# World-Ego Modeling for Long-Horizon Evolution in Hybrid Embodied Tasks

**Zuyao Lin**[1,2,3] **Jianhui Zhang**[3,4] **Peidong Jia**[5] **Xiaoguang Zhao**[1]
**Shanghang Zhang**[5] **Xingyu Chen** [3] ✉

[1]Institute of Automation, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Zhongguancun Academy [4]Shanghai Jiaotong University [5]Peking University

https://zgca-hmi-lab.github.io/WEM
https://github.com/ZGCA-HMI-Lab/WEM
https://huggingface.co/Zoorao/WEM
https://huggingface.co/datasets/Zoorao/HTEWorld




Figure 1. Conceptual illustration of World-Ego Modeling. Embodied evolution involves two heterogeneous dynamics: the world preserves persistent, instruction-agnostic scene regularities, whereas the ego captures robot-centric, instruction-conditioned motion and interaction. World-Ego Modeling decomposes these dynamics to enable stable long-horizon rollouts for hybrid navigation-manipulation tasks.

## Abstract

World models are widely explored in embodied intelligence, yet they typically predict distinct evolutions of the world and the ego within a single stream, where the world captures persistent instruction-agnostic scene regularities and the ego captures robot-centric instruction-conditioned dynamics. This world-ego entanglement leads to a degradation in long-horizon embodied scenarios, particularly in hybrid tasks with interleaved navigation and manipulation behaviors. In this paper, we introduce *World-Ego Modeling*, a new conceptual paradigm that decomposes future evolution into world and ego components. We define the world-ego boundary from three perspectives, i.e., motion-, semantic-, and intention-based views, and analyze three disentanglement strategies with post-, pre-, and full disentanglement. Further, we instantiate this paradigm as the World-Ego Model (WEM), a unified embodied world model that couples an implicit separate world-ego planner with a cascade-parallel mixture-of-experts (CP-MoE) diffusion generator. To enable rigorous evaluation, we further construct HTEWorld, the first benchmark for long-horizon world modeling with hybrid navigation-manipulation tasks, providing 125K video clips (over 4.5M frames) with fine-grained action annotations and 300 multi-turn evaluation trajectories (over 2K instructions). Extensive experiments show that WEM achieves state-of-the-art performance on HTEWorld while remaining competitive on existing manipulation-only benchmarks.

✉ *Corresponding author.*

# 1 Introduction

World models are essential to embodied AI, as they learn the physical dynamics to predict future consequences [1, 2, 3], generate synthetic data [4, 5], and serve as policy simulators [6, 7, 8, 9, 10]. Recent video-based world models [11, 12, 13, 14, 15, 16] have shown strong capabilities in generating realistic future rollouts [17, 18, 19, 20, 21, 22]. In general, an embodied world model needs to simultaneously predict the world and the ego (i.e., embodiment), while recent efforts usually ignore the distinction between them.

As known, the *world* captures persistent, instruction-agnostic scene regularities such as layout and object permanence, while the *ego* captures instruction-conditioned dynamics such as robot behavior and object interactions. Related world-ego concepts have appeared in adjacent areas: JEPA-style architectures separate environment evolution from action proposals [23, 24]; ego-vision world models gain controllability by factoring future video into ego motion, object dynamics, and scene composition [25]. Intuitively, distinguishing the two yields an interpretable decomposition of future prediction. As for video generation, separating the world and the ego avoids overloading a single predictive stream with two heterogeneous responsibilities. In terms of embodied prediction, the world and the ego correspond to fundamentally different aspects of the embodied world (i.e., intention-agnostic change vs. intention-driven behavior), so modeling them separately aligns the predictive structure with the underlying physical reality. Hence, we are motivated to explore world-ego modeling towards the next generation of the embodied world model.

Particularly, conventional world models degrade in long-horizon embodied evolution, especially for hybrid navigation-manipulation tasks [26, 27, 28, 29, 30]. We believe this challenge can be effectively handled by the paradigm of world-ego modeling, since long-horizon scene consistency required by navigation aligns with the world's persistent evolution, and the contact-rich physical dynamics required by manipulation align with the ego's instruction-driven behavior. Thereby, we aim to design a unified world evolution for long-horizon and composite embodied tasks under the world-ego concept.

As shown in Fig. 1, this paper introduces **World-Ego Model** (**WEM**) to investigate an essential question: *How should we define the "world" and the "ego", and does embodied world modeling require world-ego disentanglement?* First, we explore the world-ego boundary from three views: i) the *motion-based view* uses motion cues (e.g., optical flow) to assign contact-induced object dynamics to the ego and other regions to the world; ii) the *semantic-based view* learns an assignment mask to attribute scene regions to the world and the robot or manipulated objects to the ego; and iii) the *intention-based view* lets the world carry visual-history regularities and the ego carry instruction-conditioned dynamics so that the generator implicitly learns how each region uses the two information sources. To instantiate the world-ego concept into a concrete model architecture, we develop a general framework that supports three world-ego definitions and three disentanglement strategies. In detail, there is an implicit planner to infer separate world and ego states via role-conditioned attention (RCA) over asymmetric query groups and a generator to produce video chunks with a cascade-parallel mixture of experts (CP-MoE) and chunk-wise autoregressive diffusion paradigm [31, 32, 33, 34, 35]. As a result, by adopting the semantic world-ego view and full world-ego disentanglement, WEM shows great potential for long-horizon multi-turn embodied evolution in hybrid navigation-manipulation tasks.

This study needs to evaluate both navigation-oriented scene imagination and manipulation-oriented physical simulation under continuous multi-turn embodied instructions, but none of the recent benchmarks [36, 37, 38, 39] meet our requirements. We therefore construct a **Hybrid-Task Embodied World Benchmark (HTEWorld)** on top of BEHAVIOR-1K [40], providing 125K video clips (over 4.5M frames) as training data with fine-grained action-centric annotations and 300 evaluation trajectories that combine interleaved navigation and manipulation stages, together comprising over 2K instructions. On HTEWorld, our WEM can handle hybrid-task rollouts and outperform state-of-the-art models (fine-tuned on the same training data) by a large margin. Besides, WEM remains competitive on existing manipulation-oriented benchmarks [39]. Our contributions are summarized as follows:

- We propose World-Ego Modeling, a new conceptual paradigm for the embodied world model that decomposes future evolution into the world and the ego. We define the world-ego boundary from motion-, semantic-, and intention-based views and analyze the necessity of world-ego disentanglement for embodied evolution.
- We design WEM, a video-based embodied world model with an RCA-based planner and a CP-MoE generator that instantiates the concept of world-ego modeling, to address long-horizon and multi-turn video rollout for hybrid

navigation-manipulation tasks.

- We construct HTEWorld, the first training dataset, benchmark, and metric protocol for long-horizon world evolution with hybrid navigation-manipulation behaviors. Our WEM achieves state-of-the-art performance on HTEWorld and maintains compatibility with the previous manipulation-oriented task.

## 2 Related Work

**Video World Models.** World models predict future states from historical observations, actions, or instructions, serving as internal simulators for planning, data generation, and policy learning [1, 41, 2, 6]. Early methods learn compact latent dynamics for latent imagination. With diffusion models [42, 43] and video generation [44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55], future prediction has moved from low-dimensional state transitions to pixel-level visual rollouts. Cosmos-Predict [12] builds a post-trainable video world foundation model for Physical AI, supporting future observation prediction and synthetic data generation. Recent works further advance interactive video world models: Genie [17] learns a latent action model from unlabeled videos; The Matrix [56], Yume [57], and LIVE [58] target real-time controllable interaction, open-ended scene exploration, and long-horizon consistency modeling; PAN [19] proposes a generative latent prediction framework for general, interactable, and long-horizon world simulation conditioned on history and language actions.

**Embodied Interaction Modeling.** Embodied world models must predict not only environment evolution but also how agent behaviors change the physical world. Existing methods predict future robot observations via video generation [5, 18, 20, 59] or action-conditioned rollouts [60]. Such predictions support a wide range of downstream uses, including data generation [61], policy learning [62, 10], evaluation [39], control [8], and video-level planning [63]. WoW [64] learns a generative world model from large-scale real robot trajectories and uses inverse dynamics to translate imagined outcomes into executable actions. Ctrl-World [65] builds a controllable multi-view world model for robot manipulation, using pose-conditioned memory retrieval and frame-level action conditioning for long-horizon policy imagination, evaluation, and improvement. However, most methods couple scene evolution, robot motion, task intent, and contact dynamics in a single generative stream. Without separating persistent, instruction-agnostic world regularities from robot-centric, instruction-conditioned ego dynamics, they can suffer from temporal inconsistency and weak instruction alignment in long-horizon composite tasks. A few works explore disentangled modeling: JEPA-style architectures [24, 66, 67, 68] predict future states in representation space while separating representation learning from action-conditioned planning or policy heads, and GEM [25] decomposes future egocentric videos into ego motion, object dynamics, and scene composition. However, they mainly target action conditioning, viewpoint motion, or local object dynamics, without systematically studying the world-ego boundary or disentanglement level. To this end, we propose World-Ego Modeling, which disentangles world and ego states in video world models to separately capture scene persistence and robot-centered interaction dynamics in long-horizon composite embodied evolution.

## 3 World-Ego Modeling

### 3.1 Embodied Evolution as World-Ego Prediction

We study multi-turn embodied video generation. Let $\mathbf{O}_0$ be the initial egocentric observation, $\mathbf{V}_{<k} = \{\mathbf{V}_1, \ldots, \mathbf{V}_{k-1}\}$ the visual history before step $k$, and $a_{\leq k} = \{a_1, \ldots, a_k\}$ the instruction sequence up to step $k$. A monolithic embodied video world model predicts the next video chunk as $\hat{\mathbf{V}}_k = \mathcal{M}_\theta(\mathbf{O}_0, \mathbf{V}_{<k}, a_{\leq k})$, collapsing scene structure, viewpoint change, robot motion, and physical interaction into a single entangled predictive pathway.

