Towards *Unified World Models* for Visual Navigation via Memory-Augmented Planning and Foresight

Yifei Dong[1,*], Fengyi Wu[1,*], Guangyu Chen[1], Lingdong Kong[2], Qiyu Hu[1], Yuxuan Zhou[1], Xu Zhu[1], Jingdong Sun[3], Jun-Yan He[1], Qi Dai[4], Alexander G. Hauptmann[5], and Zhi-Qi Cheng[1,†]

[1]University of Washington [2]National University of Singapore [3]Apple
[4]Microsoft Research [5]Carnegie Mellon University

https://github.com/F1y1113/UniWM

Abstract. Enabling embodied agents to imagine future states is essential for robust and generalizable visual navigation. Yet, state-of-the-art systems typically rely on modular designs that decouple navigation planning from visual world modeling, which often induces state–action misalignment and weak adaptability in novel or dynamic scenarios. We propose UniWM, a unified, memory-augmented world model that integrates egocentric visual foresight and planning within a single multimodal autoregressive backbone. UniWM explicitly grounds action selection in visually imagined outcomes, tightly aligning prediction with control. Meanwhile, a hierarchical memory mechanism fuses short-term perceptual cues with longer-term trajectory context, supporting stable and coherent reasoning over extended horizons. Extensive experiments on four challenging benchmarks (Go Stanford, ReCon, SCAND, HuRoN) and the 1X Humanoid Dataset show that UniWM improves navigation success rates by up to 30%, substantially reduces trajectory errors against strong baselines, generalizes zero-shot to the unseen TartanDrive dataset, and scales naturally to high-dimensional humanoid navigation. These results position UniWM as a principled step toward unified, imagination-driven embodied navigation. All materials are committed to be open-sourced.

Keywords: Goal-Conditioned Navigation · World Modeling · Hierarchical Memory · Unified Representation Learning

1 Introduction

Visual navigation is a core capability for embodied AI and autonomous systems [16, 22, 46, 53], enabling agents to interpret egocentric observations and sequentially choose actions to reach goals in complex environments [32,36,67]. It

* Equal contribution, listed in random order. † Corresponding author.

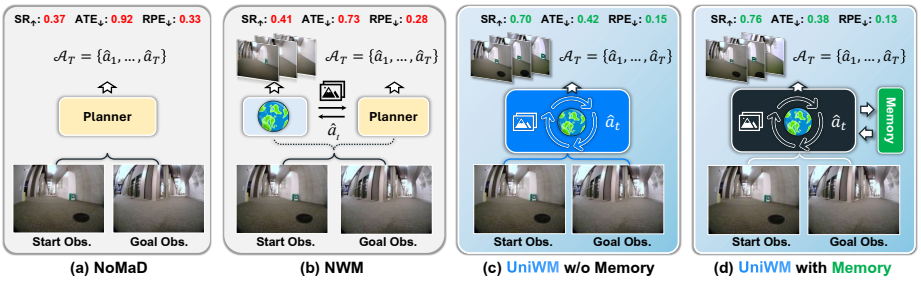


Fig. 1: Comparisons among goal-conditioned visual navigation approaches. **(a)** Navigation policy methods like NoMaD [53] directly predict action sequences A_T. **(b)** World model for navigation like NWM [9] uses a world model to visualize future observations, enhancing a separate navigation planner. **(c)** UniWM (no memory) unifies planning and visualization within one multimodal backbone, and actions are grounded in the imagined next observation while generating A_T autoregressively. **(d)** UniWM (with hierarchical memory) adds intra-step and cross-step memory banks, stabilizing longer-horizon rollouts and consistently yielding the highest SR and lowest errors (ATE/RPE). All panels use the same start/goal observations; headers report navigation performance SR↑, ATE↓, and RPE↓ on HuRoN [30] dataset.

supports real-world applications such as robotic delivery, autonomous driving, and assistive technologies [34,35,39,41,42,63], where robust perception, accurate planning, and the ability to *anticipate* how the environment will evolve under candidate actions are essential. Humans excel at this by mentally simulating future outcomes to plan efficiently in both familiar and novel settings [9, 11].

Despite rapid progress, current visual navigation systems are limited in fundamental ways (Fig. 1). **(a) Direct policy methods** (*e.g.*, GNM [50], VINT [51], NoMaD [53]) map observations to actions, but are tightly coupled to training distributions and often struggle to adapt in novel environments [52]. **(b) Modular pipelines** pair a planner with a separate world model: NavCoT [43] textualizes future observations and loses spatial fidelity, while NWM [9] uses diffusion-based rollouts and ranking. However, when prediction and control are learned in isolation and trajectory memory is absent, state-action misalignment arises and errors compound under partial observability and long horizons [19, 59]. **(c) Unified autoregressive frameworks** provide a more principled direction by interleaving *"imagining the next view"* with *"predicting the next action"*, grounding decisions in anticipated outcomes and reducing misalignment (Fig. 1c). Yet, unification alone does not prevent gradual drift in longer-horizon reasoning. **(d) Hierarchical memory** supplies the missing inductive bias: retaining both immediate perceptual cues and longer-range trajectory context promotes temporal coherence, yielding higher SR and lower errors in challenging settings (Fig. 1d).

In short, effective navigation requires not only the ability to *imagine while acting* but also to *remember over time*. The central challenge is therefore to unify

planning and imagination within a single backbone while injecting temporal structure to sustain stable long-horizon performance.

To address this challenge, we propose **UniWM**, a unified memory-augmented world model that integrates navigation planning and visual imagination within a single multimodal autoregressive backbone (Fig. 2; Sec. 3.1). During training, we interleave planner and world-model samples and jointly optimize bin-token classification for actions and reconstruction for images in a shared tokenization space spanning actions, text, pose, and vision; the framework scales naturally with parameter-efficient tuning such as LoRA (Fig. 2a; Sec. 3.2). At inference, UniWM alternates between predicting the next action and imagining the next egocentric view, explicitly grounding control in predicted visual outcomes and mitigating state–action misalignment (Fig. 2b; Sec. 3.3). We further introduce a two-level hierarchical memory that combines an intra-step cache with a cross-step trajectory store, augmenting attention via similarity gating and temporal decay to maintain coherent long-horizon rollouts and improve stability (Fig. 3). Together, UniWM unifies planning and imagination within one backbone and offers a practical recipe for memory-augmented foresight in visual navigation.

Empirically, UniWM improves Success Rate and reduces ATE/RPE on Go Stanford [29], ReCon [49], SCAND [36], and HuRoN [30] versus GNM [50], VINT [51], NoMaD [53], Anole-7B [17], and NWM [9]. For instance, Go Stanford SR rises from 0.45 to 0.75 (Table 1; Fig. 4). UniWM also generalizes zero-shot to unseen TartanDrive (SR 0.42; Table 7; Fig. 6) and scales to 25-DoF humanoid navigation on 1X Humanoid (Table 8; Fig. 7). Beyond navigation, UniWM improves one-step and rollout visualization (higher SSIM/PSNR, lower LPIPS/-DreamSim; Table 2). Ablations attribute gains to reconstruction for imagination fidelity (and downstream navigation), bin-token loss for action accuracy, and hierarchical memory for long-horizon stability, with additional effects from token budget, joint training, memory-layer choice, goal conditioning, and substep interleaving (Tables 3–6; Fig. 5).

In summary, this work provides the following key contributions:

- **Unified architecture.** We propose **UniWM**, to our knowledge, the **first** unified, memory-augmented world model that integrates visual navigation planning and imagination within a single multimodal autoregressive backbone, addressing the representational fragmentation of modular pipelines.
- **Unified training.** We introduce an end-to-end interleaved training strategy that unifies planner and world-model instances within one autoregressive backbone, jointly optimizing discretized action prediction and visual reconstruction to tightly align imagination with control.
- **Hierarchical memory.** We develop a hierarchical memory mechanism that fuses short-term perceptual cues with longer-term trajectory context via similarity-based retrieval and temporal weighting, enabling stable and coherent predictions over extended horizons.
- **Comprehensive validation.** Extensive experiments confirm consistent gains across benchmarks, stronger imagination fidelity, robust generalization to novel cases, and scalability to high-dimensional humanoid navigation.

2 Related Work

World models have emerged as a unifying paradigm for learning predictive representations of environment dynamics, supporting simulation, decision-making, *etc*. We summarize **two lines of research**: **(a)** advances in generic world-model architectures, and **(b)** world models for goal-conditioned visual navigation.

