WORLD-EGO MODELING FOR EMBODIED VIDEO GENERATION IN LONG-HORIZON NAVIGATION-MANIPULATION TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Embodied video world models typically capture both scene evolution and the robot's behavior, which we refer to as the *world* and the *ego*, respectively. The world and the ego exhibit different underlying dynamics: world prediction relies primarily on visual history and emphasizes scene stability, whereas ego prediction relies more strongly on the current instruction and emphasizes accurate instruction following. Modeling both components within a single generation stream can entangle these different dependencies, making it difficult to specialize the prediction of either component. Consequently, it becomes difficult to simultaneously maintain scene consistency and accurate instruction following, particularly in long-horizon navigation-manipulation tasks. In this paper, we propose to decompose an embodied video into the world and the ego and disentangle their generation processes. Specifically, we define the world as the background and currently unmanipulated objects, and the ego as the robot and currently manipulated objects. Based on this definition, we develop the *World-Ego Model* (WEM), which combines a vision-language state predictor using role-conditioned attention (RCA) and asymmetric query budgets with a semantic-routed mixture-of-experts (SR-MoE) diffusion generator. To enable rigorous evaluation, we further construct HTEWorld, a dataset and benchmark for long-horizon embodied video generation with hybrid navigation-manipulation tasks, providing 125K training video clips comprising over 4.5M frames with fine-grained instructions, together with 300 multi-turn evaluation trajectories covering over 2K instructions. Extensive experiments show that WEM achieves state-of-the-art performance on HTEWorld while remaining competitive on existing manipulation-oriented evaluations.

1 INTRODUCTION

World models are essential to embodied AI, as they learn physical dynamics to predict future consequences, generate synthetic data, and serve as policy simulators (Ha & Schmidhuber, 2018; Hafner et al., 2025; Jang et al., 2025; Kim et al., 2026). Recent video-based world models have demonstrated strong capabilities in generating realistic future rollouts and are playing an increasingly important role in embodied intelligence (Agarwal et al., 2025; Zhou et al., 2024; Xiang et al., 2025).

An embodied video can be viewed as comprising two components: the *world*, which captures the evolution of the environment, and the *ego*, which captures the behavior of the embodiment and its interactions with objects. Existing embodied world models generally do not explicitly distinguish the world from the ego. Instead, they use a single generation stream to jointly model scene evolution, robot behavior, and object interactions. They have shown strong performance in short-horizon and manipulation-oriented prediction (Zhou et al., 2024; Zhen et al., 2025; Shang et al., 2026). However, when applied to more complex tasks, such as long-horizon hybrid navigation-manipulation involving interleaved navigation and manipulation stages, they struggle to maintain scene consistency (Shang et al., 2025a) or follow instructions (Deng et al., 2026) accurately, as shown in Fig. 1.

We attribute this challenge partly to mixing the world and the ego within a single generation stream. The world and the ego exhibit different underlying dynamics, i.e., world evolution relies primarily on historical context, whereas ego evolution relies more strongly on the current instruction (Wan et al., 2026). Modeling both within a single generation stream entangles these different dependencies and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

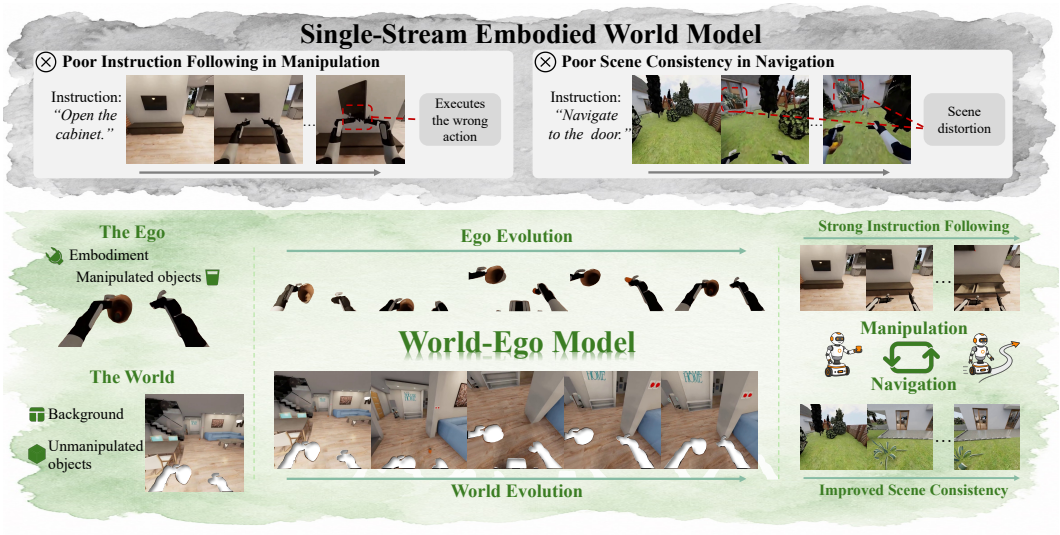


Figure 1: **Motivation of WEM.** Embodied world models usually generate video in a single stream, leading to difficulties in accurate instruction following and long-horizon scene consistency. We define the world as the background and currently unmanipulated objects, and the ego as the robot and currently manipulated objects. Further, we separately model the evolution of the world and ego, thereby improving video generation performance in long-horizon navigation-manipulation tasks.

hinders specialization for either component. This motivates us to explicitly separate the conditioning information and generation parameters associated with the world and the ego.

As illustrated in Fig. 1, we explicitly decompose an embodied video into world and ego components, defining the world as the background and currently unmanipulated objects and the ego as the robot and currently manipulated objects. Based on this decomposition, we introduce the **World-Ego Model** (**WEM**), which separates world and ego information for prediction and introduces specialized experts for generation. WEM has prediction and generation stages. In the prediction stage, a VLM-based state predictor (Bai et al., 2025) extracts world and ego states from the multimodal context using role-conditioned attention (RCA) and asymmetric query budgets. In the generation stage, we introduce a Semantic-Routed Mixture-of-Experts (SR-MoE) generator with a shared Preceding Expert and specialized World/Ego Experts. The Preceding Expert first processes all video tokens, after which a Semantic Head predicts their world-ego assignments. The assigned world and ego tokens are then routed to their corresponding specialized experts, whose outputs are merged to produce the next video chunk. Combined with chunk-wise autoregressive diffusion (Teng et al., 2025), WEM supports long-horizon multi-turn generation for hybrid navigation-manipulation tasks.

Our setting requires evaluating both navigation-oriented scene imagination and manipulation-oriented physical simulation under continuous multi-turn embodied instructions, but none of the recent benchmarks (Huang et al., 2025; Li et al., 2025; Deng et al., 2026; Shang et al., 2026) meet our requirements. We therefore construct **Hybrid-Task Embodied World Benchmark** (**HTEWorld**) on top of BEHAVIOR-1K (Li et al., 2023). HTEWorld provides 125K training video clips containing over 4.5M frames with fine-grained instructions, together with 300 evaluation trajectories containing over 2K instructions and interleaved navigation-manipulation stages. WEM achieves SOTA performance on HTEWorld and remains competitive under existing manipulation-oriented tasks (Shang et al., 2026; Khazatsky et al., 2024). Our contributions are summarized as follows:

- We show that an embodied video can be decomposed into world (i.e., background and currently unmanipulated objects) and ego (i.e., the robot and currently manipulated objects) components, providing an effective formulation for embodied video world models.
- We develop WEM to instantiate world-ego decomposition with a world-ego state predictor and an SR-MoE video generator. We demonstrate its effectiveness for long-horizon, multi-turn generation in hybrid navigation-manipulation tasks.
- We establish HTEWorld, which provides 125K training clips and 300 evaluation trajectories, as a dataset and benchmark for long-horizon embodied video generation in hybrid

navigation-manipulation tasks. Experiments show that WEM achieves SOTA performance on HTEWorld while remaining competitive on existing manipulation-oriented evaluations.

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 (Ha & Schmidhuber, 2018; Hafner et al., 2020; 2025; Wu et al., 2023). Early methods learn compact latent dynamics for latent imagination. With diffusion models (Ho et al., 2020; Rombach et al., 2022) and video generation (Liu et al., 2024; Wan et al., 2025; Yang et al., 2025; Parker-Holder et al., 2024; Zheng et al., 2024; Huang et al., 2026b; He et al., 2025; Kong et al., 2024; Wang et al., 2024; Jin et al., 2025; Jang et al., 2026; Wu et al., 2026), future prediction has moved from low-dimensional state transitions to pixel-level visual rollouts. Cosmos-Predict (Ali et al., 2025) 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 (Bruce et al., 2024) learns a latent action model from unlabeled videos; The Matrix (Feng et al., 2024), Yume (Mao et al., 2025), and LIVE (Huang et al., 2026a) target real-time controllable interaction, open-ended scene exploration, and long-horizon consistency modeling; PAN (Xiang et al., 2025) proposes a generative latent prediction framework for general, interactable, and long-horizon world simulation conditioned on history and language actions.