World-Ego Modeling makes this structure explicit by decomposing the prediction into two complementary states:

$$\mathbf{S}_k^w, \mathbf{S}_k^e = \Phi_\phi(\mathbf{O}_0, \mathbf{V}_{<k}, a_{\leq k}), \quad \hat{\mathbf{V}}_k = \mathcal{D}_\theta(\mathbf{C}_k, a_k, \mathbf{S}_k^w, \mathbf{S}_k^e), \tag{1}$$

where $\Phi_\phi$ is a vision-language state predictor, $\mathcal{D}_\theta$ is the video generator, and $\mathbf{C}_k$ is the local visual condition for the current generation window. Here $\mathbf{S}_k^w$ and $\mathbf{S}_k^e$ are not independent factors of the world; rather, they assign different predictive responsibilities—one for the *world*, one for the *ego*—to different aspects of embodied evolution.


Figure 2. Three perspectives of the world-ego definition. The motion-based view separates the world and ego by the source of visual motion; the semantic-based view separates them by the embodied role of scene entities; and the intention-based view separates them by the source of conditioning information. We adopt the semantic-based view as the default world-ego definition in WEM.

The notions of world and ego, however, are highly general and have been used with markedly different meanings across fields, ranging from the external environment versus the embodied observer in cognitive science [23, 24], to camera motion versus everything else in ego-vision video generation [25], to the robot body versus the scene in embodied manipulation [18, 20, 69]. We treat world and ego as *predictive roles* whose boundary must be specified before any disentanglement can be designed, and we examine three operational definitions next.

### 3.2 Definition of the World and the Ego

We consider three independent perspectives for drawing the world-ego boundary, each defining a self-contained criterion for assigning predictive responsibility. As illustrated in Fig. 2, these views differ in whether the boundary is determined by motion source, semantic entity role, or conditioning source.

**Motion-based view** draws the boundary by the *source of visual motion*. Under a static-scene assumption (i.e., all dynamics in the scene are induced by the embodiment), the camera's egomotion induces a predictable scene flow over the background. Pixels whose motion matches this scene flow are explained by viewpoint change alone and are assigned to the world, revealing world-related content as the camera moves. Pixels whose motion deviates from the scene flow reflect contact-driven object dynamics induced by the embodiment and are assigned to the ego. The residual between the observed and predicted flow—i.e., the object residual flow—serves as the natural proxy.

**Semantic-based view** draws the boundary by the *embodied role of scene entities*. The robot itself and any object currently being manipulated jointly constitute the *ego region*, capturing robot-centric, instruction-conditioned dynamics. The remaining scene with background and unmanipulated objects constitutes the *world region*, capturing persistent, instruction-agnostic regularities responsible for long-horizon scene consistency. The world-ego boundary is therefore interaction-dependent: a movable object belongs to the world before interaction, becomes ego-related once acted upon, and is absorbed back after the interaction completes. A semantic mask serves as the natural proxy.

**Intention-based view** draws the boundary at the *source of conditioning information*. The world reflects what is established by visual history; the ego reflects what is induced by the current instruction. Unlike the motion and semantic views, this view does not partition the future video in pixel space, but instead partitions the conditioning sources and lets the predictor implicitly learn how to extract and integrate information from world- and ego-related conditions to generate future embodied evolution. This perspective relates to Iso-Dream [70], which isolates controllable and noncontrollable visual dynamics, while we separate instruction-induced ego dynamics from history-established world regularities for embodied video generation.

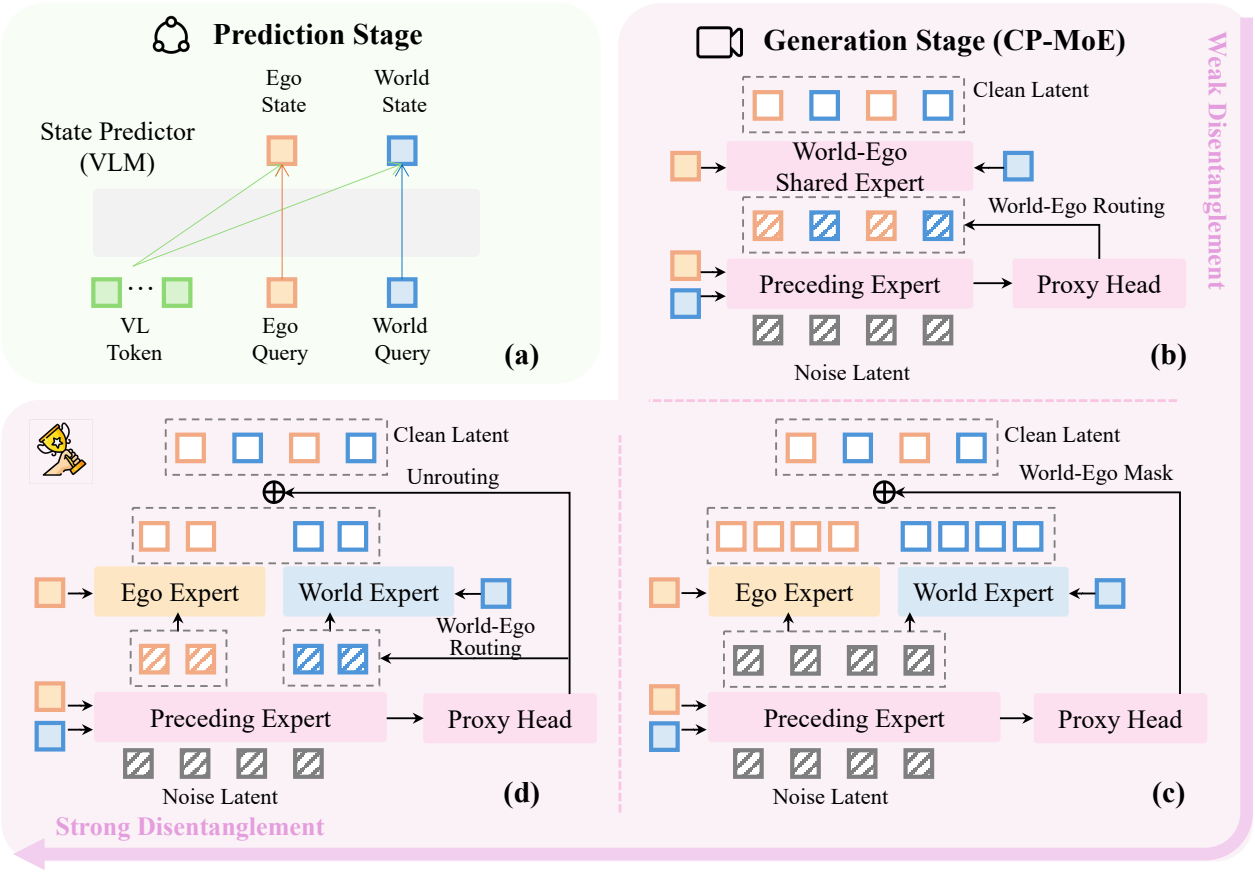


Figure 3. General framework of World-Ego Modeling. Our framework contains two stages. (a) The prediction stage uses a vision-language state predictor to infer separate ego and world states from vision-language tokens. (b–d) The generation stage instantiates different degrees of world-ego disentanglement with a cascade-parallel mixture-of-experts (CP-MoE) generator. The preceding expert predicts a world-ego proxy, which is used to separate the two predictive roles in different ways: (b) pre-disentanglement routes tokens before the rear stage, (c) post-disentanglement fuses the outputs of separate ego and world experts, and (d) full disentanglement combines routing, expert specialization, and unrouting for stronger structural separation. We adopt the semantic-based view as the default world-ego definition in WEM.

### 3.3 A General Framework for World-Ego Modeling with Disentanglement

To explore the three views and disentanglement strategies under identical conditions, we design a general framework as illustrated in Fig. 3. The framework consists of two stages that mirror Eq. (1): a *prediction stage* that infers separate world and ego states (Fig. 3(a)), and a *generation stage* based on a cascade-parallel mixture-of-experts (CP-MoE) generator (Fig. 3(b–d)).

**Prediction stage.** A vision-language state predictor $\Phi_\phi$ takes the visual-language tokens encoding $(\mathbf{O}_0, \mathbf{V}_{<k}, a_{\leq k})$ together with learnable *ego query* and *world query* to produce $\mathbf{S}_k^e$ and $\mathbf{S}_k^w$, providing the generator with two independent conditioning signals.

**Generation stage.** The CP-MoE generator splits the DiT backbone into a shared preceding expert and a specialized rear stage (i.e., ego and world experts). The preceding expert is conditioned on both states and emits a learned *proxy* that operationalizes the world-ego boundary defined in Sec. 3.2: under the semantic view it takes the form of a world-ego mask, and under the

motion view it takes the form of object flow. The proxy is then exploited differently across the three disentanglement variants in Fig. 3(b–d). Fig. 3(b) is the *pre-disentanglement*, which introduces separation at the input of the rear stage. As shown, the proxy partitions preceding-expert tokens into the world and ego groups, which are passed through a single rear module with restricted cross-attention (world tokens attend only to $\mathbf{S}_k^w$, ego tokens only to $\mathbf{S}_k^e$). Fig. 3(c) is the *post-disentanglement*, which introduces separation at the output of the rear stage. Specifically, the rear stage is duplicated into a World Expert and an Ego Expert, both processing the full token sequence under their respective states; the proxy then acts as a soft mask that fuses the two outputs. Fig. 3(d) is the *full disentanglement*, where the proxy first routes preceding-expert tokens to the World and Ego Experts, and then unroutes their outputs back into a single sequence under the same proxy, enforcing separation across input routing, branch processing, and output fusion.