World Modeling has evolved from compact recurrent dynamics [25–28] through Transformer-based designs (*e.g.*, I-JEPA [4], V-JEPA [10], DINO-WM [7]) to large-scale generative systems [8,39,47,61]. Diffusion generators (Sora [13], Cosmos [2], Genie [14]) enable high-fidelity simulation and planning [3,9,12,18,38,45,57,62,66,68,71], but often at the cost of efficiency and limited policy integration [59]. LLM-based approaches simulate dynamics via prompting [60,70], yet suffer from modality misalignment and memory degradation over long horizons. MVoT [40] generates intermediate image visualizations for spatial reasoning but is limited to narrow 2D scenarios without memory. WorldVLA [15] unifies action and world modeling for robotic manipulation but predicts action sequences in a single step, forgoing intermediate imagination and trajectory-level memory. In contrast, UniWM introduces structured memory into a unified multimodal backbone, enabling agents to both imagine while acting and remember over time, which jointly addresses the alignment and stability issues of modular designs.

Goal-Conditioned Navigation is a natural testbed for world models, as it requires tight coupling between perception and policy [20,23]. Policy-centric methods [50,51,53] map observations directly to actions without modeling environment dynamics. Navigation-oriented world models instead predict future observations to support temporally informed planning [65]: PathDreamer [37] used GAN-based simulation but relied on auxiliary inputs (*e.g.*, semantic maps), limiting generalization [43]; NWM [9] integrates video prediction into the navigation loop and selects actions via CEM-based trajectory optimization; while CEM scores complete candidate trajectories, it still proposes actions independently of the world model's visual dynamics, leaving planning and perception loosely coupled. In response, we propose a unified multimodal backbone that aligns action prediction with observation imagination, enabling end-to-end navigation through temporally grounded dynamics modeling.

3 Methodology

We present **UniWM**, a unified, memory-augmented world model that performs *planning* and *visualization* within a single autoregressive multimodal backbone. We **first** introduce preliminaries that replaces the disjoint planner-world-model pair with one multimodal LLM augmented by hierarchical memory (Sec. 3.1). We **then** detail unified training, including multimodal tokenization and role-specific objectives for planning and world modeling (Sec. 3.2), and **finally** describe hierarchical memory for stable long-horizon rollouts at inference time (Sec. 3.3).

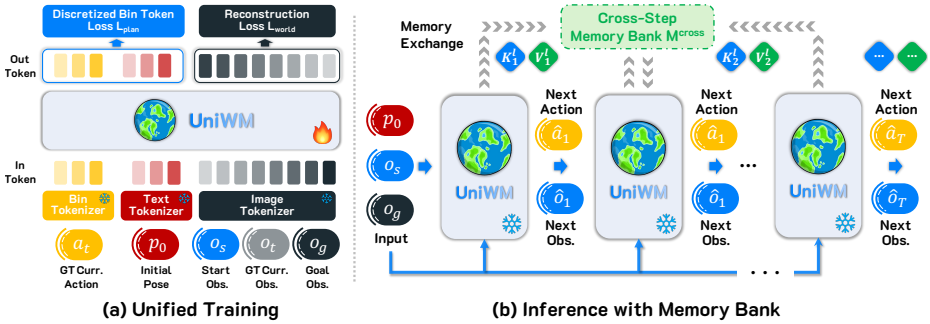


Fig. 2: Overview of the UniWM framework. (a) Training: planner and world-model samples are interleaved within a single unified multimodal autoregressive backbone, optimized jointly with the discretized bin-token loss $\mathcal{L}_{\text{plan}}$ and the reconstruction loss $\mathcal{L}_{\text{world}}$; bin/text/image tokenizers map actions, pose, and observations to tokens. **(b) Inference:** a hierarchical memory supplies intra- and cross-step KV states ($\mathcal{M}_t^{\text{intra}}$ caches the current observation; $\mathcal{M}_t^{\text{cross}}$ accumulates prior steps) to augment attention, yielding robust trajectory-consistent alternating predictions of $\hat{a}_t$ (next action) and $\hat{o}_t$ (next observation). See Fig. 3 for the detailed memory mechanism.

3.1 Preliminaries & Unified Formulation

Given an egocentric RGB observation o_s at the start, the initial agent pose $p_0 \in \mathbb{R}^3$ (position and yaw), and a goal observation o_g, the agent predicts a sequence of navigation actions $A_T = \{\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_T\}$ to reach the goal [53]. Each action $\hat{a}_t$ is either a continuous control command $(\mathbf{u}_t, \phi_t)$ or a terminal `Stop`, where $\mathbf{u}_t \in \mathbb{R}^2$ denotes planar translation (forward/backward, left/right) and $\phi_t \in \mathbb{R}$ denotes yaw rotation [9]. Actions are executed sequentially, and the agent must make monotonic progress toward o_g until issuing `Stop`.

World Models for Navigation. World models [25] predict future environment states (often image frames or video segments) conditioned on the current state and context: $\hat{s}_{t+1} = \mathcal{W}(\hat{s}_t, \mathbf{c})$, where $\hat{s}_t$ is the current state, $\hat{s}_{t+1}$ is the predicted next state, and $\mathcal{W}$ is the learned dynamics model. The context $\mathbf{c}$ may include the executed action a_t, natural-language instructions, observation history, or other factors [48]. In navigation, world models $\mathcal{W}$ serve as imagination engines that anticipate future observations to guide planning.

A common instantiation couples two modules [9]: a **planner** that selects the next action given the current observation and the goal, and a **world model** that simulates the consequent observation conditioned on the chosen action and global cues such as the start and goal views:

$$\hat{a}_{t+1} = \mathcal{P}(\hat{o}_t, o_s, o_g), \quad \hat{o}_{t+1} = \mathcal{W}(\hat{o}_t, \hat{a}_{t+1}, o_s, o_g), \tag{1}$$

where $\hat{o}_t$ is the current observation, $\hat{a}_{t+1}$ is the action proposed by the planner $\mathcal{P}$, and $\hat{o}_{t+1}$ is the next observation visualized by $\mathcal{W}$. The start and goal observations (o_s, o_g) provide the global navigation context. The two modules operate in a closed loop: $\mathcal{P}$ selects $\hat{a}_{t+1}$ conditioned on $\hat{o}_t$ and (o_s, o_g), while $\mathcal{W}$

predicts $\hat{o}_{t+1}$ given $\hat{o}_t$ and $\hat{a}_{t+1}$, which is then fed back into $\mathcal{P}$. This iterative cycle enables imagination-based planning, allowing agents to simulate prospective action-observation trajectories before execution in the real environment. However, the modular training of $\mathcal{P}$ and $\mathcal{W}$ often leads to state–action misalignment, i.e., a discrepancy between the planner's implicit state assumption and the world model's predicted state, which degrades performance in complex and partially observable settings.

Unified World Model with Memory. To address these limitations, we replace the modular pair $(\mathcal{P}, \mathcal{W})$ with a single multimodal backbone, UniWM, that tightly couples planning and visualization. UniWM is augmented with a hierarchical memory bank $\mathcal{M}_t = \{\mathcal{M}_t^{\text{intra}}, \mathcal{M}_t^{\text{cross}}\}$ that fuses short-term evidence with longer-range trajectory context (Fig. 2). At each step, UniWM performs:

$$(\hat{a}_{t+1},\, \hat{o}_{t+1}) \;=\; \textsf{UniWM}\left(\hat{o}_t,\, o_s,\, o_g,\, p_0,\, \mathcal{M}_t\right), \tag{2}$$

where UniWM alternates between two substeps within the same MLLM backbone F_θ: (i) *action prediction* and (ii) *navigation imagination*. Both are executed by F_θ and jointly learned via interleaved planner and world-model training with tailored objectives (Fig. 2a; Sec. 3.2). During inference, hierarchical memory augments attention by integrating immediate evidence with longer-horizon context (Fig. 2b; Sec. 3.3), improving temporal coherence across rollouts.