Embodied Video World Models. Embodied world models predict both scene evolution and how agent behaviors change the physical world. Existing methods predict future robot observations via video generation (Yang et al., 2024; Zhou et al., 2024; Zhen et al., 2025; Shang et al., 2025b) or action-conditioned rollouts (Walke et al., 2023). Such predictions support downstream uses, including data generation (Chen et al., 2025b), policy learning (Bi et al., 2025; Wen et al., 2024), evaluation (Shang et al., 2026), control (Kim et al., 2026), and video-level planning (Chen et al., 2025a). WoW (Chi et al., 2025) learns a generative world model from real robot trajectories and uses inverse dynamics to translate imagined outcomes into executable actions. Ctrl-World (Guo et al., 2026) 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 embodied video world models use a single generation stream to jointly model scene evolution, robot behavior, and object interactions, despite their different conditioning dependencies. Related works have explored structured modeling of visual dynamics from different perspectives. Iso-Dream (Pan et al., 2022) separates controllable and noncontrollable dynamics for model-based reinforcement learning, while GEM (Hassan et al., 2025) introduces explicit controls for ego motion and object dynamics in egocentric video generation. In contrast, we decompose embodied videos into world and ego components according to their current interaction roles and separately model their evolution for long-horizon hybrid navigation-manipulation generation.

3 Method

An embodied video world model predicts a future video $\mathbf{V}$ from the available context $\mathcal{H}$ and the current control signal a according to $\mathbf{V} \sim p_\theta(\mathbf{V} \mid \mathcal{H}, a)$, where $\mathcal{H}$ may contain the initial observation, previously generated videos, and past control signals, while a denotes the current control signal, such as a text instruction, camera trajectory, or robot action.

3.1 World-Ego Decomposition

Existing embodied video world models typically use a single generation stream. However, different parts of an embodied video follow different conditioning dependencies. The background and objects not involved in the current interaction are primarily constrained by the observed history and should preserve scene structure over time. In contrast, the robot and currently manipulated objects are driven by the current control signal and must realize the intended behavior. Asking a single predictive mapping to model these two dependencies makes it difficult to specialize for either one.

We therefore decompose an embodied video into a **world** component and an **ego** component. Specifically, the world consists of the background and currently unmanipulated objects, while the ego consists of the robot and currently manipulated objects; conceptually, $\mathbf{V} = \mathcal{C}(\mathbf{V}^w, \mathbf{V}^e)$, where $\mathcal{C}$ denotes their composition. WEM realizes this decomposition at two complementary levels. The **world-ego state predictor** separates the conditioning information required by the two roles, producing world and ego states with different contextual dependencies (Sec. 3.2); the **SR-MoE video**

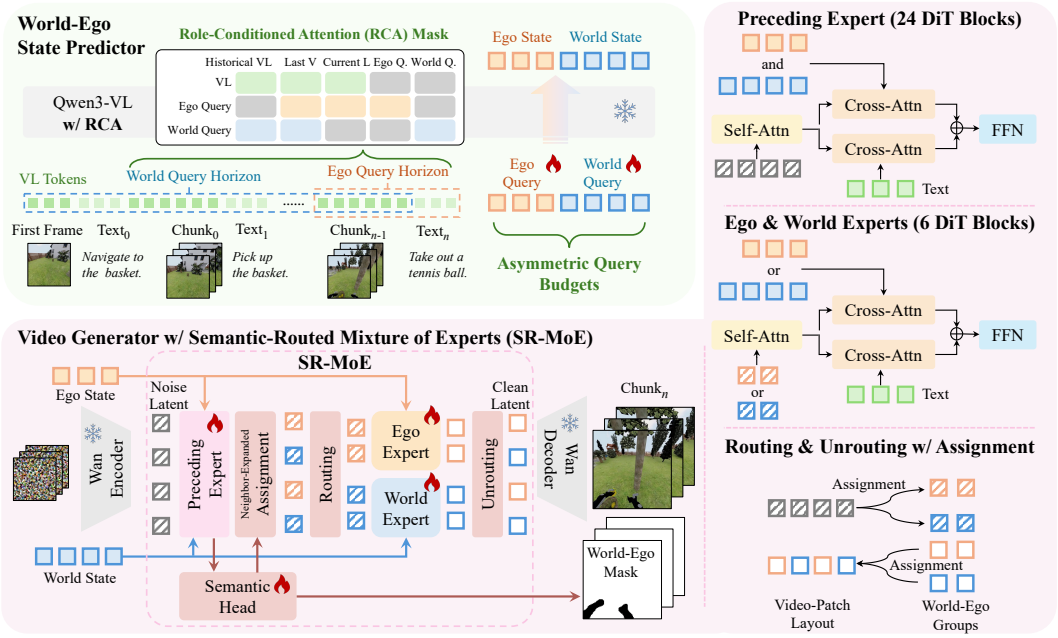


Figure 2: **Overview of WEM architecture**. A VLM-based state predictor uses asymmetric world and ego query groups with RCA to extract world and ego states from the vision-language inputs. The SR-MoE generator consists of a shared Preceding Expert and specialized World and Ego Experts. The Preceding Expert processes the noisy video latent tokens, and a Semantic Head predicts their world-ego assignment. The assigned tokens are routed to the corresponding experts, whose outputs are restored to their original positions.

generator retains shared video processing while introducing specialized generation parameters for world and ego regions (Sec. 3.3). As illustrated in Fig. 2, these two stages are applied autoregressively to generate a sequence of video chunks for long-horizon generation.

3.2 World-Ego State Predictor

The world and ego require different information from the generation history: the world benefits from long-term context to maintain persistent scene content across rounds, whereas the ego depends on the most recent visual observation and the current instruction to perform the interaction at the current round. To accommodate these different information requirements, we extract two role-specific feature groups from the same multimodal context, one for the world and one for the ego.

For the k-th generation round, we construct the multimodal context from the initial frame $\mathbf{O}_0$, previous instruction-video pairs $(a_1, \mathbf{V}_1), \ldots, (a_{k-1}, \mathbf{V}_{k-1})$, and the current instruction a_k. These inputs are arranged chronologically and encoded into a vision-language token sequence $\mathbf{X}_k \in \mathbb{R}^{L_k \times d_v}$, where L_k is the sequence length and d_v is the VLM hidden dimension. To extract separate world and ego states from this shared sequence, we append two groups of learnable queries, $\mathbf{Q}^w \in \mathbb{R}^{N_w \times d_v}$ and $\mathbf{Q}^e \in \mathbb{R}^{N_e \times d_v}$, where N_w and N_e denote the numbers of world and ego queries, respectively. A pretrained VLM Φ_ϕ (Bai et al., 2025) processes $\mathbf{X}_k$ together with these queries and produces

$$(\mathbf{S}_k^w, \mathbf{S}_k^e) = \Phi_\phi\left(\mathbf{X}_k; \mathbf{Q}^w, \mathbf{Q}^e\right). \tag{1}$$

Here, $\mathbf{S}_k^w \in \mathbb{R}^{N_w \times d_v}$ and $\mathbf{S}_k^e \in \mathbb{R}^{N_e \times d_v}$ are the VLM outputs at the world and ego query positions, which we refer to as the world and ego states, respectively.

Role-Conditioned Attention. Using different query tokens alone does not guarantee different information: if both query groups attend to the same context, they may capture largely overlapping information. We therefore explicitly control the context accessible to each query group through Role-Conditioned Attention (RCA). Specifically, both query groups extract information from the generation history, but with different attention scopes. The world queries attend to the complete historical vision-language context and to one another, while being blocked from the current instruction and the ego queries. This biases the world state toward stable scene information, rather than letting it directly depend on the current instruction. In contrast, the ego queries attend to the most recent

video chunk, the current instruction, and one another, while being blocked from earlier history and the world queries. This focuses the ego state on the interaction to be executed and reduces interference from stale information. As a result, RCA encourages $\mathbf{S}_k^w$ and $\mathbf{S}_k^e$ to capture complementary information for world and ego generation while retaining a shared VLM backbone. As illustrated in Fig. 2, RCA is implemented by reconstructing the VLM attention mask.

Asymmetric Query Budgets. The two roles also differ in the amount of context they must summarize. Since the world state aggregates a substantially longer and more diverse history, we allocate more queries to the world than to the ego. In practice, this is implemented by setting different numbers of learnable queries, N_w and N_e, which provides a flexible way to adjust the capacity allocated to the two states. This provides the world state with greater representation capacity while keeping the ego state compact and focused on the current interaction.