**Alternatives.** As shown in section 5, the *semantic-based view combined with full disentanglement* yields the best long-horizon performance on hybrid navigation-manipulation tasks. We adopt this configuration as the default instantiation of our WEM in subsequent sections.

# 4 World-Ego Model: Instantiating the Paradigm of World-Ego Modeling

WEM follows the two-stage paradigm as shown in Sec. 3.3: a vision-language state predictor $\Phi_\phi$ that maps multi-modal history to compact latent states and a video diffusion generator $\mathcal{D}_\theta$ with a CP-MoE structure that decodes the next video chunk under those states (Fig. 4). We build WEM upon a pretrained VLM [71] and a pretrained video diffusion transformer [45], retaining their priors while introducing the modules required for world-ego factorization.

## 4.1 State Predictor

The state predictor extracts world and ego latent states from the multimodal history of the embodied trajectory (Fig. 4, top). It consists of a pretrained VLM backbone augmented with ego/world queries appended to the end of the input sequence. The input sequence interleaves multimodal history in temporal order. That is, the initial frame is followed by alternating instruction texts $\{a_1, \ldots, a_k\}$ and previously generated video chunks $\{\mathbf{V}_1, \ldots, \mathbf{V}_{k-1}\}$, ending with the current instruction $a_k$ and the two query groups. After a forward pass, the hidden states at the ego- and world-query positions are extracted as $\mathbf{S}_k^e$ and $\mathbf{S}_k^w$, which serve as conditioning signals for the generator. Two design choices realize world-ego separation already at the state level.

**Asymmetric query budgets.** We allocate different numbers of queries to the two groups rather than enforcing equal budgets. Forcing equal capacity would implicitly assume that the two roles carry comparable amounts of information, but they differ in scope: the world encodes persistent scene structure accumulated across long histories, while the ego encodes instruction-conditioned dynamics local to the current step. Decoupling the budgets lets each group allocate capacity according to its own role; the exact ratio is treated as a hyperparameter.

**Role-conditioned attention.** We restrict the attention horizon of each group to the subset of inputs consistent with its predictive role:

- The world queries attend to the entire visual history, i.e., the initial frame, all previous chunks $\{\mathbf{V}_1, \ldots, \mathbf{V}_{k-1}\}$, and all past instructions $\{a_1, \ldots, a_{k-1}\}$ and to one another, but are *blocked* from the current instruction $a_k$ and from ego-query tokens. This anchors $\mathbf{S}_k^w$ to scene regularities accumulated from history, decoupled from the present instruction.
- The ego queries attend to one another, the current instruction $a_k$, and the most recent $K$ instruction-video pairs (one "turn" is one instruction together with its generated chunk). Distant history and world-query tokens are masked out. This anchors $\mathbf{S}_k^e$ to instruction-conditioned dynamics in the current local context.

We refer to the above-mentioned attention pattern as **Role-Conditioned Attention (RCA)**. By giving the two groups disjoint conditioning sources, RCA prevents $\mathbf{S}_k^w$ and $\mathbf{S}_k^e$ from collapsing into a shared representation despite sharing the same backbone.

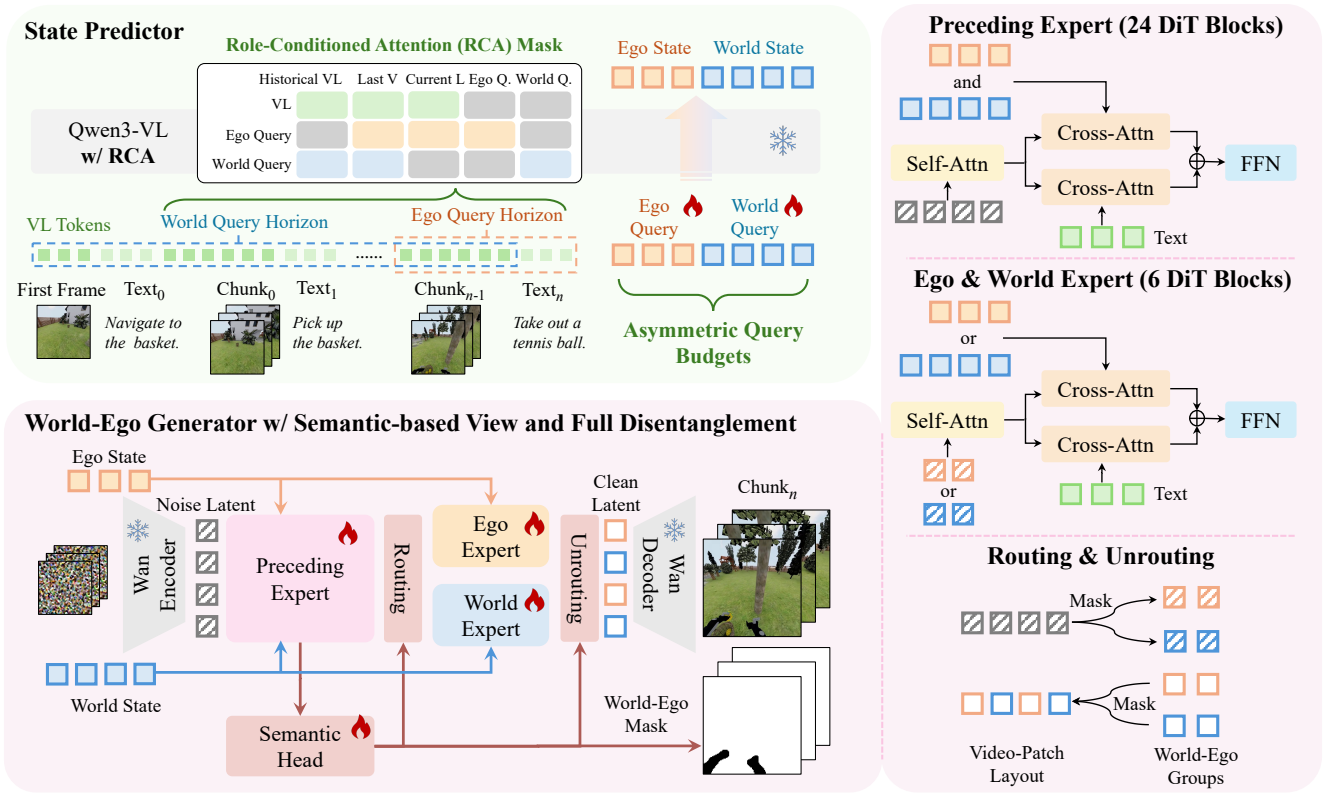


Figure 4. Overview of the World-Ego Model (WEM). WEM instantiates World-Ego Modeling with a semantic-based world-ego view and full disentanglement. The state predictor augments a pretrained VLM with role-conditioned attention (RCA) and asymmetric ego/world queries to infer separate ego and world states from multi-turn vision-language history. The generator restructures a pretrained video DiT into a CP-MoE architecture: a preceding expert jointly conditions on both states to predict a semantic world-ego mask, which routes video tokens to specialized ego and world experts and then unroutes their outputs into a clean latent for the next video chunk.

## 4.2 World-Ego Generator

The generator restructures a pretrained video DiT backbone into a CP-MoE structure (Fig. 4, bottom-left). The DiT blocks are split into an early shared group forming the preceding expert and a later group duplicated into the world and ego experts. Unlike standard sparse MoE [72, 73], all three experts are always active, and specialization arises from predefined role assignment rather than learned routing.

**Preceding expert.** The Preceding Expert serves as a shared encoder that integrates the two states into a common visual representation from which the world-ego boundary can be predicted (Fig. 4, top-right). To this end, each block performs self-attention on the noise latent, followed by two parallel cross-attention streams, attending to the text instruction (preserved from the pretrained DiT), $\mathbf{S}_k^w$, and $\mathbf{S}_k^e$. Two-stream outputs are summed before the FFN. Conditioning on both states simultaneously is what allows the Preceding Expert to produce features that capture the joint configuration of scene context and embodied interaction, a prerequisite for the downstream proxy prediction.

**Role experts.** The world expert and the ego expert realize the structural separation of full disentanglement (Fig. 4, middle-right). They share the Preceding Expert's block topology but operate strictly under a single state (i.e., $\mathbf{S}_k^w$ or $\mathbf{S}_k^e$), while the

text-conditioning cross-attention is preserved unchanged. This asymmetry with the Preceding Expert is intentional: while the Preceding Expert acts as a shared encoder, the role experts perform specialized denoising on disjoint regions of the future video, each grounded only in its corresponding state.

**Semantic head, routing, and unrouting.** The semantic head is a lightweight dense prediction transformer [74] that fuses multiple intermediate features from the preceding expert together with $\mathbf{S}_k^e$ in a coarse-to-fine manner, and outputs a world-ego mask $\mathbf{M}$ over video patches (Fig. 4, bottom). $\mathbf{M}$ serves as the routing signal: world-assigned tokens are dispatched to the world expert and ego-assigned tokens to the ego expert, with each expert's active token set expanded to the spatial neighbors of its assigned region to avoid seam artifacts at the boundary. After the role experts complete their computation, an unrouting module recomposes their per-token outputs into a single sequence under the same $\mathbf{M}$, which is then passed to the video decoder to produce the clean latent.

**Training.** WEM is trained end-to-end with two objectives: a flow-matching loss $\mathcal{L}_{\text{flow}}$ on the recomposed latents (inherited from the pretrained DiT) [75] and a mask-prediction loss $\mathcal{L}_{\text{mask}}$ on the Semantic Head's output. The mask loss combines a class-balanced binary cross-entropy term and a Dice term [76], $\mathcal{L}_{\text{mask}} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{Dice}}$, supervising the predicted mask against the ground-truth world-ego mask derived from the simulator's segmentation labels. The total loss is $\mathcal{L} = \mathcal{L}_{\text{flow}} + \lambda \mathcal{L}_{\text{mask}}$, where $\lambda$ is a tunable weight balancing the two terms.