- **Navigation Planner (Action Prediction):** Given current observation $\hat{o}_t$, conditioned on start and goal observations $(o_s,\, o_g)$, initial pose p_0, and memory bank $\mathcal{M}_t$, F_θ predicts next action $\hat{a}_{t+1}$:

$$\hat{a}_{t+1} = F_\theta(\hat{o}_t, o_s, o_g, p_0, \mathcal{M}_t). \tag{3}$$

- **World Model (Navigation Visualization):** Given current observation $\hat{o}_t$ and action $\hat{a}_{t+1}$, conditioned on (o_s, o_g), p_0, and $\mathcal{M}_t$, F_θ predicts the next observation $\hat{o}_{t+1}$ after executing $\hat{a}_{t+1}$:

$$\hat{o}_{t+1} = F_\theta(\hat{o}_t, \hat{a}_{t+1}, o_s, o_g, p_0, \mathcal{M}_t). \tag{4}$$

This design allows F_θ to act jointly as a **navigation planner** and a **world model**, alternating between roles until a terminal `Stop` is issued. During training, planner and world-model samples are interleaved so that F_θ learns both behaviors within a single autoregressive framework. At inference, a hierarchical memory bank augments F_θ by caching key–value states at both intra- and cross-step levels, enabling the integration of immediate observations with longer-range trajectory context. This unified formulation ensures consistent, memory-augmented world modeling throughout navigation.

3.2 Unified Training Scheme

Next, we describe how **UniWM** is trained as an autoregressive MLLM over text and image tokens. We build on Chameleon and Anole [17,55], which use a unified causal Transformer to jointly model multimodal token sequences (Fig. 2a).

Data Preprocessing. Each navigation trajectory provides two complementary sample types aligned with Eq. 3 and Eq. 4. For the **navigation planner**, a sample contains (o_s, o_g, o_t, p_0) with target $\hat{a}_{t+1}$. For the **world model**, the input additionally includes a_{t+1} and the target becomes $\hat{o}_{t+1}$. Visual observations are inserted as `<image>` placeholders in structured multimodal prompts, and we use a sliding window to extract multiple samples per trajectory. Due to space limits, kindly refer to **Appendix** for prompt design and examples. During training, planner and world-model samples are interleaved within the same batch to promote shared representations across roles.

Multimodal Tokenization. We employ three tokenizers to unify visual and textual inputs. Following [24, 55], a vector-quantized (VQ) image tokenizer discretizes images (o_s, o_g, o_t) into visual tokens via a learned codebook, while a byte-pair encoding (BPE) tokenizer [55] encodes pose p_0 and text prompts into text tokens. Actions a_t are mapped to discrete bin tokens using the bin tokenizer, which we discuss below. The resulting token sequences are fed to a causal Transformer for joint multimodal modeling.

Training Objective. To optimize our model for the distinct characteristics of the navigation planner and world model, we introduce tailored training objectives. At each iteration, our autoregressive MLLM jointly processes samples from both roles, producing logits across the unified vocabulary.

• **Discretized Bin Token Loss (Navigation Planner).** We propose a new classification-based approach for training the planner, which formulates continuous action prediction as multi-class classification over discretized motion bins. Each navigation action $a_t \in \mathbb{R}^3$ is represented as (x_t, y_t, ϕ_t), where x_t and y_t denote planar translations and ϕ_t denotes yaw rotation. We uniformly partition each dimension into fixed-size bins with size $b = 0.01$, computing bin index as $\lfloor |v|/b \rfloor$ for value v. We use separate positive and negative token prefixes to encode the sign and another prefix for the target dimension. For example, x-axis translation with $v = 0.03$ is encoded as `<dx_pos_bin_03>`. This scheme represents all three dimensions as special bin tokens from disjoint token sets: $\mathcal{T}_x$, $\mathcal{T}_y$, and $\mathcal{T}_\phi$. Let $P(t_i)$ denote the model's predicted distribution over all vocabulary tokens at decoding position i. We supervise the planner using **discretized bin token loss** over each action dimension:

$$\mathcal{L}_{\text{plan}} = \frac{1}{3} \sum\nolimits_{k \in \{x,y,\phi\}} \Big(-\log P\big(t_i = t_k^* \,\big|\, t_i \in \mathcal{T}_k\big) \Big) + \mathcal{L}_{\text{CE}}, \tag{5}$$

where t_k^* is the ground-truth bin token in dimension k, and $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss for output text tokens as output may also include text action `Stop`.

• **Reconstruction Loss (World Model).** We introduce a **reconstruction loss** to enforce fidelity in the predicted future observations to encourage accurate navigation visualization. Given ground-truth visual embedding $\mathbf{v}_i$ for token i (out of n tokens in the next observation $\hat{o}_{t+1}$) and the visual codebook embeddings $\mathcal{E} = \{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$ where N is the total number of visual token vocabulary:

$$\mathcal{L}_{\text{world}} = \frac{1}{n} \sum\nolimits_{i=1}^{n} \|\mathbf{v}_i, \mathcal{E}\|^2 \cdot P(t_i), \tag{6}$$

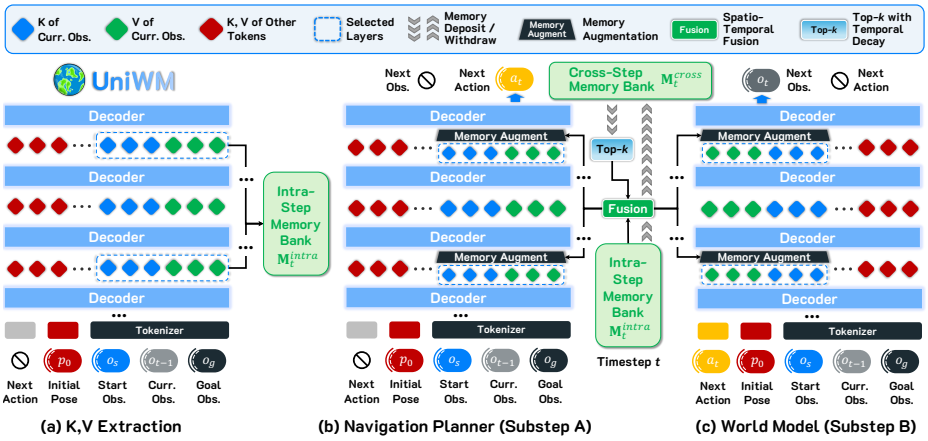


Fig. 3: Overview of hierarchical memory bank mechanism ($\mathcal{M}_t^{\text{intra}}$ & $\mathcal{M}_t^{\text{cross}}$). **(a)** KV (keys/values) extracted from selected layers are deposited into $\mathcal{M}_t^{\text{intra}}$ at the beginning of each step t (Eq. 7). **(b)(c)** $\mathcal{M}_t^{\text{intra}}$ is merged with the accumulated cross-step memory $\mathcal{M}_t^{\text{cross}}$ via top-k similarity gating (Eq. 8) and exponential temporal decay (Eq. 9), yielding a fused memory (Eq. 10) that augments attention for both the planner and the world-model substeps (Eq. 11) to promote trajectory-consistent predictions. At the end of step t, $\mathcal{M}_t^{\text{intra}}$ (with timestamp t) is appended to $\mathcal{M}_t^{\text{cross}}$ for reliable reuse at step $t{+}1$, enabling robustly efficient rollouts.

where $\sum_{i=1}^{n} \|\mathbf{v}_i, \mathcal{E}\|^2$ is the similarity vector indicating distances between $\mathbf{v}_i$ and all codebook embeddings, with lower similarity referring to larger distances, and $P(t_i) \in \mathbb{R}^{1\times N}$ denotes predicted probability distribution over visual tokens at position i. Throughout training, all tokenizers remain frozen, and only Transformer parameters are updated under autoregressive next-token prediction.

3.3 Inference with Memory Bank

At the inference phase, **UniWM** alternates between two substeps: action prediction and navigation visualization. As illustrated in Fig. 3, UniWM employs a hierarchical two-level memory bank mechanism. The **intra-step** memory $\mathcal{M}_t^{\text{intra}}$ caches key, value (K, V) pairs extracted from the current observation $\hat{o}_{t-1}$ at selected Transformer decoder layers, while the **cross-step** memory $\mathcal{M}_t^{\text{cross}}$ accumulates all past intra-step memories $\mathcal{M}_m^{\text{intra}}$, where ($m \in 1, ..., t-1$) together with their associated step indices t_m. This design allows $\mathcal{M}_t^{\text{cross}}$ to maintain a persistent trajectory-level context, thereby enabling F_θ to integrate both short-term and longer-term dependencies across steps.