3.3 SR-MoE Video Generator

Separating the conditioning states alone does not fully resolve world-ego entanglement if both regions are still modeled with the same generation parameters. However, completely separating the two generation streams would sacrifice the shared spatial and temporal structure required for coherent video synthesis. We therefore combine shared processing for common video structure with specialized processing for the world and ego. We restructure a pretrained video DiT (Peebles & Xie, 2023) into a Semantic-Routed Mixture-of-Experts (SR-MoE). The first 24 of the 30 DiT blocks form a shared Preceding Expert, while the final six are duplicated to initialize independently optimized World and Ego Experts. The shared blocks preserve global spatiotemporal interaction across the entire video, whereas the specialized blocks process world and ego tokens with separate parameters.

Preceding Expert. Let $\mathbf{Z}_k \in \mathbb{R}^{T\times N_s\times d_g}$ denote the noisy video latent tokens for generation round k, where T is the number of frames, N_s is the number of spatial tokens per frame, and d_g is the token dimension. For simplicity, we omit the diffusion timestep from the notation. Before semantic routing, all tokens in $\mathbf{Z}_k$ are jointly processed by the Preceding Expert. This shared processing is important because the world and ego are not independent visual layers: their appearance, geometry, occlusion, and motion must remain coherent within the same video. Each block additionally includes a state cross-attention branch conditioned on the concatenated world and ego states $[\mathbf{S}_k^w; \mathbf{S}_k^e]$, allowing the Preceding Expert to incorporate information from both states during shared processing. We denote the output of its ℓ-th block as $\mathbf{Z}_k^{(\ell)}$, and its final output as $\mathbf{Z}_k^P \in \mathbb{R}^{T\times N_s\times d_g}$.

Neighbor-Expanded Assignment. After shared processing, the tokens need to be assigned to the World or Ego Expert for specialized processing. Since the robot and manipulated objects move across frames, we dynamically predict the world-ego assignment using selected intermediate outputs of the Preceding Expert and the ego state. Specifically, the Semantic Head is implemented as a dense prediction transformer (Ranftl et al., 2021) that takes the intermediate outputs $\{\mathbf{Z}_k^{(\ell)}\}_{\ell\in\mathcal{J}}$ from a selected set of blocks $\mathcal{J}$ together with the ego state $\mathbf{S}_k^e$ and predicts

$$\mathbf{M}_k = \sigma\left(G_{\text{sem}}\left(\{\mathbf{Z}_k^{(\ell)}\}_{\ell\in\mathcal{J}}, \mathbf{S}_k^e\right)\right) \in [0,1]^{T\times N_s}, \tag{2}$$

where G_{sem} denotes the Semantic Head and σ denotes the sigmoid function. The resulting $\mathbf{M}_k$ is a frame-wise ego probability mask aligned with the video, with each element $M_{k,t,i}$ indicating the probability that spatial token i in latent frame t belongs to the ego region. Given a threshold τ, we define the world and ego index sets as $\mathcal{I}_k^w = \{(t,i) \mid M_{k,t,i} < \tau\}$ and $\mathcal{I}_k^e = \{(t,i) \mid M_{k,t,i} \geq \tau\}$.

A naive hard assignment scheme would send each token exclusively to one expert. This creates two related problems near the boundary separating world and ego tokens. Spatially, each expert loses access to neighboring tokens assigned to the other role, even though their visual content must remain locally coherent. Temporally, as the robot or manipulated objects move, the role assigned to a given spatial location can change across adjacent frames, causing that location to switch abruptly between independently optimized experts and potentially introducing temporal discontinuities such as flickering near the boundary. We therefore introduce Neighbor-Expanded Assignment. Rather than treating the predicted assignment as a strict separation between the two experts, we create a narrow overlapping region around the boundary. For each role $r \in \{w, e\}$, we expand its routing index set as $\widetilde{\mathcal{I}}_k^r = \mathcal{I}_k^r \cup \mathcal{N}(\mathcal{I}_k^r)$, where $\mathcal{N}(\mathcal{I}_k^r)$ denotes the indices of spatial neighbors within the same frame for positions indexed by $\mathcal{I}_k^r$. The expansion allows tokens close to the boundary to be seen by both experts, while tokens sufficiently far from the boundary remain role-specific. It therefore provides cross-boundary spatial context and reduces abrupt changes in token routing across adjacent frames by allowing tokens near the boundary to be processed by both experts.

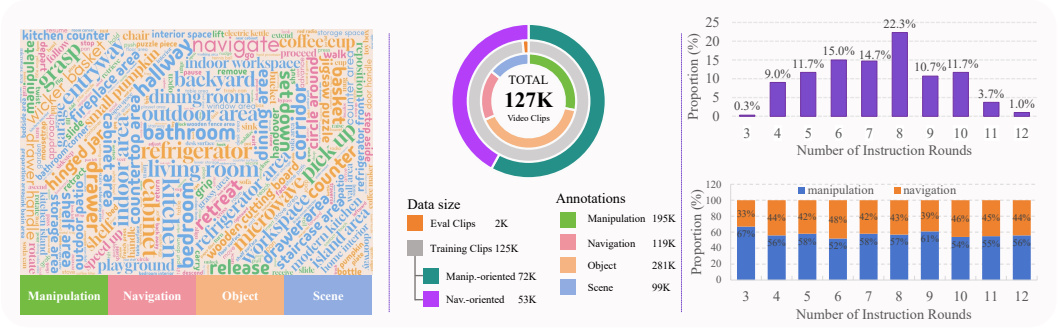


Figure 3: **Statistics of HTEWorld.** HTEWorld contains large-scale training clips and multi-turn evaluation trajectories for hybrid navigation-manipulation generation. Left: word-frequency visualization of manipulation, navigation, object, and scene annotations. Middle: dataset composition, including the training/evaluation split, manipulation/navigation training clips, and counts of actions, objects, and scenes mentioned in the instructions. Right: evaluation-trajectory statistics, showing the distribution of instruction rounds and manipulation/navigation proportions across trajectory lengths.

Routing, Role Experts, and Unrouting. The expanded token groups are gathered from the Preceding Expert output and routed to their corresponding experts:

$$\mathbf{Z}_k^r = \mathrm{Gather}\left(\mathbf{Z}_k^P, \widetilde{\mathcal{I}}_k^r\right), \qquad \widehat{\mathbf{Z}}_k^r = E_{\theta_r}^r\left(\mathbf{Z}_k^r, a_k, \mathbf{S}_k^r\right), \qquad r \in \{w, e\}. \tag{3}$$

Here, $\mathbf{Z}_k^r, \widehat{\mathbf{Z}}_k^r \in \mathbb{R}^{|\widetilde{\mathcal{I}}_k^r| \times d_g}$ denote the routed tokens before and after expert processing, respectively, and $E_{\theta_w}^w$ and $E_{\theta_e}^e$ denote the World and Ego Experts. Each expert contains a state cross-attention branch conditioned on its corresponding state $\mathbf{S}_k^r$. After expert processing, the expert outputs are restored to their original spatiotemporal positions. For positions covered by only one expanded routing index set, the corresponding expert output is used directly; for positions shared by both index sets, the two outputs are averaged. We denote this operation as

$$\widehat{\mathbf{Z}}_k = \mathrm{Restore}\left(\widehat{\mathbf{Z}}_k^w, \widetilde{\mathcal{I}}_k^w; \widehat{\mathbf{Z}}_k^e, \widetilde{\mathcal{I}}_k^e\right) \in \mathbb{R}^{T \times N_s \times d_g}. \tag{4}$$

This design preserves shared processing for global video coherence while allowing the World and Ego Experts to specialize according to the predicted world-ego assignment.

Training. WEM is trained end-to-end with a flow-matching objective (Lipman et al., 2023) for video generation and an auxiliary mask-prediction objective for the Semantic Head. The latter combines class-balanced binary cross-entropy and Dice losses (Milletari et al., 2016). The overall objective is $\mathcal{L} = \mathcal{L}_{\text{flow}} + \lambda\left(\mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{Dice}}\right)$, where λ weights the mask-prediction loss.

4 HTEWorld

Existing embodied video generation benchmarks mainly focus on short video clips or individual instructions (Huang et al., 2024; 2025; Zheng et al., 2025; Li et al., 2025; Deng et al., 2026; Shang et al., 2026). They do not adequately evaluate long-horizon generation involving interleaved navigation and manipulation. We therefore construct HTEWorld from the teleoperated demonstrations released for the 2025 BEHAVIOR Challenge in the BEHAVIOR-1K environment (Li et al., 2023). The dataset provides household-task demonstrations collected with the Galaxea R1 Pro robot and multimodal observations. We select ten representative tasks and use the head-camera RGB observations from the demonstrations to construct HTEWorld. For each task, five episodes are held out for evaluation, while the remaining episodes are used for training.

The videos are segmented into 37-frame clips at 16 FPS and a resolution of 480×480. HTEWorld contains approximately 127K video clips, including 125K training clips and 2K evaluation clips. The evaluation set contains 300 multi-turn trajectories with interleaved navigation and manipulation stages. Fig. 3 summarizes the statistics of HTEWorld. Details of the data construction and annotation pipeline are provided in Appendix A. We adopt the video quality evaluation protocol introduced by WorldArena (Shang et al., 2026), which comprises 16 metrics across six dimensions: visual quality, motion quality, content consistency, physics adherence, 3D accuracy, and controllability. We report the aggregated EWMScore as the primary metric. We additionally introduce six HTEWorld-specific metrics for multi-turn consistency and navigation-manipulation transitions, detailed in Appendix D.