# 5 Experiments

## 5.1 HTEWorld Benchmark

Existing video-based world-model benchmarks [77, 36, 78, 37, 38, 39] target short-horizon manipulation or single-prompt generation, leaving long-horizon hybrid navigation-manipulation evolution untested. We therefore construct HTEWorld on top of BEHAVIOR-1K [40], providing 125K video clips (over 4.5M frames) with fine-grained annotations and 300 multi-turn evaluation trajectories spanning over 2K instructions. We adopt the 16-metric EWMScore from WorldArena [39] as the primary metric and further introduce 6 HTEWorld-specific metrics: *Rollout Chunk-Boundary Dynamics* (*RCBD*), *Late-Prefix State Alignment* (*LPSA* ), and *Chunk Instruction-Step Retrieval* (*CISR*) for multi-turn continuous generation; and *Phase-Matched Motion Profile Alignment* (*PMPA*) , *Cross-Phase Discriminative Margin* (*CPDM* ), and *Frontier Phase-Hop State Consistency* (*FPHS*) for unified navigation-manipulation generation. Dataset construction, annotation protocol, evaluation protocol, and detailed metric definitions are provided in Appendix A–B.

## 5.2 Experimental Setup

WEM adopts the semantic view with full disentanglement (Sec. 5.3, 5.4), using a frozen Qwen3-VL-2B-Instruct [71] state predictor with 256 learnable queries, split into 192 world and 64 ego queries, and a Wan2.2-TI2V-5B [45] generator following the chunk-wise autoregressive recipe of Xiang et al. [19], Teng et al. [31], Liu et al. [33]. We compare with Cosmos-Predict 2.5 [12] (2B/14B) and WoW-7B [64], all fine-tuned on HTEWorld for 4 epochs on 16×A100 GPUs with learning rate $1 \times 10^{-5}$. Details on WEM training and baseline adaptation are provided in Appendix D.

## 5.3 Design Study I: Which Boundary Should Define World and Ego?

**Setup.** Following Sec. 3.2, we compare three operational definitions: motion-based, semantic-based, and intention-based views. The semantic-based view is implemented with the full-disentanglement architecture used by WEM, while the detailed architectures for the motion-based and intention-based views are provided in Appendix C. All variants are trained under the same protocol.

**Results.** As shown in Table 1, the semantic-based view achieves the best EWMScore, outperforming the motion-based and intention-based views by 2.12 and 2.79 points, respectively. This result supports our choice of semantic world-ego assignment as the default boundary definition of WEM.


Figure 5. Qualitative comparison on HTEWorld. Given the same initial observation and five-step instruction sequence, each model autoregressively generates a long-horizon hybrid navigation–manipulation rollout. The model size reported in parentheses denotes the parameter count of the video-generation DiT backbone. Red circles denote visible artifacts or instruction failures, and the overall column summarizes whether the full rollout is completed successfully. WEM better preserves scene geometry, object consistency, and instruction alignment across the complete trajectory.

**Analysis.** The intention-based view lacks an explicit spatial boundary and may fail to induce effective separation. The motion-based view provides an optical-flow proxy, but camera/object flow decomposition is noisy under large viewpoint changes and contact interactions. In contrast, the semantic-based view directly separates interaction regions from persistent scene regions, while allowing both to move under egomotion.

## 5.4 Design Study II: How Should World and Ego Be Disentangled?

**Setup.** We fix the semantic-based boundary and compare different architectural strategies for world-ego disentanglement. The pre-disentanglement variant separates tokens before the role-specific stage but keeps the downstream computation shared. The post-disentanglement variants use separate world and ego branches and fuse their outputs afterwards; one variant removes the semantic proxy by replacing it with a constant assignment. The full-disentanglement variant uses the semantic proxy for both routing and unrouting, corresponding to our final WEM.

**Results.** Table 2 shows that full disentanglement performs best. Post-disentanglement with a semantic proxy is already strong (EWMScore 61.09), while removing the semantic proxy drops the score to 58.59. Full disentanglement further improves to 61.48.

**Analysis.** Pre-disentanglement is limited because separated tokens are later processed by shared computation. Post-disentanglement uses separate branches, but each branch still sees the full token sequence, causing cross-role interference. Full disentanglement gives the most consistent separation by using the semantic assignment for both routing and recomposition.

Table 1. Design study on world-ego definitions. We report EWMScore on HTEWorld.

| World-ego definition | EWMScore |
|---|---|
| Intention-based view | 58.69 |
| Motion-based view | 59.36 |
| Semantic-based view | **61.48** |

Table 2. Design study on world-ego disentanglement strategies. All variants use the semantic-based view.

| Variant | EWMScore |
|---|---|
| w/o Disent. | 58.40 |
| Pre-Disent. | 58.85 |
| Post-Disent. w/o or w/ Semantic Proxy | 58.59/61.09 |
| Full Disent. (WEM) | **61.48** |

Table 3. Comparison on HTEWorld under the WorldArena's normalized metrics. Higher is better.

| Model | EWMScore | Visual | | | Motion | | | Consistency | | | 3D | | | Control | | | Physics |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | IQ | AQ | JEPA | DD | Flow | MS | BC | SC | Photo | DA | TA | Persp | Sem | Act | Inst | Inter |
| WoW-7B [64] | 53.44 | 64.72 | 49.74 | 1.30 | 22.76 | 25.49 | 67.74 | 66.86 | 63.06 | 38.08 | 82.41 | 28.16 | 95.14 | 86.75 | 3.47 | 78.42 | 80.90 |
| Cosmos-Predict 2.5-2B [12] | 54.83 | 64.40 | 50.21 | 1.26 | 24.62 | 27.23 | 69.33 | 71.54 | 66.88 | **42.53** | 82.11 | 28.78 | 95.28 | 87.05 | 3.52 | 79.60 | 83.02 |
| Cosmos-Predict 2.5-14B [12] | 55.41 | 62.14 | 50.02 | 1.38 | 29.37 | 32.34 | 71.63 | 73.85 | 68.65 | 35.48 | 82.60 | 28.81 | 94.40 | 87.46 | 3.55 | 80.20 | 84.70 |
| PAN-style Baseline [19] | 58.40 | 65.48 | 49.00 | 1.70 | 38.14 | 47.43 | 79.47 | 80.24 | 74.79 | 33.15 | 82.39 | 28.75 | 95.08 | 87.67 | 4.42 | 80.40 | 86.33 |
| WEM | **61.48** | **66.82** | **50.30** | **2.49** | **41.52** | **49.21** | **82.70** | **87.92** | **82.07** | 35.95 | **84.55** | **34.51** | **97.60** | **90.74** | **4.50** | **82.00** | **90.80** |

Table 4. Comparison on HTEWorld w/ navigation-manipulation metrics. Scores are reported in their original scale, and higher is better.

| Model | RCBD | LPSA | CISR | PMPA | CPDM | FPHS |
|---|---|---|---|---|---|---|
| WoW-7B [64] | 0.23 | 0.83 | 0.49 | 0.45 | 0.47 | 0.85 |
| Cosmos-2B [12] | 0.24 | 0.83 | 0.50 | 0.47 | 0.48 | 0.86 |
| Cosmos-14B [12] | 0.26 | 0.83 | 0.51 | 0.48 | 0.49 | 0.85 |
| PAN-style [19] | 0.27 | 0.86 | 0.49 | 0.50 | 0.46 | 0.88 |
| Intention-based | 0.28 | 0.86 | 0.53 | 0.52 | 0.50 | 0.88 |
| Motion-based | 0.29 | 0.87 | 0.53 | 0.54 | 0.51 | 0.89 |
| WEM | **0.31** | **0.87** | **0.57** | **0.54** | **0.52** | **0.89** |

Table 5. Comparison on WorldArena. Related scores are from the WorldArena paper; WEM uses the same protocol.

| Model | EWMScore |
|---|---|
| Wan2.2 [45] | 54.54 |
| WoW [64] | 54.88 |
| Cosmos-Predict 2.5 (Action) [12] | 55.91 |
| CogVideoX [46] | 57.90 |
| WEM (ours) | 58.10 |
| IRASim [79] | 58.12 |
| CtrlWorld [65] | **59.70** |

## 5.5 Comparison with Prior Arts

We compare WEM with representative baselines on HTEWorld using the WorldArena metric suite and the 6 HTEWorld-specific metrics introduced in Sec. 5.1. All models are fine-tuned on the HTEWorld training split and evaluated under the same autoregressive multi-turn rollout protocol. As shown in Table 3, WEM achieves the best EWMScore (61.48), outperforming the PAN-style baseline by 3 points and all compared methods by a larger margin. The gains are especially clear on motion, consistency, 3D, control, and physics-related metrics, suggesting that world-ego disentanglement improves coherent scene evolution and instruction-aligned interaction beyond local visual quality. Table 4 further shows consistent gains across all 6 HTEWorld-specific metrics: RCBD, LPSA, and CISR reflect stronger chunk continuity, instruction alignment, and layout preservation, while PMPA, CPDM, and FPHS indicate better phase-matched motion, camera/object coordination, and long-horizon stability. Finally, Table 5 shows that WEM remains competitive on the original WorldArena benchmark despite being optimized for hybrid tasks, suggesting compatibility with conventional manipulation-oriented evaluation. Qualitative comparisons are shown in Fig. 5.

## 5.6 Role Expert Specialization

We verify that the world and ego experts in CP-MoE develop distinct specializations rather than redundant representations. As illustrated in Fig. 1 and visualized in detail in Fig. 7, the ego expert focuses on robot body parts and manipulated objects while the world expert reproduces stable background structure; the Semantic Head accurately localizes the boundary between the two roles across diverse scenes and phases. We redirect readers to Appendix E for more ablations of WEM.