Two-level Cache Design. At the beginning of each step t, the intra-step memory bank $\mathcal{M}_t^{\text{intra}}$ is reset to avoid contamination from the previous step: $\mathcal{M}_t^{\text{intra}} \leftarrow \varnothing$. Given the tokenized multimodal input, we identify the span of the current observation $\hat{o}_{t-1}$ by marking its token sequence with two special boundary tokens, <boss> and <eoss>, thereby yielding the index set $\mathcal{I}_t$. We

then extract K, V pairs to form $\mathcal{M}_t^{\text{intra}}$ only from this span at a selected subset of decoder layers $L_{\text{save}} \subseteq L = \{l_0, \ldots, l_{31}\}$:

$$\mathcal{M}_t^{\text{intra}} = \{K_t^{(l)}, V_t^{(l)}\} \;=\; \big\{f_K^{(l)}(\mathbf{x}_{\mathcal{I}_t}),\; f_V^{(l)}(\mathbf{x}_{\mathcal{I}_t})\big\}, \quad \textit{where} \;\; l \in L_{\text{save}}, \tag{7}$$

where $\{K_t^{(l)}, V_t^{(l)}\}$ denotes keys and values obtained from the l-th decoder layer at step t, $\mathbf{x}$ represents the hidden states of the multimodal input sequence at that layer, $\mathbf{x}_{\mathcal{I}_t}$ refers to the slice of hidden states indexed by $\mathcal{I}_t$, and $f_K^{(l)}$ and $f_V^{(l)}$ are the key and value projection mappings in layer l. In parallel, as demonstrated in Fig. 3, the cross-step memory $\mathcal{M}_t^{\text{cross}}$ aggregates selected intra-step caches from previous $t-1$ steps with timestamps t_m: $\mathcal{M}_t^{\text{cross}} = \{(K_m^{(l)}, V_m^{(l)}, t_m)\}_{l \in L_{\text{save}}}$.

Spatio-temporal Fusion. At each action prediction substep of step t, the intra-step memory $\mathcal{M}_t^{\text{intra}}$ is merged with the accumulated cross-step memory $\mathcal{M}_t^{\text{cross}}$ to construct a fused memory $\tilde{\mathcal{M}}_t$, which subsequently enhances the attention mechanism for both substeps. This fusion incorporates spatial similarity selection and temporal recency weighting as shown in Fig. 3:

(i) Similarity gating. We flatten both current and historical keys and compute entry-wise cosine similarity $s_m^{(l)}$. The indices of the top-k most similar entries are collected into the set $h_t^{(l)}$:

$$s_m^{(l)} \;=\; \cos\big(K_t^{(l)},\, K_m^{(l)}\big), \quad h_t^{(l)} \;=\; \text{top-}k\big(\, s_m^{(l)}\big), \quad \text{where} \;\; m \in \{1, \ldots, t-1\}. \tag{8}$$

(ii) Temporal decay. Each selected entry is weighted by an exponential decay factor determined by its recency gap $\Delta t_m = t - t_m$, that larger weights correspond to a stronger influence on subsequent predictions. Here we set $\gamma = 0.2$, which biases the weighting toward more recent steps:

$$\alpha_m^{(l)} \;=\; \frac{\exp\big(-\gamma\, \Delta t_m\big)}{\sum_{j \in h_t^{(l)}} \exp\big(-\gamma\, \Delta t_j\big)}\,. \tag{9}$$

(iii) Memory fusion. The fused memory $\tilde{\mathcal{M}}_t = \{\tilde{K}_t^{(l)}, \tilde{V}_t^{(l)}\}_{l \in L_{\text{save}}}$ is formed by concatenating the current intra-step memory with the weighted historical entries so that historical contributions are explicitly modulated by both spatial similarity and temporal recency. Thus, for $h \in h_t^{(l)},\; l \in L_{\text{save}}$, we have:

$$\tilde{K}_t^{(l)} = \textbf{Concat}\left(K_t^{(l)}, \alpha_h^{(l)} K_h^{(l)}\right), \quad \tilde{V}_t^{(l)} = \textbf{Concat}\left(V_t^{(l)},\, \alpha_h^{(l)} V_h^{(l)}\right). \tag{10}$$

Memory-Augmented Attention. The fused memory $\tilde{\mathcal{M}}_t$ then directly engages in cross-attention computation. The attention mechanism can be formally described as scaled dot-product attention:

$$\tilde{Q}_t^{(l)} = \text{Att}(Q_t^{(l)}, \tilde{K}_t^{(l)}, \tilde{V}_t^{(l)}) = \text{softmax}\left(\frac{Q_t^{(l)} \tilde{K}_t^{(l)\top}}{\sqrt{d_k}}\right) \tilde{V}_t^{(l)}, \tag{11}$$

where $Q_t^{(l)}$ denotes the current query at layer l, and d_k is the key dimension. $\tilde{Q}_t^{(l)}$ subsequently propagates through predictions. This mechanism equips UniWM

with trajectory-consistent reasoning by leveraging both current observations and temporally structured historical memories.

Rollout Procedure. The full inference loop (detailed algorithm in **Appendix**) is: at each step t, UniWM resets $\mathcal{M}_t^{\text{intra}}$, extracts and fuses KV states via Eqs. 7–10, then alternates between action prediction (Eq. 3) and observation generation (Eq. 4) under memory-augmented attention (Eq. 11). Current $\mathcal{M}_t^{\text{intra}}$ is then appended to $\mathcal{M}_t^{\text{cross}}$, and process repeats until a `Stop` action is emitted.

4 Experiments

Datasets. We evaluate in two settings using six open-source datasets.

- **Setting (i)** *Egocentric Navigation with Wheeled/Legged Robots*: **Go Stanford** [29], **ReCon** [49], **SCAND** [36], and **HuRoN** [30] are used for training and in-domain evaluation, spanning indoor corridors to socially compliant outdoor paths. **TartanDrive** [56] is held out for unseen testing; its visible ego-robot structures induce a realistic distribution shift.
- **Setting (ii)** *Humanoid Navigation*: we use the navigation subset of the **1X Humanoid Dataset** [1] with 25-DoF joint-angle actions. Following [5], we train a separate UniWM instance due to incompatible action representations.

For Setting (i), we normalize per-frame displacement by the average step size, filter backward motions [9, 53] and trajectories shorter than three steps, and segment visual streams into sub-scenes via Qwen2.5-VL [6]. For 1X Humanoid, we retain only navigation episodes. After preprocessing, the trajectory counts are: Go Stanford (4457/496), ReCon (4652/517), SCAND (2560/285), HuRoN (4642/516), 1X Humanoid (6599/733), and TartanDrive (eval only: 500).

Evaluation Metrics. We report two suites of metrics. **(1) Navigation quality:** Absolute Trajectory Error (ATE), Relative Pose Error (RPE) [54], and Success Rate (SR). SR counts success when the final distance to the goal is below the agent's average step size (meters). **(2) Visualization quality:** SSIM [58], PSNR [31], LPIPS [69], and DreamSim [21]. To assess long-horizon stability under rollout, we also compute SSIM@n, PSNR@n, LPIPS@n, and DreamSim@n.

Implementation Details. UniWM is fine-tuned from GAIR Anole-7B [17] (4096-token context) with all tokenizers frozen. Images are resized to 448×448 and discretized into 784 visual tokens. We update only LoRA [33] adapters (rank = 16) on the Transformer's *qkv* projections [44]. Training uses AdamW for 20 epochs (lr 2×10^{-4}, batch size 8) on 4×A100-80GB GPUs. At inference, two boundary tokens (<boss>/<eoss>, IDs 8196/8197) trigger KV deposit into the intra-step memory bank, and we extract KV states from decoder layers $\{l_0, l_7, l_{15}, l_{23}, l_{31}\}$. All baselines are **retrained from scratch** on the same four training datasets (Go Stanford, ReCon, SCAND, HuRoN), except Anole-7B (zero-shot prompting). All models are evaluated zero-shot on TartanDrive under identical conditions. Since Aether [72] does not release training code, we use its checkpoint and report it only for visualization, as it operates in a camera-pose action space incompatible with our robot-action benchmarks. Due to space limits, kindly refer to **Appendix** for additional implementation and metrics details.

Table 1: Comparisons with SOTA Methods upon Goal-Conditioned Visual Navigation on evaluation splits of Go Stanford, ReCon, SCAND, and HuRoN with SR, ATE, and RPE. Best results for each metric are in bold.