Table 1: **Comparison on HTEWorld** using the WorldArena metrics and EWMScore. Metric abbreviations follow WorldArena. 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 (Chi et al., 2025)	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 (Ali et al., 2025)	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 (Ali et al., 2025)	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 (Wan2.1) (Xiang et al., 2025)	56.58	64.87	49.01	1.20	39.65	44.00	77.01	82.99	75.95	29.87	79.64	26.26	95.23	86.22	3.12	69.14	81.12
PAN (Wan2.2) (Xiang et al., 2025)	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 (Wan2.1)	60.11	65.02	49.01	1.92	38.95	47.40	82.61	**88.19**	81.95	34.22	83.69	31.91	95.33	88.49	4.02	**87.72**	81.32
WEM (Wan2.2)	**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**

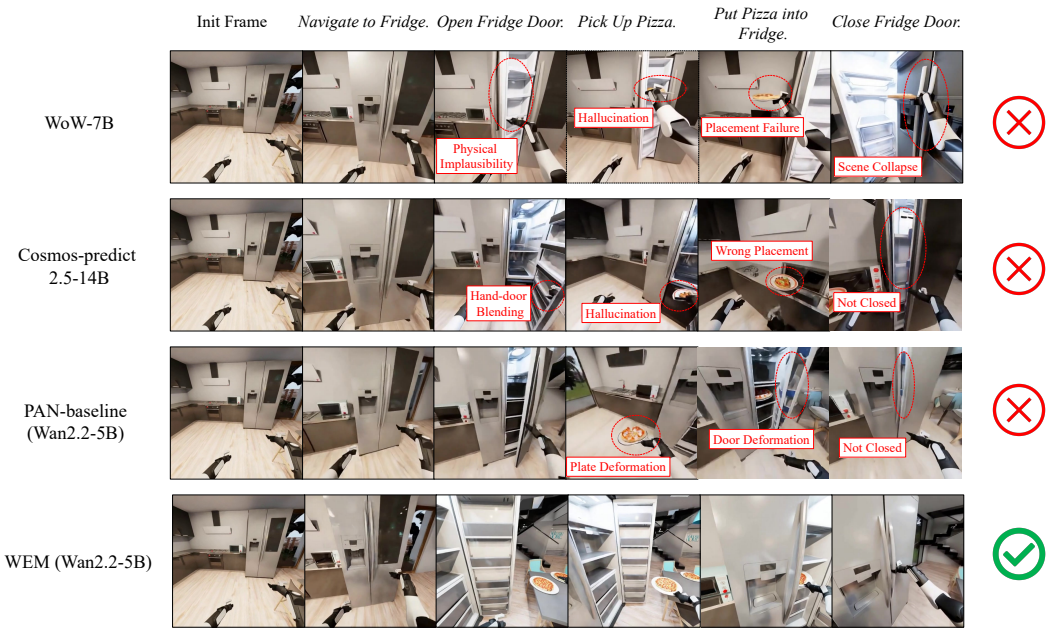


Figure 4: **Qualitative comparison on HTEWorld.** Given the same initial observation and five-step instruction sequence, each model autoregressively generates a long-horizon multi-turn navigation-manipulation rollout. Red circles indicate scene inconsistency or instruction-following failures. WEM can better maintain scene consistency and instruction following throughout the rollout.

5 Experiments

5.1 Experimental Setup

We build the state predictor on Qwen3-VL-2B-Instruct (Bai et al., 2025). We empirically set the query budgets to $N_w = 192$ for the world and $N_e = 64$ for the ego. The SR-MoE video generator is initialized from either Wan2.1-T2V-1.3B or Wan2.2-TI2V-5B (Wan et al., 2025), resulting in WEM (Wan2.1) and WEM (Wan2.2), respectively. We freeze the Qwen3-VL backbone and jointly optimize all remaining WEM parameters. We set the routing threshold τ to 0.5 and anneal the mask-loss weight λ from 0.3 to 0.06 over training, emphasizing semantic-assignment learning early and video generation later. WEM is trained for four epochs with a learning rate of 1×10^{-5}. Both variants are evaluated on HTEWorld, while all other experiments use WEM (Wan2.2).

We compare WEM with Cosmos-Predict2.5-2B and Cosmos-Predict2.5-14B (Ali et al., 2025), WoW-7B (Chi et al., 2025), and two PAN-style baselines (Xiang et al., 2025). PAN (Wan2.1) and PAN (Wan2.2) use the same Qwen3-VL and Wan initializations as their corresponding WEM variants, but predict a unified state and retain a single generation stream. Unless otherwise specified, all methods are trained on the same data and evaluated under the same protocol for each benchmark. Complete implementation, training, and evaluation details are provided in Appendix B.

5.2 Main Results

Evaluation on HTEWorld. As shown in Table 1, WEM performs strongly on HTEWorld for long-horizon multi-turn generation in hybrid navigation-manipulation tasks. WEM (Wan2.2) and WEM (Wan2.1) achieve the highest and second-highest EWMScores of 61.48 and 60.11, respectively.

Table 2: **Comparison on WorldArena.** The condition denotes the control signal provided in addition to the visual context.

Model	Condition	EWMScore ↑
Wan2.2 (Wan et al., 2025)	Text	54.54
WoW (Chi et al., 2025)	Text	54.88
CogVideoX (Yang et al., 2025)	Text	57.90
WEM (Wan2.2)	Text	58.10
Cosmos-Predict 2.5 (Ali et al., 2025)	Action	55.91
IRASim (Zhu et al., 2025)	Action	58.12
Ctrl-World (Guo et al., 2026)	Action	**59.70**

Table 3: **Comparison on DROID.** PAN and WEM use the Wan2.2 backbone.

Model	EWMScore ↑
Wan2.2 (Wan et al., 2025)	46.44
PAN (Xiang et al., 2025)	50.83
WEM	**53.70**

Table 4: **Ablation of state prediction and generation designs in WEM**, using PAN as the baseline and mask supervision as a control. Results are reported on HTEWorld. All variants use Wan2.2 (Wan et al., 2025) as the video-generation backbone.

Model	State Predictor			Video Generator		Training	EWMScore ↑
	None	Unified	World-Ego	Single Stream	SR-MoE	Mask Supervision	
PAN (Xiang et al., 2025)		✓		✓			58.40
PAN + Mask Supervision		✓		✓		✓	58.61
PAN w/o State Predictor	✓			✓			53.49
WEM w/o State Predictor	✓				✓	✓	57.71
PAN + World-Ego States			✓	✓			58.69
WEM			✓		✓	✓	**61.48**

More importantly, the gains concentrate on the two capabilities that motivate world-ego decomposition: scene consistency and instruction following. Compared with PAN (Wan2.2), WEM (Wan2.2) improves background consistency (BC) by 7.68 points and subject consistency (SC) by 7.28 points, showing substantially more stable scene evolution. It also improves semantic alignment (Sem) by 3.07 points, instruction following (Inst) by 1.60 points, and interaction quality (Inter) by 4.47 points, indicating better adherence to the instructed behavior. The qualitative results in Fig. 4 show the same pattern: while the compared models gradually suffer from scene distortion or fail to follow later instructions, WEM better preserves the scene and completes the full instruction sequence. These improvements are consistent across different video generation backbones: WEM outperforms its PAN counterpart by 3.53 points with Wan2.1 and by 3.08 points with Wan2.2, demonstrating the effectiveness of world-ego decomposition across different video-generation backbones. Results under the HTEWorld-specific metrics are reported in Appendix D.3.

Evaluation on WorldArena. We further evaluate WEM on WorldArena (Shang et al., 2026), which focuses on short-horizon tabletop manipulation. As shown in Table 2, WEM achieves an EWMScore of 58.10, the highest score among the compared text-conditioned methods. Using text instructions alone, it also performs on par with methods using stronger action-based control signals. These results show that world-ego decomposition remains effective for manipulation-only tasks, where separating the persistent scene from robot-object interactions is still beneficial.

Evaluation on DROID. We evaluate WEM on DROID (Khazatsky et al., 2024), a large-scale real-world dataset covering diverse scenes and robot manipulation tasks. As shown in Table 3, WEM (Wan2.2) achieves an EWMScore of 53.70, outperforming PAN (Wan2.2) by 2.87 points and the Wan2.2 baseline by 7.26 points. These results extend our findings from simulated environments to real-world robot videos, showing that world-ego decomposition remains effective under the greater visual diversity and less controlled conditions of real-world data.

5.3 Ablation Studies

Table 4 shows our ablation studies on state prediction, video generation, and training strategy.