## 6 Conclusion

We presented World-Ego Modeling, a paradigm that decomposes embodied video prediction into persistent world regularities and robot-centric ego dynamics. We studied motion-, semantic-, and intention-based boundaries, analyzed disentanglement strategies, and instantiated the best design as WEM, coupling a vision-language state predictor with a CP-MoE diffusion generator. We further constructed HTEWorld with interleaved navigation-manipulation trajectories and dedicated multi-turn metrics to evaluate long-horizon hybrid embodied evolution. Experiments show that WEM outperforms representative baselines on HTEWorld while remaining competitive on manipulation-only benchmarks, suggesting explicit world-ego separation as a promising direction for structured and controllable embodied world models.

**Limitations.** WEM is only an initial instantiation of the broader World-Ego Modeling paradigm, and the current study explores a limited set of boundary definitions, disentanglement designs, and simulated embodied settings. More general boundaries, architectures, and real-world applications remain open directions; we provide a detailed discussion in Appendix G.

## References

[1] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *NeurIPS*, 2018.

[2] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 2025.

[3] Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, Yujun Shen, and Yinghao Xu. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026.

[4] Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, et al. Dreamgen: Unlocking generalization in robot learning through video world models. In *CoRL*, 2025.

[5] Mengjiao Yang, Yilun Du, Seyed Kamyar Seyed Ghasemipour, Jonathan Tompson, Leslie Pack Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *ICLR*, 2024.

[6] Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: World models for physical robot learning. In *CoRL*, 2023.

[7] Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loïc Magne, Avnish Narayan, You Liang Tan, Guanzhi Wang, Qi Wang, Jiannan Xiang, Yinzhen Xu, Seonghyeon Ye, Jan Kautz, Furong Huang, Yuke Zhu, and Linxi Fan. FLARE: Robot learning with implicit world modeling. In *CoRL*, 2025.

[8] Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, and Jinwei Gu. Cosmos policy: Fine-tuning video models for visuomotor control and planning. In *ICLR*, 2026.

[9] Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026.

[10] Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning. In *RSS*, 2024.

[11] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

[12] Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical AI. *arXiv preprint arXiv:2511.00062*, 2025.

[13] Yuanhui Huang, Wenzhao Zheng, Yuan Gao, Xin Tao, Pengfei Wan, Di Zhang, Jie Zhou, and Jiwen Lu. Owl-1: Omni world model for consistent long video generation. *arXiv preprint arXiv:2412.09600*, 2024.

[14] Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. WORLDMEM: Long-term consistent world simulation with memory. In *NeurIPS*, 2025.

[15] Ying Yang, Zhengyao Lv, Tianlin Pan, Haofan Wang, Binxin Yang, Hubery Yin, Chen Li, Ziwei Liu, and Chenyang Si. StableWorld: Towards stable and consistent long interactive video generation. *arXiv preprint arXiv:2601.15281*, 2026.

[16] Zile Wang, Zexiang Liu, Jaixing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, et al. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026.

[17] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *ICML*, 2024.

[18] Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. Robodreamer: Learning compositional world models for robot imagination. In *ICML*, 2024.

[19] Jiannan Xiang, Yi Gu, Zihan Liu, Zeyu Feng, Qiyue Gao, Yiyan Hu, Benhao Huang, Guangyi Liu, Yichi Yang, Kun Zhou, et al. PAN: A world model for general, interactable, and long-horizon world simulation. *arXiv preprint arXiv:2511.09057*, 2025.

[20] Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. Tesseract: Learning 4d embodied world models. In *ICCV*, 2025.

[21] Taiye Chen, Xun Hu, Zihan Ding, and Chi Jin. Learning world models for interactive video generation. In *NeurIPS*, 2025.

[22] Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. VideoREPA: Learning physics for video generation through relational alignment with foundation models. In *NeurIPS*, 2025.

[23] Yann LeCun. A path towards autonomous machine intelligence, 2022. URL https://openreview.net/forum?id=BZ5a1r-kVsf. OpenReview position paper, version 0.9.2.

[24] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

[25] Mariam Hassan, Sebastian Stapf, Ahmad Rahimi, Pedro Rezende, Yasaman Haghighi, David Brüggemann, Isinsu Katircioglu, Lin Zhang, Xiaoran Chen, Suman Saha, et al. Gem: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control. In *CVPR*, 2025.

[26] Yu Shang, Lei Jin, Yiding Ma, Xin Zhang, Chen Gao, Wei Wu, and Yong Li. Longscape: Advancing long-horizon embodied world models with context-aware moe. *arXiv preprint arXiv:2509.21790*, 2025.

[27] Byungjun Kim, Taeksoo Kim, Junyoung Lee, and Hanbyul Joo. Dexterous world models. In *CVPR*, 2026.

[28] Xinshuai Song, Weixing Chen, Yang Liu, Vincent Chan, Guanbin Li, and Liang Lin. Towards long-horizon vision-language navigation: Platform, benchmark and method. In *CVPR*, 2025.

[29] Zhuo Xu, Hao-Tien Lewis Chiang, Zipeng Fu, Mithun George Jacob, Tingnan Zhang, Tsang-Wei Edward Lee, Wenhao Yu, Connor Schenck, David Rendleman, Dhruv Shah, Fei Xia, Jasmine Hsu, Jonathan Hoech, Pete Florence, Sean Kirmani, Sumeet Singh, Vikas Sindhwani, Carolina Parada, Chelsea Finn, Peng Xu, Sergey Levine, and Jie Tan. Mobility VLA: Multimodal instruction navigation with long-context VLMs and topological graphs. In *CoRL*, 2025.

[30] Zhenyu Wu, Yuheng Zhou, Xiuwei Xu, Ziwei Wang, and Haibin Yan. MoManipVLA: Transferring vision-language-action models for general mobile manipulation. In *CVPR*, 2025.

[31] Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

[32] Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *NeurIPS*, 2024.

[33] Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. In *ICLR*, 2026.

[34] Haodong Li, Shaoteng Liu, Zhe Lin, and Manmohan Chandraker. Rolling sink: Bridging limited-horizon training and open-ended testing in autoregressive video diffusion. *arXiv preprint arXiv:2602.07775*, 2026.

[35] Hidir Yesiltepe, Tuna Han Salih Meral, Adil Kaan Akan, Kaan Oktay, and Pinar Yanardag. Infinity-RoPE: Action-controllable infinite video generation emerges from autoregressive self-rollout. *arXiv preprint arXiv:2511.20649*, 2025.

[36] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. VBench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE TPAMI*, 2025.

[37] Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E Gonzalez, et al. Worldmodelbench: Judging video generation models as world models. In *NeurIPS*, 2025.

[38] Yufan Deng, Zilin Pan, Hongyu Zhang, Xiaojie Li, Ruoqing Hu, Yufei Ding, Yiming Zou, Yan Zeng, and Daquan Zhou. Rethinking video generation model for the embodied world. In *ICML*, 2026.

[39] Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyou Wang, Lei Jin, Xin Zhang, Yinzhou Tang, Haisheng Su, et al. Worldarena: A unified benchmark for evaluating perception and functional utility of embodied world models. *arXiv preprint arXiv:2602.08971*, 2026.

[40] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, et al. BEHAVIOR-1K: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *CoRL*, 2023.

[41] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *ICLR*, 2020.

[42] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[44] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, Lifang He, and Lichao Sun. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024.

[45] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[46] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2025.

[47] Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, et al. Genie 2: A large-scale foundation world model, 2024. URL https://deepmind.google/blog/genie-2-a-large-scale-foundation-world-model/. Google DeepMind blog.

[48] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-Sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

[49] Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2World: Crafting video diffusion models to interactive world models. In *ICLR*, 2026.

[50] Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, et al. Matrix-game 2.0: An open-source, real-time, and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025.

[51] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. HunyuanVideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[52] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. MotionCtrl: A unified and flexible motion controller for video generation. In *SIGGRAPH*, 2024.

[53] Wonjoon Jin, Qi Dai, Chong Luo, Seung-Hwan Baek, and Sunghyun Cho. FloVD: Optical flow meets video diffusion model for enhanced camera-controlled video synthesis. In *CVPR*, 2025.

[54] Wonbong Jang, Shikun Liu, Soubhik Sanyal, Juan Camilo Perez, Kam Woh Ng, Sanskar Agrawal, Juan-Manuel Perez-Rua, Yiannis Douratsos, and Tao Xiang. Rays as pixels: Learning a joint distribution of videos and camera trajectories. *arXiv preprint arXiv:2604.09429*, 2026.

[55] Xindi Wu, Despoina Paschalidou, Jun Gao, Antonio Torralba, Laura Leal-Taixé, Olga Russakovsky, Sanja Fidler, and Jonathan Lorraine. Motion attribution for video generation. In *ICML*, 2026.

[56] Ruili Feng, Han Zhang, Zhantao Yang, Jie Xiao, Zhilei Shu, Zhiheng Liu, Andy Zheng, Yukun Huang, Yu Liu, and Hongyang Zhang. The matrix: Infinite-horizon world generation with real-time moving control. *arXiv preprint arXiv:2412.03568*, 2024.

[57] Xiaofeng Mao, Shaoheng Lin, Zhen Li, Chuanhao Li, Wenshuo Peng, Tong He, Jiangmiao Pang, Mingmin Chi, Yu Qiao, and Kaipeng Zhang. Yume: An interactive world generation model. *arXiv preprint arXiv:2507.17744*, 2025.

[58] Junchao Huang, Ziyang Ye, Xinting Hu, Tianyu He, Guiyu Zhang, Shaoshuai Shi, Jiang Bian, and Li Jiang. LIVE: Long-horizon interactive video world modeling. *arXiv preprint arXiv:2602.03747*, 2026.

[59] Yu Shang, Xin Zhang, Yinzhou Tang, Lei Jin, Chen Gao, Wei Wu, and Yong Li. Roboscape: Physics-informed embodied world model. In *NeurIPS*, 2025.