Method	Go Stanford			ReCon			SCAND			HuRoN		
	SR ↑	ATE ↓	RPE ↓	SR ↑	ATE ↓	RPE ↓	SR ↑	ATE ↓	RPE ↓	SR ↑	ATE ↓	RPE ↓
GNM [50]	0.27	1.11	0.31	0.72	0.70	0.20	0.49	0.51	0.21	0.36	1.07	0.35
VINT [51]	0.29	1.09	0.35	0.68	0.84	0.28	0.45	0.58	0.28	0.30	1.19	0.43
NoMaD [53]	0.33	0.94	0.30	0.71	0.77	0.21	0.50	0.54	0.23	0.37	0.92	0.33
Anole-7B [17]	0.18	2.18	0.73	0.41	1.74	0.69	0.29	1.37	0.71	0.20	1.92	0.78
NWM [9]	0.45	0.80	0.27	0.79	0.58	0.17	0.55	0.41	0.19	0.41	0.73	0.28
UniWM (w/o $\mathcal{M}$)	0.71	0.32	0.10	0.82	0.35	0.12	0.61	0.36	0.14	0.70	0.42	0.15
UniWM (w/ only $\mathcal{M}_t^{\text{intra}}$)	0.73	0.29	**0.09**	0.85	0.38	0.13	0.64	0.33	**0.13**	0.74	0.44	0.15
UniWM (w/ $\mathcal{M}_t^{\text{intra}}$ & $\mathcal{M}_t^{\text{cross}}$)	**0.75**	**0.22**	**0.09**	**0.93**	**0.34**	**0.11**	**0.68**	**0.32**	**0.13**	**0.76**	**0.38**	**0.13**

4.1 Comparison to State-of-the-Art Approaches

Navigation Performance. Table 1 reports goal-conditioned navigation results on four in-domain datasets (Go Stanford, ReCon, SCAND, HuRoN), and Fig. 4 shows qualitative comparisons (additional results are placed in the **Appendix**). We compare UniWM against policy-centric baselines GNM [50], VINT [51], and NoMaD [53], as well as Anole-7B [17] under zero-shot prompting. We also include NWM [9], which performs planning via MPC with a CDiT world model. UniWM consistently outperforms all baselines. Even without memory, UniWM yields large gains in SR and improves ATE/RPE across datasets. Adding intra-step memory further stabilizes predictions, while cross-step memory strengthens long-horizon consistency, producing the best overall performance.

Visualization Performance. Table 2 evaluates imagination quality. We compare with Diamond [3] (UNet), Aether [72] (geometry-aware model built on CogVideoX [64]), and NWM [9] (CDiT). UniWM is competitive across all metrics. For one-step prediction, it achieves the best SSIM and DreamSim. Under open-loop rollouts, UniWM remains stable with SSIM@5 = 0.350, preserving semantic consistency and reducing compounding errors over longer horizons.

4.2 Ablation Study & Analyses

1. How do context size and token length affect performance? Table 3 varies both factors under Anole-7B's fixed 4096-token window, revealing a trade-off between temporal coverage and spatial resolution. Increasing either context or token length improves results (*e.g.*, 2×484→4×484 or 2×484→2×625). Comparing 1×784 with 2×625 and 4×484 suggests that, under a fixed token budget, spatial resolution contributes more than additional context frames.

2. Do $\mathcal{L}_{\text{plan}}$ and $\mathcal{L}_{\text{world}}$ help training? We evaluate both losses against a label-smoothing baseline $\mathcal{L}_{\text{LS}}$ (used only in this ablation). As shown in Fig. 5, replacing $\mathcal{L}_{\text{LS}}$ with $\mathcal{L}_{\text{world}}$ substantially improves visualization quality (including rollouts) and also benefits navigation. Combining $\mathcal{L}_{\text{plan}}$ and $\mathcal{L}_{\text{world}}$ performs best overall: $\mathcal{L}_{\text{plan}}$ yields larger navigation gains than $\mathcal{L}_{\text{world}}$ (SR +0.12 vs. +0.10), suggesting that $\mathcal{L}_{\text{world}}$ helps navigation primarily via improved imagination, whereas $\mathcal{L}_{\text{plan}}$ directly optimizes action accuracy.

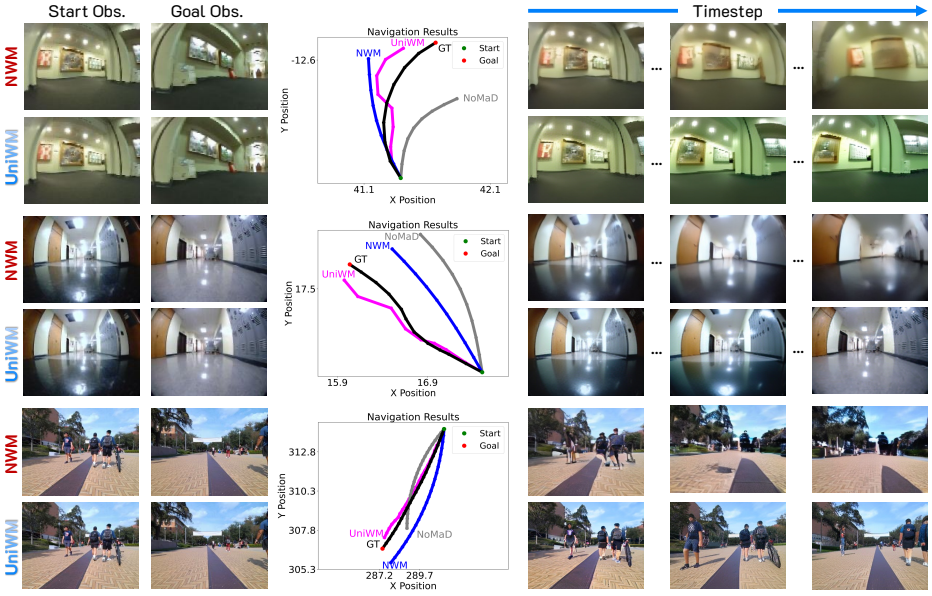


Fig. 4: Qualitative Comparisons on Go Stanford and HuRoN datasets across UniWM, NWM, and NoMaD. Qualitative results here include both static indoor environments and outdoor scenarios with moving pedestrians. The central trajectory plots highlight the difference between predicted A_T and the ground-truth.

Table 2: Comparisons with SOTA methods on visual quality assessment, averaged over evaluation splits of Go Stanford, ReCon, SCAND, and HuRoN.

Method	SSIM ↑	PSNR ↑	LPIPS ↓	DreamSIM ↓	SSIM@5 ↑	PSNR@5 ↑	LPIPS@5 ↓	DreamSIM@5 ↓
Diamond [3]	0.311	9.837	0.410	0.131	0.186	6.352	0.582	0.252
Aether [72]	0.373	10.952	0.271	0.068	0.235	7.183	0.514	0.158
NWM [9]	0.389	11.420	0.318	0.089	0.256	7.755	0.494	0.174
UniWM	**0.457**	**13.607**	**0.254**	**0.041**	**0.350**	**10.874**	**0.435**	**0.126**

3. Should the navigation planner and world model be trained jointly? Table 4 compares unified UniWM against a variant in which the planner and world model are trained separately with identical data and schedule. The unified model consistently outperforms across both navigation and visualization metrics, confirming that joint training more effectively aligns imagination with control.

4. Do we need both intra-step and cross-step memory at inference? Table 1 compares three variants: no memory, intra-step only, and intra+cross. Intra-step memory improves SR and stabilizes pose estimates across datasets; adding cross-step memory further improves long-horizon performance and achiev es the best SR/RPE (0.78/0.11), showing that cross-step context provides complementary gains beyond intra-step stabilization.

5. How many layers should be included in the memory bank? Table 5 varies the number of memory-augmented layers (with both $\mathcal{M}_t^{\text{intra}}$ and

Table 3: Impact of Context Size and Image Token Length on both navigation and visualization performance, averaged over four datasets. All settings are evaluated without memory banks. Best results for each metric are highlighted in **bold**.

Context	Token Len.	Navigation			Visualization					
		SR ↑	ATE ↓	RPE ↓	SSIM ↑	LPIPS ↓	DreamSIM ↓	SSIM@5 ↑	LPIPS@5 ↓	DreamSIM@5 ↓
1	784	**0.71**	**0.36**	**0.13**	**0.457**	0.254	**0.041**	**0.350**	**0.435**	**0.126**
2	625	0.68	0.39	0.15	0.448	**0.247**	0.051	0.336	0.451	0.137
2	484	0.55	0.53	0.26	0.365	0.328	0.084	0.258	0.515	0.192
4	484	0.64	0.44	0.19	0.425	0.285	0.052	0.315	0.462	0.141

Table 4: Comparison of the unified UniWM and the separate planner + world-model setting on both navigation (SR, ATE, RPE) and visualization metrics (SSIM, LPIPS, DreamSim). Best results for each metric are highlighted in **bold**.