Effect of world-ego state predictor. Comparing rows 3 and 1 of Table 4, introducing a unified state predictor improves EWMScore from 53.49 to 58.40 under the same single-stream generator without mask supervision. This substantial improvement shows the benefit of explicitly predicting a state to condition future generation. Comparing rows 1 and 5, replacing the unified state with separate world and ego states further improves the score to 58.69, with the generator and training scheme unchanged. This controlled comparison shows that separating the conditioning information required by the world and ego roles provides an additional benefit beyond state prediction itself.

Table 5: **Ablation of SR-MoE design choices on HTEWorld.** Token Separation indicates whether world and ego tokens are separated according to the predicted assignment before subsequent expert processing. Expert Specialization indicates whether separate World and Ego Experts are used. Assignment Proxy specifies whether the world-ego assignment is derived from semantic regions or motion induced by the robot and manipulated objects. The shaded row denotes the configuration adopted by WEM. All models use Wan2.2.

Token Separation	Expert Specialization	Assignment Proxy	EWMScore ↑
✓		Motion	58.77
	✓	Motion	59.36
✓	✓	Motion	57.04
✓		Semantic	58.85
	✓	Semantic	61.09
✓	✓	Semantic	**61.48**

We further examine the state predictor's Role-Conditioned Attention and asymmetric query budget using full WEM with the Wan2.2 video-generation backbone (61.48 EWMScore). Removing Role-Conditioned Attention and allowing both query groups to access the complete multimodal context reduces the score to 60.64. Replacing the 192 world and 64 ego queries with 128 queries for each role reduces the score to 60.50. These results support allocating different capacities and attention scopes to the two states, consistent with their different contextual dependencies described in Sec. 3.2.

Effect of SR-MoE. Comparing rows 3 and 4 of Table 4, replacing the single-stream generator with SR-MoE improves EWMScore from 53.49 to 57.71 without the state predictor. Comparing rows 5 and 6, the same replacement improves the score from 58.69 to 61.48 with world-ego states. Since SR-MoE also introduces mask supervision, we examine this factor separately by comparing rows 1 and 2: adding the same auxiliary supervision to PAN improves its score only from 58.40 to 58.61. This smaller gain suggests that auxiliary mask supervision alone does not explain the improvements associated with SR-MoE, supporting the benefit of role-specific generation.

We further examine the key design choices of SR-MoE in Table 5. Under semantic assignment, the full SR-MoE achieves an EWMScore of 61.48 (row 6). Removing token separation reduces the score to 61.09 (row 5), while removing expert specialization decreases it to 58.85 (row 4). These results show that both token separation and expert specialization contribute to the effectiveness of SR-MoE. To validate the semantic assignment used in SR-MoE, we compare it with a motion-based alternative that identifies ego regions from motion induced by the robot and manipulated objects, as illustrated in Fig. 5. Comparing rows 1 and 4, 2 and 5, and 3 and 6, semantic assignment improves EWMScore by 0.08, 1.73, and 4.44 points, respectively. This advantage can be attributed to semantic masks being directly tied to visual appearance and therefore easier to predict reliably from video latents, as shown in Fig. 7, where the predicted assignments remain accurate across the generated sequence. The SR-MoE variants are illustrated in Fig. 6, with implementation details provided in Appendix C.

We additionally examine Neighbor-Expanded Assignment within SR-MoE using the Wan2.2 video-generation backbone. Removing it reduces EWMScore from 61.48 to 59.57 and produces visible discontinuities in the generated videos. This result supports providing overlapping context near the boundary between world- and ego-assigned regions, consistent with the motivation in Sec. 3.3.

6 Conclusion

This work studies embodied video generation by decomposing a video into world and ego components. We define the world as the background and currently unmanipulated objects, and the ego as the robot and currently manipulated objects. Based on this formulation, we develop WEM, which separates their conditioning information with a VLM-based state predictor and separates their generation parameters with an SR-MoE generator. We construct HTEWorld for long-horizon multi-turn generation with interleaved navigation and manipulation. WEM achieves state-of-the-art performance on HTEWorld while remaining competitive on existing manipulation-oriented evaluations.

Limitations. DROID allows us to evaluate WEM in a real-world tabletop manipulation setting, but the lack of real-world datasets with interleaved navigation and manipulation limits the evaluation of long-horizon hybrid generation in real environments. Moreover, although WEM improves long-horizon generation, errors may still accumulate and earlier context may be lost as generation extends over more instruction rounds or longer durations. A detailed discussion is provided in Appendix F.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing of the manuscript, including improving wording, grammar, and clarity. We also used Gemini 2.5 Flash to assist in generating annotations for the HTEWorld dataset. All AI-assisted outputs were reviewed and verified by the authors before use. We did not use generative AI tools for literature retrieval or discovery, research ideation or execution, drafting substantive sections of the paper, or proving mathematical claims. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

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.

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.

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.

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.

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.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *ICLR*, 2026.

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*, 2025a.

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*, 2025b.

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.

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.

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.

Gemini Team, Google. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025.

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

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

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

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

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.

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.

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

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*, 2026a.

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

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.

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.

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.

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.

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.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, et al. DROID: A large-scale in-the-wild robot manipulation dataset. In *RSS*, 2024.

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.

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.

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.

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.

Gen Li, Bo Zhao, Jianfei Yang, and Laura Sevilla-Lara. Mask2iv: Interaction-centric video generation via mask trajectories. In *AAAI*, 2026.

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

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.

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.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 2016.

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

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.

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

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.

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

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

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*, 2025a.

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

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.

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

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.

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.

Chi Wan, Kangrui Wang, Yuan Si, Pingyue Zhang, and Manling Li. WorldAgen: Unified state-action prediction with test-time world model training. In *AAAI*, pp. 18584–18592, 2026.

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.

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.

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.

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

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.

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.

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.

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.

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.

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

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.

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.

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.

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.

Table 6: Summary of HTEWorld.

Item	Description
Source environment	BEHAVIOR-1K (Li et al., 2023)
Embodiment	Wheeled bimanual Galaxea R1 Pro
Task coverage	Ten household tasks
Training videos	125K clips; over 4.5M frames
Clip format	37 frames; 480×480; 16 FPS
Evaluation set	300 trajectories; over 2K instructions
Annotations	Language instructions, semantic masks, optical flow

A Data Construction and Annotation

A.1 HTEWorld Data Construction

HTEWorld is constructed from the teleoperated demonstrations released for the 2025 BEHAVIOR Challenge in the BEHAVIOR-1K environment (Li et al., 2023), targeting embodied video generation with interleaved navigation and manipulation. We select ten household tasks: turning on radio; picking up trash; putting away Halloween decorations; cleaning up plates and food; can meat; setting mousetraps; hiding Easter eggs; picking up toys; rearranging kitchen furniture; and setting up a coffee station in the kitchen. These tasks cover diverse navigation and manipulation behaviors, including approaching target locations, opening containers, grasping and carrying objects, and placing objects at designated locations. A representative hybrid sequence is *navigate to a trash bin* $\rightarrow$ *pick up the bin* $\rightarrow$ *navigate to the trash* $\rightarrow$ *place the trash in the bin*.

The trajectories are recorded from the egocentric head camera of the Galaxea R1 Pro robot. For each task, five episodes are held out for evaluation and excluded from training. We divide the remaining trajectories into fixed-length video clips while preserving the parent action and the temporal position of each clip within that action. Since a single action may span multiple clips, the language annotation describes the behavior visible in each clip rather than directly reusing its parent action label.

Each clip contains 37 frames at 16 FPS and a resolution of 480×480, and incomplete final segments are discarded. The resulting training set contains 125K clips with over 4.5M frames. From the held-out episodes, we select 300 representative multi-turn trajectories with over 2K instructions for evaluation, preserving the ordered navigation and manipulation stages for continuous generation. Table 6 summarizes the resulting data.

A.2 Language Instruction Annotation

We use Gemini 2.5 Flash (Gemini Team, Google, 2025) to generate one action-centric language instruction for each video clip. The annotation describes the robot behavior and object interaction visible in the current clip rather than directly copying the corresponding task or action label.

The annotation prompt provides the robot embodiment and egocentric viewpoint, the ordered action context of the episode, the current action, and the temporal position of the clip within that action. The Galaxea R1 is described as a wheeled bimanual robot observed from its head camera, with its left and right arms identified by their locations in the image. The action context helps identify objects involved in successive interactions, while the temporal position helps distinguish different stages of a long action, such as approaching an object, manipulating it, and retracting the gripper.

The output is constrained to a single sentence of at most 30 words. It describes observable physical behavior, uses egocentric directions when appropriate, and specifies the acting arm when visible. The model is instructed not to copy the original action label verbatim, infer unobserved future actions, or refer to the recording or annotation process.

Before prompt construction, corrupted action labels are sanitized by removing trailing character sequences introduced by truncated labels. Failed annotation requests are retried, and completed annotations are saved incrementally. The same language annotations are used across all compared world-ego definitions and generator variants.