[60] Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. BridgeData V2: A dataset for robot learning at scale. In *CoRL*, 2023.

[61] Sheng Chen, Peiyu He, Jiaxin Hu, Ziyang Liu, Yansheng Wang, Tao Xu, Chi Zhang, et al. Astra: Toward general-purpose mobile robots via hierarchical multimodal learning. *arXiv preprint arXiv:2506.06205*, 2025.

[62] Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, Hongyan Zhao, Hanyu Liu, Zhizhong Su, Lei Ma, Hang Su, and Jun Zhu. Motus: A unified latent action world model. *arXiv preprint arXiv:2512.13030*, 2025.

[63] Boyuan Chen, Tianyuan Zhang, Haoran Geng, Kiwhan Song, Caiyi Zhang, Peihao Li, William T. Freeman, Jitendra Malik, Pieter Abbeel, Russ Tedrake, Vincent Sitzmann, and Yilun Du. Large video planner enables generalizable robot control. *arXiv preprint arXiv:2512.15840*, 2025.

[64] Xiaowei Chi, Peidong Jia, Chun-Kai Fan, Xiaozhu Ju, Weishi Mi, Kevin Zhang, Zhiyuan Qin, Wanxin Tian, Kuangzhi Ge, Hao Li, Zezhong Qian, Anthony Chen, Qiang Zhou, Yueru Jia, Jiaming Liu, Yong Dai, Qingpo Wuwu, Chengyu Bai, Yu-Kai Wang, Ying Li, Lizhang Chen, Yong Bao, Zhiyuan Jiang, Jiacheng Zhu, Kai Tang, Ruichuan An, Yulin Luo, Qiuxuan Feng, Siyuan Zhou, Chi-min Chan, Chengkai Hou, Wei Xue, Sirui Han, Yike Guo, Shanghang Zhang, and Jian Tang. WoW: Towards a world omniscient world model through embodied interaction. *arXiv preprint arXiv:2509.22642*, 2025.

[65] Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-world: A controllable generative world model for robot manipulation. In *ICLR*, 2026.

[66] Jingwen Sun, Wenyao Zhang, Zekun Qi, Shaojie Ren, Zezhi Liu, Hanxin Zhu, Guangzhong Sun, Xin Jin, and Zhibo Chen. VLA-JEPA: Enhancing vision-language-action model with latent world model. *arXiv preprint arXiv:2602.10098*, 2026.

[67] Shuang Zeng, Dekang Qi, Xinyuan Chang, Feng Xiong, Shichao Xie, Xiaolong Wu, Shiyi Liang, Mu Xu, and Xing Wei. JanusVLN: Decoupling semantics and spatiality with dual implicit memory for vision-language navigation. In *ICLR*, 2026.

[68] Jiahui Zhang, Yurui Chen, Yueming Xu, Ze Huang, Yanpeng Zhou, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, and Li Zhang. 4D-VLA: Spatiotemporal vision-language-action pretraining with cross-scene calibration. In *NeurIPS*, 2025.

[69] Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. In *RSS*, 2025.

[70] Minting Pan, Xiangming Zhu, Yunbo Wang, and Xiaokang Yang. Iso-Dream: Isolating and leveraging noncontrollable visual dynamics in world models. In *NeurIPS*, 2022.

[71] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[72] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *JMLR*, 2022.

[73] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[74] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021.

[75] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.

[76] Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. Dice loss for data-imbalanced NLP tasks. In *ACL*, 2020.

[77] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. VBench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024.

[78] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, et al. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

[79] Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. IRASim: A fine-grained world model for robot manipulation. In *ICCV*, 2025.

[80] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020.

[81] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[82] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[83] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

[84] Kyunghyun Cho, Bart Van Merriënboer, Çağlar Gulçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *EMNLP*, 2014.

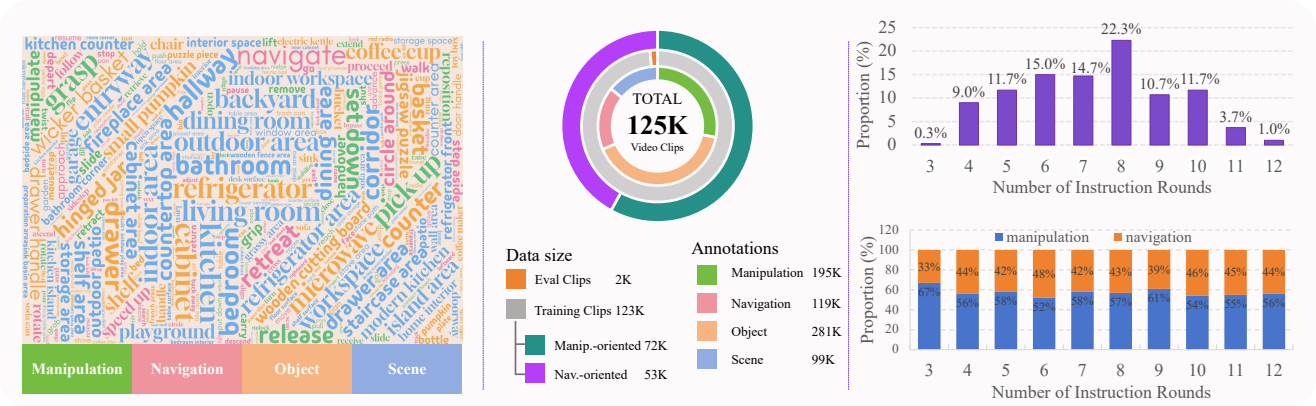


Figure 6. Statistics of the proposed HTEWorld benchmark. HTEWorld provides large-scale training clips and multi-turn evaluation trajectories for hybrid embodied world modeling. We visualize: Left. hybrid-task vocabulary spanning manipulation, navigation, objects, and scenes; Middle. training-set composition, including training/evaluation scale, action-oriented clip types, and annotation categories; and Right. evaluation-trajectory composition, including instruction-round distribution and the manipulation/navigation proportion at each length.

# A HTEWorld Annotation Pipeline

Each training clip is annotated with three complementary signals: a semantic world-ego mask, decomposed optical flow, and a language caption. The semantic world-ego mask is obtained directly from the BEHAVIOR-1K simulator's instance segmentation, which labels the robot body, end-effectors, and currently manipulated objects as the ego region, and the remaining scene as the world region. The flow and caption annotations require dedicated pipelines described below.

**Optical flow extraction and decomposition.** We estimate dense optical flow for each consecutive frame pair using RAFT [80]. The raw flow $\mathbf{F}$ mixes two physically distinct components: camera-induced flow from the robot's ego-motion and contact-induced object flow from manipulation interactions. To separate them, we estimate the camera ego-motion as a homography $\mathbf{H}$ by fitting matched feature pairs with RANSAC across each frame pair. The camera-induced flow field $\mathbf{F}_{\text{cam}}$ is rendered by applying $\mathbf{H}$ to every pixel coordinate, and the residual object flow is obtained as $\mathbf{F}_{\text{obj}} = \mathbf{F} - \mathbf{F}_{\text{cam}}$. Regions with large $\|\mathbf{F}_{\text{obj}}\|$ correspond to objects undergoing contact-driven motion, while regions dominated by $\mathbf{F}_{\text{cam}}$ reflect viewpoint change from navigation. This decomposition yields the motion-based world-ego proxy used in Design Study I (Sec. 5.3) and provides the flow signals for the RCBD and PMPA metrics (Appendix B).

**Language caption generation.** We generate one action-centric caption per clip using `google/gemini-3-flash-preview`. A key technical contribution of our annotation pipeline is **dynamic prompt construction**: rather than using a fixed template, we construct a distinct, context-rich prompt for each clip by fusing four heterogeneous sources of information.

1. **Robot grounding.** A detailed description of the Galaxea R1 embodiment (wheeled bimanual humanoid, egocentric head camera) with explicit arm-placement conventions ("lower-left and lower-right of frame"), ensuring the model correctly interprets egocentric observations without hallucinating a third-person perspective.
2. **Episode trajectory context.** The full ordered list of action labels for the current episode is injected, with the current step highlighted. This allows the model to leverage long-horizon context for object identification (e.g., knowing a plate was placed on a shelf two steps earlier) while being explicitly instructed to *only describe what is visible in the current clip*, preventing trajectory labels from contaminating the visual description.

3. **Temporal phase hint.** The clip's index within its parent action (e.g., "clip 3 of 5, 60% through this action") is provided so the model can reason about whether the robot is approaching, actively manipulating, or retracting, yielding temporally grounded descriptions across multi-clip actions.
4. **Structured output rules.** Strict constraints enforce a single sentence of at most 30 words describing only observable physical motion, require egocentric directional language ("forward", "left", "downward"), mandate arm specificity ("its left gripper", "both arms"), and prohibit verbatim copying of the action label or any meta-reference to camera or frames.

We further apply **intent sanitization** as a pre-generation quality filter: action labels from the simulator occasionally contain garbled or incomplete tokens (e.g., consonant-only suffixes from truncated transcriptions). Before each prompt is built, label tokens are scanned and trailing sequences consisting entirely of consonants with four or more characters are stripped, preventing noisy labels from misleading the model's scene understanding.

Caption generation is parallelized across 24 worker threads, with multiple API keys distributed via round-robin scheduling to maximize throughput while respecting per-key rate limits. Failed requests are retried with exponential backoff (up to 5 attempts, capped at 30 seconds), and each clip's output is written atomically so the pipeline supports incremental restarts without re-processing completed clips. The resulting captions serve as the language supervision signal for all three world-ego views.