Method	SR ↑	ATE ↓	RPE ↓	SSIM ↑	LPIPS ↓	DreamSIM ↓	SSIM@5 ↑	LPIPS@5 ↓	DreamSIM@5 ↓
UniWM (Separate)	0.65	0.41	0.16	0.443	0.280	0.055	0.329	0.470	0.154
UniWM	**0.71**	**0.36**	**0.13**	**0.457**	**0.254**	**0.041**	**0.350**	**0.435**	**0.126**

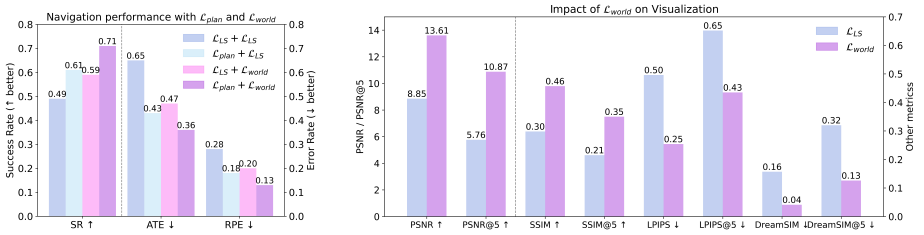


Fig. 5: Impact of discretized bin-token loss ($\mathcal{L}_{\text{plan}}$) and reconstruction Loss ($\mathcal{L}_{\text{world}}$) on navigation (left) and visualization (right) performance, averaged over evaluation splits of Go Stanford, ReCon, SCAND, and HuRoN. X-axis arrows indicate whether higher or lower values are preferable.

$\mathcal{M}_t^{\text{cross}}$ enabled). Moderate multi-depth integration (3–7 layers) steadily improves SR/ATE/RPE, with 5 layers offering the best trade-off. Dense integration (16–32 layers) degrades performance and increases compute and KV overhead; we therefore adopt 5 layers in all experiments.

6. Does goal conditioning affect generalization? We retrain UniWM without the goal image in the visualization substep, keeping all other settings fixed. On unseen TartanDrive, removing goal conditioning reduces performance (SR 0.33 *vs*. 0.35, ATE 1.37 *vs*. 1.20, RPE 0.51 *vs*. 0.46), indicating that goal conditioning improves generalization rather than hindering it.

7. Why UniWM predicts action and observation at different substeps? Table 6 compares two strategies: *predict both* (jointly outputting $\hat{a}_{t+1}$ and $\hat{o}_{t+1}$ in one forward pass) *vs*. *interleave* (alternating planner and world-model substeps during both training and inference). Across all datasets, interleaving yields higher SR and lower ATE/RPE, empirically validating our design choice.

Table 5: Impact of number of selected layers included in memory bank on navigation performance of UniWM (with $\mathcal{M}_t^{\text{intra}}$ & $\mathcal{M}_t^{\text{cross}}$) on evaluation splits of four in-domain datasets. Best results for each metric are in bold.

Layer Num	Go Stanford			ReCon			SCAND			HuRoN		
	SR ↑	ATE ↓	RPE ↓	SR ↑	ATE ↓	RPE ↓	SR ↑	ATE ↓	RPE ↓	SR ↑	ATE ↓	RPE ↓
1	0.71	0.30	0.10	0.84	0.36	0.12	0.62	0.35	0.14	0.72	0.41	0.14
3	0.74	0.27	0.09	0.89	0.35	0.11	0.66	0.33	0.13	0.75	0.39	0.14
5	**0.75**	**0.22**	**0.09**	**0.93**	**0.34**	**0.11**	0.68	0.32	**0.13**	**0.76**	**0.38**	**0.13**
7	0.74	0.25	0.09	0.91	0.35	0.12	**0.69**	**0.31**	0.13	0.74	0.38	0.14
16	0.61	0.46	0.23	0.70	0.58	0.26	0.52	0.49	0.24	0.57	0.55	0.22
32	0.58	0.52	0.26	0.67	0.64	0.29	0.49	0.55	0.27	0.54	0.61	0.25

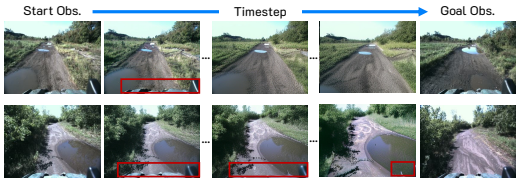
Table 6: Comparison of navigation performance under different step strategies across four datasets. Best results for each metric are in bold.

Step Strategy	Go Stanford			ReCon			SCAND			HuRoN		
	SR ↑	ATE ↓	RPE ↓	SR ↑	ATE ↓	RPE ↓	SR ↑	ATE ↓	RPE ↓	SR ↑	ATE ↓	RPE ↓
Predict both	0.65	0.38	0.13	0.80	0.41	0.15	0.57	0.39	0.17	0.63	0.47	0.20
Interleave	**0.71**	**0.32**	**0.10**	**0.82**	**0.35**	**0.12**	**0.61**	**0.36**	**0.14**	**0.70**	**0.42**	**0.15**

Table 7: Zero-shot navigation performance evaluated on TartanDrive (unseen) without finetuning.

Method	SR ↑	ATE ↓	RPE ↓
GNM [50]	0.16	2.45	0.79
VINT [51]	0.13	2.38	0.79
NoMaD [53]	0.18	2.23	0.77
Anole-7B [17]	0.15	2.12	0.83
NWM [9]	0.27	1.61	0.62
UniWM (w/o $\mathcal{M}$)	0.35	1.20	0.46
UniWM (w/ only $\mathcal{M}_t^{\text{intra}}$)	0.38	1.04	0.41
UniWM (w/ $\mathcal{M}_t^{\text{intra}}$ & $\mathcal{M}_t^{\text{cross}}$)	**0.42**	**0.95**	**0.37**

Fig. 6: Qualitative Results in unseen environments (from the TartanDrive dataset) with UniWM. Red boxes denote ego-robot parts.



4.3 Generalization in Unseen Environments

We evaluate zero-shot generalization on the unseen TartanDrive split (no finetuning) in Table 7 and Fig. 6. UniWM generalizes strongly: even without memory it improves SR and reduces pose errors. Adding $\mathcal{M}_t^{\text{intra}}$ stabilizes predictions, and further enabling $\mathcal{M}_t^{\text{cross}}$ improves long-horizon consistency, yielding best overall results. This confirms that UniWM transfers reliably to unseen environments.

Error Cases & Limitations. TartanDrive observations occasionally contain visible ego-robot parts. As illustrated in Fig. 6, UniWM preserves these cues in the first-step prediction, but they may fade during rollouts. We attribute this to domain gap: the training datasets rarely include ego-robot regions, so the model treats them as background and effectively "inpaints" them away, leading to inconsistency with ground-truth frames in unseen settings.

Table 8: Humanoid navigation performance assessment with UniWM and other baselines trained from scratch on the 1X Humanoid Dataset.

Method	SR ↑	ATE ↓	RPE ↓
GNM [50]	0.41	0.82	0.29
VINT [51]	0.38	0.89	0.33
NoMaD [53]	0.44	0.76	0.27
Anole-7B [17]	0.20	1.95	0.73
NWM [9]	0.51	0.57	0.24
UniWM (w/o $\mathcal{M}$)	0.78	0.31	0.15
UniWM (w/ only $\mathcal{M}_t^{\text{intra}}$)	0.79	0.29	0.14
UniWM (w/ $\mathcal{M}_t^{\text{intra}}$ & $\mathcal{M}_t^{\text{cross}}$)	**0.81**	**0.28**	**0.14**

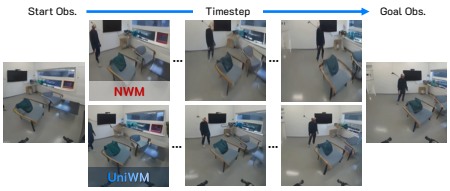


Fig. 7: Qualitative Results on the 1X Humanoid Dataset with UniWM and NWM. The model generates egocentric observations consistent with 25-DoF joint-angle navigation commands.