A.3 Semantic World-Ego Mask Annotation

The Semantic Head is trained using binary world-ego masks. Following the semantic definition in the main text, the robot and currently manipulated objects are assigned to the ego, while the background and currently unmanipulated objects are assigned to the world. These binary masks provide supervision for the ego probability mask predicted by the Semantic Head.

Instance masks and interaction-dependent roles. For HTEWorld, instance segmentation is obtained directly from the BEHAVIOR-1K simulator. Robot pixels are always assigned to the ego. Object instances are assigned according to their role in the current interaction: an object belongs to the world before manipulation, becomes part of the ego while being manipulated or carried, and returns to the world after the interaction. Therefore, instance segmentation determines which pixels belong to each entity, while interaction information determines its current world-ego role.

Interaction-consistency filtering. We determine whether a candidate object is currently manipulated using complementary interaction cues, including robot-object overlap, proximity to the end-effector, consistency of motion between robot and object regions, mask confidence, connected-component size, and temporal consistency. Temporal tracking suppresses short-lived or identity-switching assignments and helps preserve object identity through brief occlusions. A candidate object is assigned to the ego only when the available cues consistently support an interaction.

The robot masks and the masks of currently manipulated objects are combined into the ego target, while all remaining pixels are assigned to the world. These annotations are used only for training. During inference, WEM predicts the world-ego assignment internally and does not require external masks.

A.4 Optical Flow Annotation

Optical-flow supervision is used for the motion-based alternative described in Appendix C.1. This definition requires the optical flow caused by the robot and manipulated objects rather than the total apparent motion in the image.

We estimate dense optical flow $\mathbf{F}$ between consecutive frames using RAFT (Teed & Deng, 2020). Camera-induced motion is approximated by fitting a homography $\mathbf{H}$ to matched feature correspondences using RANSAC and converting the estimated transformation into a camera-induced flow field $\mathbf{F}_{\text{cam}}$. We then compute the residual flow as

$$\mathbf{F}_{\text{obj}} = \mathbf{F} - \mathbf{F}_{\text{cam}}, \tag{5}$$

which serves as a proxy for the optical flow caused by robot and manipulated-object motion.

The residual flow is approximate because a single homography cannot fully explain viewpoint-induced motion in scenes with varying depth, and additional errors can arise from occlusion and optical-flow estimation. The resulting residual-flow targets are used to supervise the Flow Head in the motion-based variants.

A.5 Annotation on WorldArena and DROID

WorldArena (Shang et al., 2026) and DROID (Khazatsky et al., 2024) do not provide simulator-exported instance segmentation. We therefore use SAM 3 (Carion et al., 2026) to segment and track the robot and candidate objects in the videos. The resulting masks are processed using the same interaction-consistency procedure described in Appendix A.3 to identify currently manipulated objects and construct binary world-ego targets.

Language instructions are generated using the procedure in Appendix A.2, with the embodiment description and available task context adapted to each dataset. Neither SAM 3 nor Gemini is required during WEM inference.

For DROID, we construct a trajectory-disjoint subset with 5,591 training trajectories (approximately 40K clips) and 294 validation trajectories, corresponding to approximately 2K clips. We use only the left view of the first exterior camera (`exterior_image_1_left`) for all experiments. Videos

are processed at 15 FPS and segmented into 37-frame clips with a resolution of 480×704. All compared methods are trained and evaluated on the same subset, camera view, and video preprocessing configuration.

B Training and Evaluation Protocols

B.1 Training Configuration

The state predictor is initialized from Qwen3-VL-2B-Instruct, and the video generator is initialized from Wan2.1-T2V-1.3B or Wan2.2-TI2V-5B according to the evaluated model configuration. The Qwen3-VL backbone is frozen, while the learnable world and ego queries and the SR-MoE video generator are optimized during training.

We use $N_w = 192$ world queries and $N_e = 64$ ego queries. For both Wan backbones, the first 24 DiT blocks form the shared Preceding Expert, while the final six blocks are duplicated to initialize independently optimized World and Ego Experts.

On HTEWorld, training is performed for four epochs on $16\times$ NVIDIA A100 80GB GPUs with a global batch size of 64 and a learning rate of 1×10^{-5}. We maintain an exponential moving average of the model weights with decay 0.99 and use the averaged checkpoint for evaluation. Unless otherwise specified, the additional analyses in this appendix use WEM with the Wan2.2 backbone.

PAN uses the same Qwen3-VL and Wan initialization but predicts a unified state and retains a single generation stream. All compared video generators are fine-tuned rather than adapted only through LoRA. On DROID, the compared methods follow the same optimization protocol on the trajectory-disjoint subset described in Appendix A.5.

B.2 Multi-Turn Generation and Baseline Adaptation

WEM generates video chunks autoregressively. The first round is conditioned on the initial observation and the current instruction. For subsequent rounds, the state predictor receives the initial observation, previous instruction-video pairs, and the current instruction. The visual history consists of previously generated videos rather than reference future frames.

Cosmos-Predict2.5 and WoW-7B require adaptation to this multi-turn setting. During training, both models use a mixed conditioning strategy: with probability 0.9, a clip is conditioned on the preceding chunk, and with probability 0.1, it is initialized from the first frame. This exposes the models to both trajectory continuation and trajectory initialization during training.

At evaluation, Cosmos-Predict2.5 uses `image2world` for the first generation round and `video2world` for subsequent rounds. Later rounds are conditioned on the last ten latent frames of the preceding generated chunk.

WoW-7B uses `video2world` throughout generation. For the first round, the conditioning video is constructed by repeating the initial frame 41 times. For subsequent rounds, a 41-frame tail from the preceding generated video is used. WoW generates 82 frames internally and discards the 41 conditioning frames. The remaining 41 generated frames are retained as the conditioning video for the next round and uniformly subsampled to 37 frames for evaluation.

On HTEWorld, all methods are evaluated using 37 output frames per chunk at 480×480 resolution and 16 FPS. We use 35 diffusion steps. The guidance scale is set to 5.0 for all methods.

B.3 Evaluation Protocol

We adopt the video quality evaluation protocol introduced by WorldArena (Shang et al., 2026), which comprises 16 metrics across six dimensions: visual quality, motion quality, content consistency, physics adherence, 3D accuracy, and controllability. We report the aggregated EWMScore as the primary metric. The six HTEWorld-specific diagnostics introduced in Appendix D evaluate complementary properties of multi-turn generation, including cross-chunk continuity, state preservation, and navigation-manipulation transitions.

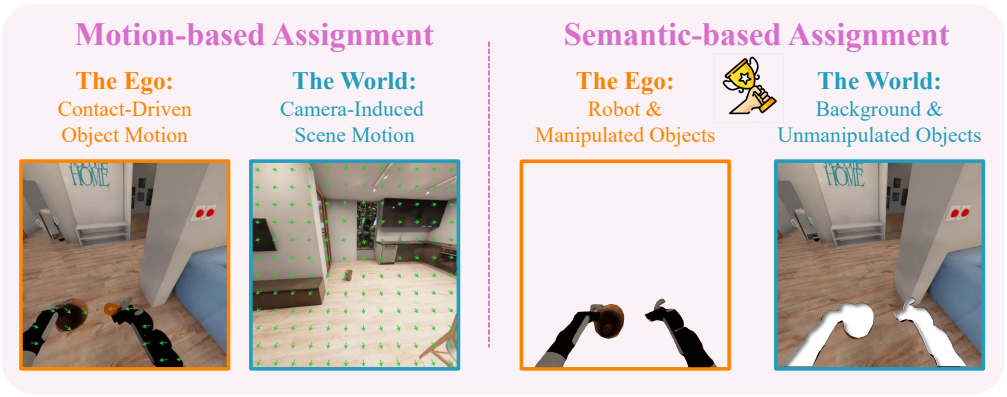


Figure 5: **Comparison of motion-based and semantic-based world-ego assignments.** The motion-based assignment derives the world-ego assignment from a motion proxy, where the ego corresponds to residual motion induced by the robot and manipulated objects after camera-motion compensation, while the world corresponds to camera-induced background motion. The semantic-based assignment derives the world-ego assignment from a semantic ego mask, where the ego consists of the robot and currently manipulated objects, while the world consists of the background and unmanipulated objects.

On WorldArena, WEM is evaluated following the benchmark protocol. The other results reported in the main-text WorldArena comparison are taken from the original WorldArena paper. The additional Mask2IV comparison in Appendix E.2 is evaluated by us. On DROID, Wan2.2, PAN, and WEM are trained and evaluated on the same trajectory-disjoint subset described in Appendix A.5.