## B HTEWorld-Specific Metric Definitions

We introduce six HTEWorld-specific metrics along two axes. Let $\mathbf{V} = (\mathbf{V}_1, \ldots, \mathbf{V}_K)$ denote the autoregressive rollout of $K$ generated chunks, and $\mathbf{V}^* = (\mathbf{V}_1^*, \ldots, \mathbf{V}_K^*)$ the corresponding ground-truth chunks. Each chunk $\mathbf{V}_k$ carries a phase label $p_k \in \{\text{Nav}, \text{Manip}\}$. We denote the first/last frame of chunk $\mathbf{V}_k$ as $\mathbf{V}_k^{(1)}$ and $\mathbf{V}_k^{(-1)}$, respectively. Let $d_p(\cdot,\cdot)$ denote LPIPS [81] perceptual distance, $\mathcal{F}(\cdot,\cdot)$ optical flow magnitude (RAFT [80]), and $\mathcal{H}(\cdot)$ a pretrained video encoder (CLIP [82]).

**Multi-Turn Continuous Generation.** **RCBD** (Rollout Chunk-Boundary Dynamics) measures how faithfully the generated model reproduces the appearance and motion dynamics at each chunk boundary, without rewarding over-smoothing. For each consecutive pair $(\mathbf{V}_k, \mathbf{V}_{k+1})$, we compute an appearance gap $b_k = d_p(\mathbf{V}_k^{(-1)}, \mathbf{V}_{k+1}^{(1)})$ and a motion gap $m_k$ as the optical flow discontinuity across the boundary (flow between the last two frames of $\mathbf{V}_k$ vs. the first two frames of $\mathbf{V}_{k+1}$). A symmetric match score is computed for each gap as $\mathcal{S}(x,y) = \exp\big(-|\log(x/y)|\big)$, measuring the ratio alignment between generated and GT dynamics. The boundary score is the geometric mean $\sqrt{\mathcal{S}(b_k, b_k^*) \cdot \mathcal{S}(m_k, m_k^*)}$, and RCBD is the mean over all $K-1$ boundaries.

**LPSA** (Late-Prefix State Alignment) measures how well the visual state at the end of each generated chunk aligns with the corresponding GT state, with emphasis on later rollout steps. For each chunk $k$, the last $W = 4$ frames of $\mathbf{V}_k$ and $\mathbf{V}_k^*$ are encoded by $\mathcal{H}(\cdot)$, and their cosine similarity $r_k$ is computed. LPSA is defined as the linearly weighted mean $\sum_k k \cdot r_k \,/\, \sum_k k$, so later chunks (which accumulate more rollout error) contribute more to the final score.

**CISR** (Chunk Instruction-Step Retrieval) measures instruction-step alignment via a retrieval task. For each generated chunk $\mathbf{V}_k$, we ask: given its video features $\mathcal{H}(\mathbf{V}_k)$, can the model correctly retrieve the GT chunk $\mathbf{V}_k^*$ among all GT chunks $\{\mathbf{V}_1^*, \ldots, \mathbf{V}_K^*\}$ in the same trajectory? We rank all GT chunks by cosine similarity to $\mathcal{H}(\mathbf{V}_k)$ and report the reciprocal rank of the correct match. CISR is the Mean Reciprocal Rank (MRR) over all $K$ chunks.

**Navigation–Manipulation Generation.** **PMPA** (Phase-Matched Motion Profile Alignment) measures whether the temporal evolution of motion within each chunk matches the corresponding GT chunk, separately for navigation and manipulation phases. For each chunk, we compute a 4-dimensional motion profile over consecutive frame pairs: $[\bar{u}/L,\ \tilde{u}/L,\ \log(1+\tilde{u}/\bar{u}),\ \mathcal{E}(u)]$, where $\bar{u}$ is the median optical flow magnitude, $\tilde{u}$ is the top-20% mean, $L$ is the frame diagonal length, and $\mathcal{E}(u)$ is the normalized flow entropy. Both the generated and GT profiles are resampled to 16 time steps, and their point-wise L2 distance $\delta$ is converted to a score $\exp(-\delta/\tau)$. PMPA is the mean score over all chunks.

**CPDM** (Cross-Phase Discriminative Margin) measures whether the generated video is more similar to its own phase than to the opposite phase. For each generated chunk $\mathbf{V}_k$ with phase $p_k$, we compute a positive similarity $r^+ = \langle \mathcal{H}(\mathbf{V}_k), \mathcal{H}(\mathbf{V}_k^*) \rangle$

(same-phase GT chunk) and a hard-negative similarity $r^- = \max_{j:\, p_j \neq p_k} \langle \mathcal{H}(\mathbf{V}_k), \mathcal{H}(\mathbf{V}_j^*) \rangle$ (most confusable opposite-phase GT chunk). The chunk score is $\sigma\big((r^+ - r^-)/\tau\big)$, where $\sigma$ is the sigmoid and $\tau = 0.05$. CPDM is the mean score over all chunks.

**FPHS** (Frontier Phase-Hop State Consistency) measures visual state consistency in the *spatially localized change region* at each phase-transition boundary. For each boundary where the phase switches ($p_k \neq p_{k+1}$), a window of $R = 4$ frames on each side is extracted from both the generated and GT rollouts. The change region is localized by accumulating GT optical flow magnitudes and selecting the top-20% spatial area. Both generated and GT windows are cropped to this region, and their video feature similarity $\mathcal{H}(\cdot)$ is computed. FPHS is the mean score over all phase-switch boundaries.

## C Motion- or Intention-based Model Variant

This section details the concrete architectural realizations of the motion-based and intention-based world-ego views used in Design Study I (Sec. 5.3). Both variants share the same state predictor and CP-MoE generator backbone as WEM, and differ only in the proxy signal used for routing and fusion.

**Motion-based view.** The motion-based view adopts a post-disentanglement layout: a shared general expert processes the full noisy latent, followed by two parallel role experts (world expert and ego expert) that each consume the shared representation. Rather than using a semantic label as the proxy, a lightweight convolutional flow head tapped from the general expert features predicts a dense object residual flow map $\hat{\mathbf{F}}_{\text{obj}}$. Concretely, features from four intermediate general expert blocks are fused via learned per-block weights and decoded through a multi-scale U-Net head into a per-token flow magnitude. The flow magnitude is converted to a continuous fusion weight $\alpha \in [0, 1]$ per token via $\alpha = \sigma\big((\tau - \|\hat{\mathbf{F}}_{\text{obj}}\|)/\delta\big)$, where $\sigma$ is the sigmoid, $\tau$ is a threshold, and $\delta$ controls sharpness. Tokens with large residual flow (contact-driven ego motion) receive low $\alpha$ and are weighted toward the ego expert output; tokens with small residual flow (stable scene content) receive high $\alpha$ and are weighted toward the world expert output. The final output is the soft combination $\alpha \cdot \mathbf{X}^w + (1 - \alpha) \cdot \mathbf{X}^e$, applied before the output head. The flow head is supervised with the ground-truth object residual flow from the homography-decomposition pipeline (Appendix A) using an L1 loss, jointly with the standard flow-matching loss.

**Intention-based view.** The intention-based view uses a single unified decoder with no explicit proxy signal, mask head, or role-expert split. World-ego separation is instead induced through two complementary design choices: an asymmetric state injection mechanism and a recurrent world-state update.

For state injection, the world state $\mathbf{S}_k^w$ is fed into every decoder layer as cross-attention memory tokens, allowing spatially distributed scene context to selectively influence each token position. The ego state $\mathbf{S}_k^e$ is injected as an AdaLN [83] modulation signal: it is mean-pooled into a compact summary vector that modulates the per-layer scale and shift of each transformer block, providing a global action-level control that uniformly shifts the generation dynamics.

For state update, $\mathbf{S}_k^w$ is not re-extracted from scratch at each turn. Instead, a GRU-style [84] updater produces the final world state by gating between the previous world state $\mathbf{S}_{k-1}^w$ and a new proposal $\hat{\mathbf{S}}_k^w$ extracted from the current context:

$$\mathbf{S}_k^w = \mathbf{G}_k \odot \mathbf{S}_{k-1}^w + (1 - \mathbf{G}_k) \odot \tilde{\mathbf{S}}_k^w, \tag{2}$$

where $\mathbf{G}_k \in [0, 1]^{N \times D}$ is a keep gate dynamically computed from $\mathbf{S}_{k-1}^w$, $\hat{\mathbf{S}}_k^w$, and a mean-pooled summary of the ego state $\mathbf{S}_k^e$, and $\tilde{\mathbf{S}}_k^w$ is a candidate update formed analogously with a separate reset gate. Because $\mathbf{G}_k$ is conditioned on the ego state, the gating can suppress world-state updates triggered by transient ego actions and preferentially retain stable scene information across turns. No auxiliary proxy loss is added; the model is trained with the standard flow-matching loss alone.

## D Training and Evaluation Protocol

**Unified Training Setup.** All models—WEM and all baselines—are fine-tuned on the HTEWorld training split under the same protocol: full-parameter fine-tuning for 4 epochs on 16×NVIDIA A100 80GB GPUs with learning rate $1 \times 10^{-5}$. We

initially attempted LoRA for Cosmos-Predict2.5 and WoW-7B but found it consistently underperformed full fine-tuning, which we attribute to the large domain gap between internet-video pretraining and the egocentric manipulation scenarios in HTEWorld; all baselines therefore use full fine-tuning to match WEM. An EMA of model weights (decay 0.99) is maintained for all models and used for evaluation.

**WEM Training Objectives.** Beyond the shared flow-matching loss $\mathcal{L}_{\text{flow}}$, WEM adds a mask-prediction loss $\mathcal{L}_{\text{mask}}$ on the semantic head:

$$\mathcal{L}_{\text{mask}} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{Dice}}, \tag{3}$$

giving a total loss $\mathcal{L} = \mathcal{L}_{\text{flow}} + \lambda\,\mathcal{L}_{\text{mask}}$, where $\lambda$=0.3 is annealed to 20% of its initial value over training so that semantic supervision is strongest in the early stages. The DPT semantic head taps encoder features at layers $\{5, 9, 13, 17, 21, 23\}$.

