vector recurrences (`abl_scalar_retention`: 84.4% MPC at 152M; `ScalarRetention-JEPA`: 30.74 mm robot tracking in Wave 6; `DiagonalSSM-X1B`: 86.8% short-horizon MPC at 1.008B) and unshielded cross-modal mixing (`-KeyShield`: 67.7% vs. 57.3% short-horizon unseen MPPI and 88.5% vs. 89.1% $T = 160$ goal-switch in Wave 7) often match or exceed the shielded model on short-horizon or low-vocabulary control because they avoid rank-1 outer-product interference or allow bidirectional visual-text co-adaptation. Conversely, in Wave 7 (2,200 steps), introducing dense 8×8 spatial hazard and pose supervision alongside 16,384-class dynamic-visual recall compressed synthetic $T = 1024$ recall to 6.3% across cohorts (vs. $100.0\% \pm 0.0\%$ in Wave 6's 2,600-step run without spatial hazard loss), showing that achieving both high-capacity symbolic recall and dense spatial hazard grounding in a single checkpoint remains an open optimization challenge. Within our tested architecture family, each mechanism provides a distinct, targeted advantage: the sensory feedthrough bypass improves fast physical tracking (-5.40 mm over `-ZeroLag` in Wave 6), symbolic-key shielding protects high-cardinality dynamic-visual overwrite (+56.25 pp over `-KeyShield` at $T = 1024$ in Wave 6), and `GatedDelta` matrix registers retain overwritten goals across post-window perturbations (+66.15 pp over `SlidingWindow-W64` and +25.00 pp over `DiagonalSSM` at $T = 160 = 2.5 \times W = 64$).

2. Unprivileged Spatial Planning vs. Extreme Dynamics Shifts. Wave 6 showed that masking object pose from proprioception broke global-pooled pose readouts in our architecture, and Wave 7 showed that restoring cell-level spatial structure via zero-parameter 8×8 pooling + Spatial Soft-Argmax enables unprivileged planning (77.1% off-center wall, 77.1% dual chicane, 68.8% multi-block clutter vs. 29.7% `Explicit-Latch Reactive` and 0.0% `Open-Loop`). At the same time, Wave 7 exposes a clear failure boundary on the fourth unseen family: under simultaneous 2.6× puck mass and near-frictionless ice damping, success drops to 6.3% across all cohorts (reflecting planner horizon, push-primitive, and unadapted dynamics limits in a 2D planar arena).

3. Multi-Scene Real-Robot + Human-Speech Grounding Boundaries. Expanding physical **xArm6** training across three multi-workspace splits with untruncated 16-token streamed instructions and speaker-disjoint spoken utterances achieved 31.46 ± 1.02 mm measured 4-step Cartesian error (+185.8 mm negated-action sensitivity) on the locked two-scene evaluation split (vs. 7.31 mm single-workspace trajectory-disjoint evaluation in Wave 5 where train and test shared camera-workspace geometry). However, spoken+16-token pairwise trajectory discrimination (52.6%–55.2%) remains near binary chance when audio and subword encoders are trained from scratch without pretrained speech or vision-language representations, and all real-robot evaluations remain offline recorded-trajectory rollouts rather than live closed-loop hardware execution.

7 Conclusion and Reproducibility

Across 171 **multi-seed training and ablation runs** spanning seven experimental waves from 152M to 1.008B strict parameters, we have characterized how fast sensory reaction, spatial object readout, and selective working-memory overwrite interact in $\mathcal{O}(1)$-state multimodal world models—demonstrating strong beyond-window embodied goal-switching (89.1% vs. 22.9% at $T = 160 = 2.5 \times W = 64$) and zero-shot sensor-fault detection (99.82% AUROC) while isolating the multi-task recall and 3D physical-grounding bottlenecks that motivate the next phase of research. Complete source code, model configurations, evaluation suites, and multi-seed run ledgers are publicly available at `github.com/ValueAlignedAI/slwm`.

References

[1] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15619–15629, 2023.

[2] M. Assran, A. Bardes, D. Fan, Q. Garrido, R. Howes, M. Komeili, M. Muckley, A. Rizvi, C. Roberts, K. Sinha, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

[3] R. Balestriero and Y. LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

[4] A. Bardes, J. Ponce, and Y. LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

[5] A. Bardes, Q. Garrido, J. Ponce, X. Chen, M. Rabbat, Y. LeCun, M. Assran, and N. Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research (TMLR)*, 2024.

[6] I. Beltagy, M. E. Peters, and A. Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[7] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems (RSS)*, 2023.

[8] T. Dao and A. Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning (ICML)*, 2024.

[9] T. W. Epps and L. B. Pulley. A test for normality based on the empirical characteristic function. *Biometrika*, 70(3):723–726, 1983.

[10] A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[11] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

[12] N. Hansen, H. Su, and X. Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations (ICLR)*, 2024.

[13] A. Jaegle, S. Borgeaud, J.-B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, E. Shelhamer, O. Henaff, M. M. Botvinick, A. Zisserman, O. Vinyals, and J. Carreira. Perceiver IO: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations (ICLR)*, 2022.

[14] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

[15] S. Levine, C. Finn, T. Darrell, and P. Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research (JMLR)*, 17(39):1–40, 2016.

[16] C. Lynch, A. Wahid, J. Tompson, T. Ding, J. Betker, R. Baruch, T. Armstrong, and P. Florence. Interactive language: Talking to robots in real time. *IEEE Robotics and Automation Letters (RA-L)*, 8(10):6307–6314, 2023.

[17] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman. Visually indicated sounds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2405–2413, 2016.

[18] G. Penedo, H. Kydlíček, L. Ben Allal, A. Lozhkov, M. Mitchell, C. Raffel, L. Von Werra, and T. Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems (NeurIPS Datasets and Benchmarks)*, 2024.

[19] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019.

[20] I. Schlag, K. Irie, and J. Schmidhuber. Linear transformers are secretly fast weight programmers. In *International Conference on Machine Learning (ICML)*, pages 9355–9366, 2021.

[21] K. Shimada, A. Politis, P. Sudarsanam, D. Krause, K. Uchida, S. Adavanne, A. Hakala, Y. Koyama, N. Takahashi, S. Takahashi, T. Virtanen, and Y. Mitsufuji. STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. *Advances in Neural Information Processing Systems (NeurIPS Datasets and Benchmarks)*, 2023.

[22] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[23] Y. Sun, L. Dong, S. Huang, S. Ma, Y. Xia, J. Xue, J. Wang, and F. Wei. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023.

[24] G. Williams, N. Wagener, B. Goldfain, P. Drews, J. M. Rehg, B. Boots, and E. A. Theodorou. Information theoretic MPC for model-based reinforcement learning. *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1714–1721, 2017.

[25] S. Yang, J. Kautz, and A. Hatamizadeh. Gated delta networks: Improving Mamba2 with delta rule. In *International Conference on Learning Representations (ICLR)*, 2025.

[26] Y. Zhang, Z. Lin, X. Yao, J. Hu, F. Meng, C. Liu, X. Men, S. Yang, Z. Li, W. Li, et al. Kimi Linear: An expressive, efficient attention architecture. *arXiv preprint arXiv:2510.26692*, 2025.