C Alternative Designs

This section provides the implementation details of the variants used in the SR-MoE ablation study in Sec. 5.3. The ablation examines two architectural design choices: *Token Separation*, which determines whether world and ego tokens are separated according to the predicted assignment before subsequent expert processing, and *Expert Specialization*, which determines whether separate World and Ego Experts are used. We additionally compare two choices of *Assignment Proxy* used to derive the world-ego assignment: a semantic ego mask identifying the robot and currently manipulated objects, and a motion proxy capturing motion induced by the robot and manipulated objects after camera-motion compensation. The two assignment views are illustrated in Fig. 5. We use *Proxy Head* to collectively refer to the Semantic Head and Flow Head used in these two assignment variants. This allows us to examine whether the effectiveness of SR-MoE depends on the choice of assignment proxy. The corresponding generator variants are illustrated in Fig. 6.

C.1 Motion-Based World-Ego Assignment

The default SR-MoE derives the world-ego assignment from a semantic proxy, namely the ego probability mask predicted by the Semantic Head. To evaluate the choice of assignment proxy, we construct a motion-based alternative that instead predicts a motion proxy corresponding to motion induced by the robot and manipulated objects.

The motion-based variant replaces the Semantic Head with a Flow Head while keeping the remaining generator structure unchanged. The Flow Head takes intermediate features from four selected Preceding Expert blocks, fuses them using learned per-block weights, and decodes the fused representation with a multi-scale U-Net head to predict a two-dimensional flow vector

$$\widehat{\mathbf{F}}^{\mathrm{obj}}_{k,t,i} \in \mathbb{R}^2, \tag{6}$$

where $\widehat{\mathbf{F}}^{\mathrm{obj}}_{k,t,i}$ denotes the predicted residual flow at token (t, i) in generation round k. We convert the predicted flow vector into a scalar motion magnitude

$$\widehat{f}_{k,t,i} = \left\| \widehat{\mathbf{F}}^{\mathrm{obj}}_{k,t,i} \right\|_2 . \tag{7}$$

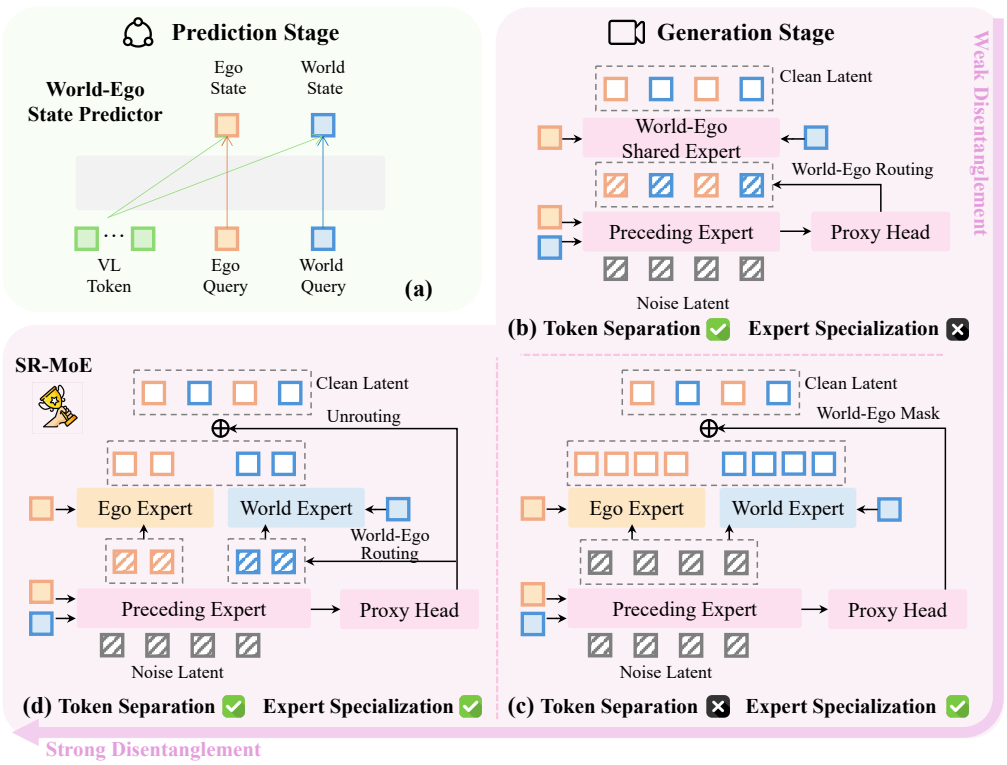


Figure 6: **Architectural variants of SR-MoE evaluated in the ablation study.** Our framework contains a shared prediction stage and three generator variants with different combinations of Token Separation and Expert Specialization. (a) The prediction stage uses world and ego queries to extract two role-specific states from vision-language tokens. A Proxy Head predicts a world-ego proxy from the Preceding Expert features, instantiated as either a semantic ego mask or motion induced by the robot and manipulated objects. The world-ego assignment is then derived from this proxy. (b) With Token Separation only, the assignment separates the Preceding Expert output into world and ego token groups, which are processed independently by the same shared expert under their corresponding states. (c) With Expert Specialization only, the complete token sequence is processed by separate World and Ego Experts, and the proxy-derived assignment is used to combine their outputs. (d) The full SR-MoE combines both designs by routing the assigned world and ego token groups to their corresponding specialized experts and restoring the processed tokens to their original spatiotemporal positions.

Given a motion threshold τ_{flow}, the corresponding ego and world token sets are defined as

$$\mathcal{I}_k^e = \{(t,i) \mid \widehat{f}_{k,t,i} \geq \tau_{\text{flow}}\}, \qquad \mathcal{I}_k^w = \{(t,i) \mid \widehat{f}_{k,t,i} < \tau_{\text{flow}}\}. \tag{8}$$

For variants with Token Separation, we apply the same Neighbor-Expanded Assignment as in the default SR-MoE to obtain $\widetilde{\mathcal{I}}_k^w$ and $\widetilde{\mathcal{I}}_k^e$ before subsequent expert processing. For variants without Token Separation, the motion proxy is used only to combine the outputs of the World and Ego Experts after expert processing. We therefore convert the predicted motion magnitude into a continuous world weight

$$\alpha_{k,t,i} = \sigma\left(\frac{\tau_{\text{flow}} - \widehat{f}_{k,t,i}}{\delta_{\text{flow}}}\right), \tag{9}$$

where δ_{flow} is a constant controlling the transition sharpness. Small residual motion places greater weight on the World Expert, while large residual motion places greater weight on the Ego Expert. The expert outputs are combined as

$$\widehat{\mathbf{Z}}_k = \alpha_k \odot \widehat{\mathbf{Z}}_k^w + (1 - \alpha_k) \odot \widehat{\mathbf{Z}}_k^e, \tag{10}$$

where α_k collects the token-wise weights $\alpha_{k,t,i}$, $\odot$ denotes element-wise multiplication, and α_k is broadcast along the feature dimension to match the expert outputs.

The semantic- and motion-based variants therefore differ only in the proxy used to derive the world-ego assignment: an ego probability mask in the semantic variant and robot- and manipulated-object-induced residual flow in the motion-based variant.

C.2 SR-MoE Ablation Variants

The three generator variants shown in Figure 6 independently control Token Separation and Expert Specialization while sharing the same state predictor and Preceding Expert.

Token Separation only. This variant enables Token Separation while removing Expert Specialization. Starting from the Preceding Expert output $\mathbf{Z}_k^P$, the proxy-derived world and ego assignments are expanded using Neighbor-Expanded Assignment, and the corresponding token groups are processed by the same shared expert. Although both groups share the same expert parameters, they are processed separately within the expert. Tokens belonging exclusively to the world group are not visible to the ego group, and tokens belonging exclusively to the ego group are not visible to the world group. Tokens in the overlapping region introduced by Neighbor-Expanded Assignment are included in both groups. In addition, the world group is conditioned only on $\mathbf{S}_k^w$, while the ego group is conditioned only on $\mathbf{S}_k^e$. The two outputs are then restored to their original spatiotemporal positions, with overlapping positions averaged. This variant therefore isolates assignment-based token separation while keeping the generation parameters shared between the world and ego.

Expert Specialization only. This variant removes Token Separation while retaining Expert Specialization. The complete Preceding Expert output $\mathbf{Z}_k^P$ is sent to independently optimized World and Ego Experts, each conditioned on its corresponding state. The proxy-derived assignment is applied only after expert processing to combine the two outputs. For semantic assignment, the predicted ego probability mask $\mathbf{M}_k$ provides the fusion weights:

$$\widehat{\mathbf{Z}}_k = (1 - \mathbf{M}_k) \odot \widehat{\mathbf{Z}}_k^w + \mathbf{M}_k \odot \widehat{\mathbf{Z}}_k^e, \tag{11}$$

where $\mathbf{M}_k$ is broadcast along the feature dimension to match the expert outputs. For motion-based assignment, the flow-derived weights α_k defined in Appendix C.1 are used instead. This variant therefore isolates the effect of specialized World and Ego parameters without restricting either expert to a role-specific token group.