**Baseline Conditioning Adaptation.** Cosmos-Predict2.5 and WoW-7B are single-turn models that require adaptation to HTEWorld's multi-turn rollout. To keep the training distribution consistent with inference, both models are trained with a **mixed conditioning schedule**: with probability 0.9 a clip is conditioned on the preceding chunk's latent representation (multi-turn continuation), and with probability 0.1 on the first-frame initialization (trajectory start). This 90/10 mixture is the same for both baselines and mirrors the evaluation rollout without requiring separate training stages.

**Multi-Turn Evaluation Protocol.** WEM is natively chunk-autoregressive and requires no inference adaptation. For Cosmos-Predict2.5, chunk $k$=0 uses the `image2world` mode (first frame + instruction), and each subsequent chunk $k$>0 uses `video2world` mode conditioned on the last $L$=10 latent frames of the preceding generated chunk. For WoW-7B, which operates exclusively in `video2world` mode, chunk $k$=0 uses a repeated-frame conditioning clip (first scene frame tiled 41 times); subsequent chunks condition on a tail clip of the last 41 frames of the preceding chunk. The model generates 82 frames internally, discards the 41 conditioning frames, and uniformly subsamples the remainder to 37 output frames. All models generate 37 frames per chunk at $480\times480$, 16 FPS, 35 diffusion steps; guidance scales are 5 for Cosmos-Predict2.5 and 7 for WoW-7B. EWMScore is computed per-chunk and averaged over all $K$ chunks and all 300 evaluation trajectories.

## E Ablation Study

We ablate three key components of WEM: asymmetric query budgets, role-conditioned attention (RCA), and neighbor-expanded routing. In the first variant, we use equal query budgets (128 queries each for world and ego). In the second variant, we relax the role-specific attention masks so each query group can additionally access the other role's inputs. In the third variant, we disable neighbor-expanded routing, so each role expert only processes tokens assigned by the predicted semantic mask. As shown in Table 6, removing each component degrades performance, confirming that effective world-ego separation requires role-aware state extraction and boundary-aware expert routing.

Table 6. Component ablation of WEM on HTEWorld under the WorldArena evaluation protocol. Higher is better.

| Variant | EWMScore |
|---|---|
| w/o Asymmetric Query Budget | 60.50 |
| w/o Role-Conditioned Attention | 60.64 |
| w/o Neighbor-Expanded Routing | 59.57 |
| WEM | **61.48** |

## F Role Expert Specialization

Fig. 7 visualizes the outputs of WEM's world and ego experts under the mask predicted by the Semantic Head, illustrating that each expert specializes in its assigned region.

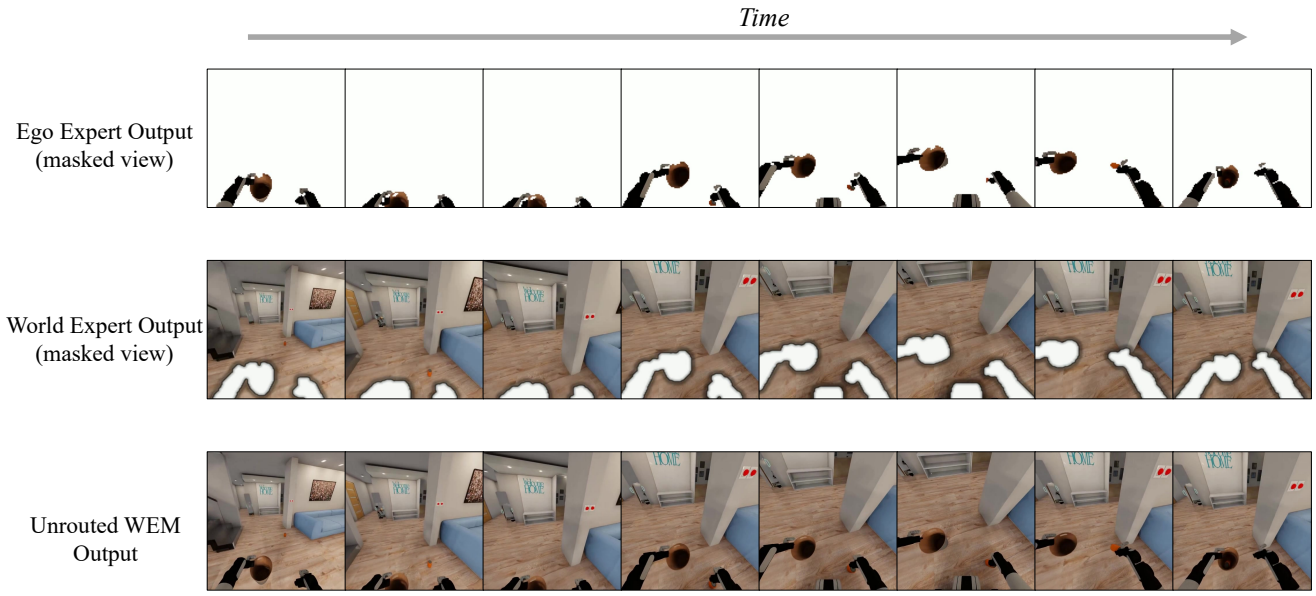


Figure 7. Mask-guided visualization of role expert specialization. We visualize the world expert and ego expert outputs in WEM under the world-ego mask $\mathbf{M}$ predicted by the semantic head. The ego expert output is shown in ego-assigned regions, while the world expert output is shown in world-assigned regions; complementary regions are filled with light gray for clarity. The last row shows the final WEM output obtained by unrouting the role-expert outputs under the same $\mathbf{M}$. This masking is used only for visualization and does not affect quantitative evaluation.

## G Limitations and Future Work

**Evaluation on simulated environments.** Our experiments are conducted on simulator-based embodied datasets. This setting enables controlled evaluation of long-horizon navigation-manipulation evolution, provides access to fine-grained annotations, and allows consistent comparison across model variants. However, simulated environments cannot fully capture the visual diversity, sensing noise, object variability, and contact dynamics of real-world robot scenarios. As a result, the sim-to-real generalization of World-Ego Modeling remains an important open question. Future work should evaluate the proposed paradigm on real-world embodied video data and real-robot trajectories, where perception errors, imperfect actuation, and more complex physical interactions may place stronger demands on the world-ego decomposition.

**Dependence on boundary construction.** World-Ego Modeling requires an operational definition of the boundary between world and ego. In this paper, we study motion-, semantic-, and intention-based views and find that the semantic view works best in our setting. The default WEM instantiation therefore constructs semantic world-ego masks from instance segmentation results. This design gives the model a clear and interpretable routing signal, but it also introduces an additional preprocessing step for datasets where instance-level annotations are not directly available. The motion-based view can avoid semantic masks by using optical flow, but it still depends on preprocessing and is less robust under large viewpoint changes or contact-rich interactions. The intention-based view removes explicit spatial annotation, but currently provides weaker separation in our experiments. A key future direction is therefore to develop stronger weakly supervised or self-supervised boundary estimation methods, allowing the model to infer world-ego structure directly from video, language, and interaction signals.

**Residual long-horizon degradation.** WEM is designed to reduce the interference between persistent scene evolution and transient robot-centric dynamics. By assigning scene-level regularities to the world state and instruction-driven interaction

dynamics to the ego state, the model improves long-horizon consistency compared with single-stream generation. Nevertheless, the current autoregressive rollout process still accumulates errors across chunks. When the generation horizon becomes very long, early mistakes can propagate into later predictions, and the consistency between distant chunks remains more difficult to maintain than local temporal coherence within a chunk. This limitation is not unique to WEM, but it indicates that world-ego disentanglement alone is not sufficient to fully solve long-horizon embodied generation. Future work may combine World-Ego Modeling with hierarchical planning, explicit memory refresh, uncertainty-aware rollout, or periodic anchoring to stable observations.

**Broader design space of World-Ego Modeling.** This work should be viewed as an initial study of the World-Ego Modeling paradigm rather than an exhaustive exploration of its design space. We define the world-ego boundary from three practical perspectives and instantiate one effective model architecture, but many alternatives remain unexplored. For example, the boundary may be defined through 3D geometry, affordances, controllability, causal interaction, or task progress rather than purely through motion, semantics, or conditioning sources. Similarly, the current CP-MoE generator is only one possible way to impose world-ego disentanglement. Future architectures may use adaptive routing, recurrent memory, structured latent states, or different levels of expert sharing to achieve a more natural separation between world and ego.

**State prediction and representation.** WEM uses a pretrained vision-language model as the state predictor, leveraging its visual understanding and language grounding to infer world and ego states from multi-turn history. While effective, this choice is not necessarily optimal for compact state extraction in video world models. A general-purpose VLM may introduce unnecessary computation and may not be specialized for preserving the state variables most relevant to future visual evolution. Future work could explore dedicated recurrent state predictors, latent-space memory modules, or lightweight video state encoders trained directly with the world-ego objective. In addition, the current query-based representation uses a fixed capacity allocation between world and ego states. Adaptive state representations, such as slot-based memory, variable-length tokens, or dynamically allocated state capacity, may better match the varying complexity of different scenes, instructions, and interaction stages.

**Applications beyond video prediction.** This paper evaluates World-Ego Modeling primarily as a video-based embodied world model. However, the underlying idea is broader: an active agent often needs to separate persistent environmental structure from self-conditioned dynamics. This distinction may be useful for downstream policy learning, planning, autonomous driving, interactive simulation, and embodied reasoning. For instance, autonomous driving models may benefit from separating road-scene regularities from ego-vehicle behavior, while robot planners may benefit from separating stable scene memory from action-conditioned interaction dynamics. Exploring these application scenarios would help determine whether World-Ego Modeling is only a useful inductive bias for video generation or a more general principle for embodied intelligence.