4.4 Extension to Humanoid Navigation

To assess scalability to high-dimensional action spaces, we evaluate on the navigation subset of the 1X Humanoid Dataset [1] with 25-DoF joint-angle commands. All baselines except Anole-7B (evaluated via direct prompting) are retrained from scratch under identical conditions. As shown in Table 8, UniWM consistently outperforms all baselines across SR, ATE, and RPE, and hierarchical memory provides further gains. Qualitative comparison in Fig. 7 shows that UniWM produces more spatially coherent rollouts than NWM, better preserving complex structures such as room layouts and ongoing human activities in the scene. UniWM also faithfully retains the humanoid robot's visible body parts (*e.g.*, arms) in the egocentric view, which NWM tends to blur over time.

5 Conclusion

We presented UniWM, a unified memory-augmented world model that couples visual imagination and navigation planning within a single multimodal autoregressive backbone. By jointly modeling perception, prediction, and control, UniWM mitigates state-action misalignment, while hierarchical memory fuses short-term observations with longer-range trajectory context to stabilize long-horizon rollouts. Across six benchmarks, including zero-shot evaluation on TartanDrive and 25-DoF humanoid navigation on the 1X Humanoid Dataset, UniWM achieves higher SR and lower ATE/RPE than strong baselines. Future work can explore adaptive token allocation, uncertainty-aware planning, and closed-loop deployment on real robots.

Acknowledgments

This work was supported in part by the University of Washington Faculty Startup Fund, the Carwein–Andrews Ph.D. Fellowship, the UW Graduate School Top Scholar Award, and PacTrans seed and small project funding through the U.S. Department of Transportation Region 10 University Transportation Center.

References

1. 1X Technologies: 1X World Model Challenge (Jun 2024)
2. Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al.: Cosmos world foundation model platform for physical AI. arXiv preprint arXiv:2501.03575 (2025)
3. Alonso, E., Jelley, A., Micheli, V., Kanervisto, A., Storkey, A.J., Pearce, T., Fleuret, F.: Diffusion for world modeling: Visual details matter in Atari. Advances in Neural Information Processing Systems **37**, 58757–58791 (2024)
4. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15619–15629 (2023)
5. Bagchi, A., Bao, Z., Bharadhwaj, H., Wang, Y.X., Tokmakov, P., Hebert, M.: Walk through paintings: Egocentric world models from internet priors. arXiv preprint arXiv:2601.15284 (2026)
6. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923 (2025)
7. Baldassarre, F., Szafraniec, M., Terver, B., Khalidov, V., Massa, F., LeCun, Y., Labatut, P., Seitzer, M., Bojanowski, P.: Back to the features: DINO as a foundation for video world models. arXiv preprint arXiv:2507.19468 (2025)
8. Ball, P.J., Bauer, J., Belletti, F., et al.: Genie 3: A new frontier for world models. https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/, last accessed on March 2, 2026 (2025)
9. Bar, A., Zhou, G., Tran, D., Darrell, T., LeCun, Y.: Navigation world models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15791–15801 (2025)
10. Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., Ballas, N.: Revisiting feature prediction for learning visual representations from video. arXiv preprint arXiv:2404.08471 (2024)
11. Bartoccioni, F., Ramzi, E., Besnier, V., Venkataramanan, S., Vu, T.H., Xu, Y., Chambon, L., Gidaris, S., Odabas, S., Hurych, D., Marlet, R., Boulch, A., Chen, M., Éloi Zablocki, Bursuc, A., Valle, E., Cord, M.: VaViM and VaVAM: Autonomous driving through video generative modeling. arXiv preprint arXiv:2502.15672 (2025)
12. Bian, H., Kong, L., Xie, H., Pan, L., Qiao, Y., Liu, Z.: DynamicCity: Large-scale 4D occupancy generation from dynamic scenes. In: International Conference on Learning Representations (2025)
13. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., et al.: Video generation models as world simulators. OpenAI Blog **1**(8), 1 (2024)
14. Bruce, J., Dennis, M.D., Edwards, A., Parker-Holder, J., Shi, Y., Hughes, E., Lai, M., Mavalankar, A., Steigerwald, R., Apps, C., et al.: Genie: Generative interactive environments. In: International Conference on Machine Learning (2024)
15. Cen, J., Yu, C., Yuan, H., Jiang, Y., Huang, S., Guo, J., Li, X., Song, Y., Luo, H., Wang, F., et al.: Worldvla: Towards autoregressive action world model. arXiv preprint arXiv:2506.21539 (2025)
16. Chaplot, D.S., Gandhi, D., Gupta, S., Gupta, A., Salakhutdinov, R.: Learning to explore using active neural slam. arXiv preprint arXiv:2004.05155 (2020)

17. Chern, E., Su, J., Ma, Y., Liu, P.: ANOLE: An open, autoregressive, native large multimodal models for interleaved image-text generation. arXiv preprint arXiv:2407.06135 (2024)
18. Chi, H., Qiu, D., Su, H., Liu, H., Li, Z., Zhang, H., Lv, C.: Driver-wm: A driver-centric traffic-conditioned latent world model for in-cabin dynamics rollout. arXiv preprint arXiv:2605.05092 (2026)
19. Ding, J., Zhang, Y., Shang, Y., Zhang, Y., Zong, Z., Feng, J., Yuan, Y., Su, H., Li, N., Sukiennik, N., et al.: Understanding world or predicting future? a comprehensive survey of world models. ACM Computing Surveys (2024)
20. Frey, J., Mattamala, M., Chebrolu, N., Cadena, C., Fallon, M., Hutter, M.: Fast traversability estimation for wild visual navigation. In: Robotics: Science and Systems. vol. 19 (2023)
21. Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., Isola, P.: DreamSim: Learning new dimensions of human visual similarity using synthetic data. arXiv preprint arXiv:2306.09344 (2023)
22. Fu, Z., Kumar, A., Agarwal, A., Qi, H., Malik, J., Pathak, D.: Coupling vision and proprioception for navigation of legged robots. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17273–17283 (2022)
23. Fung, P., Bachrach, Y., Celikyilmaz, A., Chaudhuri, K., Chen, D., Chung, W., Dupoux, E., Gong, H., Jégou, H., Lazaric, A., et al.: Embodied AI agents: Modeling the world. arXiv preprint arXiv:2506.22355 (2025)
24. Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., Taigman, Y.: Make-a-scene: Scene-based text-to-image generation with human priors. In: European Conference on Computer Vision. pp. 89–106. Springer (2022)
25. Ha, D., Schmidhuber, J.: World models. arXiv preprint arXiv:1803.10122 **2**(3) (2018)
26. Hafner, D., Lillicrap, T., Ba, J., Norouzi, M.: Dream to control: Learning behaviors by latent imagination. arXiv preprint arXiv:1912.01603 (2019)
27. Hafner, D., Lillicrap, T., Norouzi, M., Ba, J.: Mastering Atari with discrete world models. arXiv preprint arXiv:2010.02193 (2022)
28. Hafner, D., Pasukonis, J., Ba, J., Lillicrap, T.: Mastering diverse domains through world models. arXiv preprint arXiv:2301.04104 (2024)
29. Hirose, N., Sadeghian, A., Vázquez, M., Goebel, P., Savarese, S.: GONet: A semi-supervised deep learning approach for traversability estimation. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 3044–3051 (2018)
30. Hirose, N., Shah, D., Sridhar, A., Levine, S.: SACSoN: Scalable autonomous control for social navigation. IEEE Robotics and Automation Letters **9**(1), 49–56 (2023)
31. Hore, A., Ziou, D.: Image quality metrics: PSNR vs. SSIM. In: IEEE International Conference on Pattern Recognition. pp. 2366–2369 (2010)
32. Hu, A., Russell, L., Yeo, H., Murez, Z., Fedoseev, G., Kendall, A., Shotton, J., Corrado, G.: GAIA-1: A generative world model for autonomous driving. arXiv preprint arXiv:2309.17080 (2023)
33. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: LoRA: Low-rank adaptation of large language models. International Conference on Learning Representations (2022)
34. Hu, T., Liu, X., Wang, S., Zhu, Y., Liang, A., Kong, L., Zhao, G., Gong, Z., Cen, J., Huang, Z., Hao, X., Li, L., Song, H., Li, X., Ma, J., Shen, S., Zhu, J., Tao, D., Liu, Z., Liang, J.: Vision-language-action models for autonomous driving: Past, present, and future. arXiv preprint arXiv:2512.16760 (2025)

35. Hu, X., Yin, W., Jia, M., Deng, J., Guo, X., Zhang, Q., Long, X., Tan, P.: DrivingWorld: Constructing world model for autonomous driving via video GPT. arXiv preprint arXiv:2412.19505 (2024)
36. Karnan, H., Nair, A., Xiao, X., Warnell, G., Pirk, S., Toshev, A., Hart, J., Biswas, J., Stone, P.: Socially compliant navigation dataset (SCAND): A large-scale dataset of demonstrations for social navigation. IEEE Robotics and Automation Letters **7**(4), 11807–11814 (2022)
37. Koh, J.Y., Lee, H., Yang, Y., Baldridge, J., Anderson, P.: PathDreamer: A world model for indoor navigation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14738–14748 (2021)
38. Kong, L., Xie, S., Gong, Z., Li, Y., Chu, M., Liang, A., Dong, Y., Hu, T., Qiu, R., Li, R., Hu, H., Lu, D., Yin, W., Ding, W., Li, L., Song, H., Zhang, W., Ma, Y., et al.: The RoboSense challenge: Sense anything, navigate anywhere, adapt across platforms. arXiv preprint arXiv:2601.05014 (2026)
39. Kong, L., Yang, W., Mei, J., Liu, Y., Liang, A., Zhu, D., Lu, D., Yin, W., Hu, X., Jia, M., Deng, J., Zhang, K., Wu, Y., Yan, T., Gao, S., Wang, S., Li, L., Pan, L., Liu, Y., Zhu, J., Ooi, W.T., Hoi, S.C.H., Liu, Z.: 3D and 4D world modeling: A survey. arXiv preprint arXiv:2509.07996 (2025)
40. Li, C., Wu, W., Zhang, H., Xia, Y., Mao, S., Dong, L., Vulić, I., Wei, F.: Imagine while reasoning in space: Multimodal visualization-of-thought. arXiv preprint arXiv:2501.07542 (2025)
41. Liang, A., Kong, L., Yan, T., Liu, H., Yang, W., Huang, Z., Yin, W., Zuo, J., Hu, Y., Zhu, D., Lu, D., Liu, Y., Jiang, G., Li, L., Li, X., Zhuo, L., Ng, L.X., Cottereau, B.R., Gao, C., Pan, L., Ooi, W.T., Liu, Z.: WorldLens: Full-spectrum evaluations of driving world models in real world. arXiv preprint arXiv:2512.10958 (2025)
42. Liang, A., Liu, Y., Yang, Y., Lu, D., Li, L., Kong, L., Zhao, H., Ooi, W.T.: LiDARCrafter: Dynamic 4D world modeling from LiDAR sequences. In: AAAI Conference on Artificial Intelligence. vol. 40 (2026)
43. Lin, B., Nie, Y., Wei, Z., Chen, J., Ma, S., Han, J., Xu, H., Chang, X., Liang, X.: NavCoT: Boosting LLM-based vision-and-language navigation via learning disentangled reasoning. arXiv preprint arXiv:2403.07376 (2024)
44. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in Neural Information Processing Systems **36**, 34892–34916 (2023)
45. Mei, J., Yang, Y., Yang, X., Wen, L., Lv, J., Shi, B., Liu, Y.: Vision-centric 4d occupancy forecasting and planning via implicit residual world models. arXiv preprint arXiv:2510.16729 (2025)
46. Mirowski, P., Pascanu, R., Viola, F., Soyer, H., Ballard, A.J., Banino, A., Denil, M., Goroshin, R., Sifre, L., Kavukcuoglu, K., et al.: Learning to navigate in complex environments. arXiv preprint arXiv:1611.03673 (2016)
47. Parker-Holder, J., et al.: Genie 2: A large-scale foundation world model. https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model, last accessed on March 2, 2026 (2024)
48. Russell, L., Hu, A., Bertoni, L., Fedoseev, G., Shotton, J., Arani, E., Corrado, G.: GAIA-2: A controllable multi-view generative world model for autonomous driving. arXiv preprint arXiv:2503.20523 (2025)
49. Shah, D., Eysenbach, B., Kahn, G., Rhinehart, N., Levine, S.: Rapid exploration for open-world navigation with latent goal models. arXiv preprint arXiv:2104.05859 (2021)
50. Shah, D., Sridhar, A., Bhorkar, A., Hirose, N., Levine, S.: GNM: A general navigation model to drive any robot. arXiv preprint arXiv:2210.03370 (2022)

51. Shah, D., Sridhar, A., Dashora, N., Stachowicz, K., Black, K., Hirose, N., Levine, S.: ViNT: A foundation model for visual navigation. arXiv preprint arXiv:2306.14846 (2023)
52. Song, M., Deng, X., Zhou, Z., Wei, J., Guan, W., Nie, L.: A survey on diffusion policy for robotic manipulation: Taxonomy, analysis, and future directions. Authorea Preprints (2025)
53. Sridhar, A., Shah, D., Glossop, C., Levine, S.: NoMaD: Goal masked diffusion policies for navigation and exploration. In: IEEE International Conference on Robotics and Automation. pp. 63–70 (2024)
54. Sturm, J., Burgard, W., Cremers, D.: Evaluating egomotion and structure-from-motion approaches using the TUM RGB-D benchmark. In: IEEE/RSJ International Conference on Intelligent Robot Systems Workshops. vol. 13, p. 6 (2012)
55. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024)
56. Triest, S., Sivaprakasam, M., Wang, S.J., Wang, W., Johnson, A.M., Scherer, S.: TartanDrive: A large-scale dataset for learning off-road dynamics models. In: IEEE International Conference on Robotics and Automation. pp. 2546–2552 (2022)
57. Valevski, D., Leviathan, Y., Arar, M., Fruchter, S.: Diffusion models are real-time game engines. arXiv preprint arXiv:2408.14837 (2024)
58. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
59. Xiao, Z., Lan, Y., Zhou, Y., Ouyang, W., Yang, S., Zeng, Y., Pan, X.: WorldMem: Long-term consistent world simulation with memory. arXiv preprint arXiv:2504.12369 (2025)
60. Xing, E., Deng, M., Hou, J., Hu, Z.: Critiques of world models. arXiv preprint arXiv:2507.05169 (2025)
61. Xiong, Z., Song, Y., Kang, H., Yan, Q., Jiang, L., Yang, J., Fu, Z., Fotiadis, S., Wang, A., Liu, Z., et al.: Actworld: From explorable to interactive world model via action-aware memory. arXiv preprint arXiv:2606.17730 (2026)
62. Xiong, Z., Ye, X., Yaman, B., Cheng, S., Lu, Y., Luo, J., Jacobs, N., Ren, L.: Unidrive-wm: Unified understanding, planning and generation world model for autonomous driving. arXiv preprint arXiv:2601.04453 (2026)
63. Xu, X., Liang, A., Liu, Y., Li, L., Kong, L., Liu, Z., Liu, Q.: U4D: Uncertainty-aware 4D world modeling from LiDAR sequences. arXiv preprint arXiv: 2512.02982 (2025)
64. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: CogVideoX: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
65. Yao, X., Gao, J., Xu, C.: NavMorph: A self-evolving world model for vision-and-language navigation in continuous environments. arXiv preprint arXiv:2506.23468 (2025)
66. Yu, J., Qin, Y., Wang, X., Wan, P., Zhang, D., Liu, X.: GameFactory: Creating new games with generative interactive videos. arXiv preprint arXiv:2501.08325 (2025)
67. Yue, J., Huang, Z., Chen, Z., Wang, X., Wan, P., Liu, Z.: Simulating the world model with artificial intelligence: A roadmap. arXiv preprint arXiv:2511.08585 (2025)
68. Zhang, K., Tang, Z., Hu, X., Pan, X., Guo, X., Liu, Y., Huang, J., Yuan, L., Zhang, Q., Long, X.X., Cao, X., Yin, W.: Epona: Autoregressive diffusion world model for autonomous driving. In: IEEE/CVF International Conference on Computer Vision. pp. 27220–27230 (2025)

69. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)
70. Zhao, G., Wang, X., Zhu, Z., Chen, X., Huang, G., Bao, X., Wang, X.: DriveDreamer-2: LLM-enhanced world models for diverse driving video generation. In: AAAI Conference on Artificial Intelligence. vol. 39, pp. 10412–10420 (2025)
71. Zhu, D., Hu, Y., Liu, Y., et al.: SPIRAL: Semantic-aware progressive LiDAR scene generation. arXiv preprint arXiv:2505.22643 (2025)
72. Zhu, H., Wang, Y., Zhou, J., Chang, W., Zhou, Y., Li, Z., Chen, J., Shen, C., Pang, J., He, T.: Aether: Geometric-aware unified world modeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8535–8546 (2025)