Token Separation + Expert Specialization. The proxy-derived assignment is expanded by Neighbor-Expanded Assignment, and the resulting world and ego token groups are routed to independently optimized World and Ego Experts, each conditioned only on its corresponding state. The expert outputs are subsequently restored to their original spatiotemporal positions. Tokens covered by only one expanded assignment use the corresponding expert output directly, while tokens in the overlapping region are processed by both experts and their outputs are averaged. This variant combines assignment-based token separation with specialized generation parameters. Together with the two variants above, it allows the contributions of Token Separation and Expert Specialization to be evaluated independently under the same overall framework.

D HTEWorld-Specific Metrics and Evaluation

D.1 Metric Definitions

The WorldArena metrics remain our primary evaluation protocol. HTEWorld additionally evaluates continuous generation across multiple instruction rounds containing both navigation and manipulation. We therefore introduce six complementary diagnostics. RCBD, LPSA, and CISR examine multi-turn continuity and state evolution, while PMPA, CPDM, and FPHS focus on navigation-manipulation behavior and transitions between the two phases.

Let $\mathbf{V} = (\mathbf{V}_1, \ldots, \mathbf{V}_K)$ denote a generated trajectory and $\mathbf{V}^* = (\mathbf{V}_1^*, \ldots, \mathbf{V}_K^*)$ its reference, with chunks aligned by instruction round. Each chunk has a phase label $p_k \in \{\mathrm{Nav}, \mathrm{Manip}\}$. The first

and last frames of $\mathbf{V}_k$ are $\mathbf{V}_k^{(1)}$ and $\mathbf{V}_k^{(-1)}$. Let d_p denote LPIPS distance (Zhang et al., 2018) and $\phi_{\text{CLIP}}(\cdot)$ the CLIP-based visual representation (Radford et al., 2021) of a frame set or video chunk. Feature comparisons use cosine similarity. Each metric is computed per trajectory and then averaged over the evaluation set.

Rollout Chunk-Boundary Dynamics (RCBD). RCBD evaluates whether the appearance and motion changes across adjacent generated chunks match those of the reference trajectory. For each adjacent pair, we compute the appearance gap

$$b_k = d_p\left(\mathbf{V}_k^{(-1)}, \mathbf{V}_{k+1}^{(1)}\right). \tag{12}$$

To characterize the corresponding motion change, let $\mathbf{F}_k^-$ denote the optical flow between the last two frames of $\mathbf{V}_k$, and $\mathbf{F}_{k+1}^+$ the optical flow between the first two frames of $\mathbf{V}_{k+1}$. We define the motion gap as

$$m_k = \frac{1}{|\Omega|}\sum_{\mathbf{p}\in\Omega}\left\|\mathbf{F}_k^-(\mathbf{p}) - \mathbf{F}_{k+1}^+(\mathbf{p})\right\|_2, \tag{13}$$

where Ω denotes the image domain. Let b_k^* and m_k^* denote the corresponding appearance and motion gaps computed from the reference trajectory.

We compare the generated and reference gaps using the symmetric ratio score

$$\mathcal{S}(x,y) = \exp\left(-\left|\log\frac{x+\epsilon}{y+\epsilon}\right|\right), \tag{14}$$

where ϵ is a small constant for numerical stability. RCBD is then defined as

$$\text{RCBD} = \frac{1}{K-1}\sum_{k=1}^{K-1}\sqrt{\mathcal{S}(b_k, b_k^*)\mathcal{S}(m_k, m_k^*)}. \tag{15}$$

Matching reference boundary dynamics avoids rewarding frozen or overly smooth transitions.

Late-Prefix State Alignment (LPSA). LPSA measures whether the generated trajectory reaches visual states similar to the reference trajectory after each instruction round, with greater weight placed on later rounds. Let $\mathbf{V}_k^{[-W:]}$ and $\mathbf{V}_k^{*[-W:]}$ denote the last $W = 4$ frames of the generated and reference chunks, respectively. Their state similarity is

$$r_k = \cos\left(\phi_{\text{CLIP}}(\mathbf{V}_k^{[-W:]}), \phi_{\text{CLIP}}(\mathbf{V}_k^{*[-W:]})\right), \tag{16}$$

where $\cos(\cdot,\cdot)$ denotes cosine similarity. We compute

$$\text{LPSA} = \frac{\sum_{k=1}^{K} k\, r_k}{\sum_{k=1}^{K} k}. \tag{17}$$

The weighting emphasizes later states, where errors accumulated over previous rounds become increasingly important.

Chunk Instruction-Step Retrieval (CISR). CISR evaluates whether each generated chunk corresponds to the correct stage of the reference trajectory. For each generated chunk $\mathbf{V}_k$, we rank all reference chunks $\{\mathbf{V}_j^*\}_{j=1}^K$ from the same trajectory by $\cos\big(\phi_{\text{CLIP}}(\mathbf{V}_k), \phi_{\text{CLIP}}(\mathbf{V}_j^*)\big)$. Let ρ_k denote the rank of the corresponding reference chunk $\mathbf{V}_k^*$, where rank 1 indicates the highest similarity. We define

$$\text{CISR} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{\rho_k}. \tag{18}$$

A higher CISR indicates better alignment with the correct instruction step along the reference trajectory.

Phase-Matched Motion Profile Alignment (PMPA). PMPA evaluates whether the generated trajectory exhibits motion patterns similar to the reference within the same instruction phase. Since navigation and manipulation differ substantially in their characteristic motion, we compare phase-matched generated and reference chunks.

For each consecutive frame pair, we estimate optical flow using RAFT (Teed & Deng, 2020). Let u denote the resulting flow-magnitude field. We construct a four-dimensional motion descriptor

$$\mathbf{q} = \left[\bar{u}/D,\ \widetilde{u}/D,\ \log\left(1 + \frac{\widetilde{u}}{\bar{u} + \epsilon}\right),\ \mathcal{E}(u) \right], \tag{19}$$

where $\bar{u}$ is the median flow magnitude, $\widetilde{u}$ is the mean of the largest 20% of flow magnitudes, D is the frame diagonal, $\mathcal{E}(u)$ denotes the normalized entropy of the flow-magnitude distribution, and ϵ is a small constant for numerical stability.

The generated and reference motion-profile sequences are temporally resampled to 16 positions. Let $\mathbf{q}_{k,s}$ and $\mathbf{q}^*_{k,s}$ denote the generated and reference descriptors at resampled position s of chunk k, respectively. Their discrepancy is

$$\delta_k = \frac{1}{16} \sum_{s=1}^{16} \left\| \mathbf{q}_{k,s} - \mathbf{q}^*_{k,s} \right\|_2 . \tag{20}$$

We define the PMPA score of chunk k as

$$s_k^{\mathrm{PMPA}} = \exp\left(-\delta_k / \tau_{\mathrm{PMPA}}\right), \tag{21}$$

where τ_{PMPA} controls the sensitivity to motion-profile discrepancies. The trajectory-level PMPA is then

$$\mathrm{PMPA} = \frac{1}{K} \sum_{k=1}^{K} s_k^{\mathrm{PMPA}}, \tag{22}$$

where K is the number of evaluated chunks. Each generated chunk is compared with its corresponding reference chunk from the same navigation or manipulation phase.

Cross-Phase Discriminative Margin (CPDM). CPDM evaluates whether a generated chunk is more similar to its corresponding reference than to reference chunks from the opposite phase. For chunk k, we define the positive similarity as

$$r_k^+ = \cos\left(\phi_{\mathrm{CLIP}}(\mathbf{V}_k), \phi_{\mathrm{CLIP}}(\mathbf{V}^*_k)\right), \tag{23}$$

and the strongest opposite-phase similarity as

$$r_k^- = \max_{j:\, p_j \neq p_k} \cos\left(\phi_{\mathrm{CLIP}}(\mathbf{V}_k), \phi_{\mathrm{CLIP}}(\mathbf{V}^*_j)\right). \tag{24}$$

We compute

$$\mathrm{CPDM} = \frac{1}{K} \sum_{k=1}^{K} \sigma\left(\frac{r_k^+ - r_k^-}{\tau_{\mathrm{CPDM}}} \right), \qquad \tau_{\mathrm{CPDM}} = 0.05, \tag{25}$$

where σ denotes the sigmoid function. A higher score indicates that the generated trajectory better preserves the distinction between navigation and manipulation phases.

Frontier Phase-Hop State Consistency (FPHS). FPHS focuses specifically on transitions between navigation and manipulation. For each transition boundary satisfying $p_k \neq p_{k+1}$, we extract $R = 4$ frames on both sides of the boundary from both the generated and reference trajectories.

We compute optical-flow magnitudes over the reference transition window and accumulate them temporally to obtain a spatial change map. The top-20% locations in this map define the region with the strongest reference transition dynamics. The generated and reference transition windows are restricted to this region and encoded using $\phi_{\mathrm{CLIP}}(\cdot)$. Their cosine similarity provides the score for the corresponding transition boundary. FPHS is obtained by averaging these scores over all navigation-manipulation transition boundaries in the trajectory.

Table 7: Pearson correlation between HTEWorld-specific metrics and human ratings of complete multi-turn rollouts. Task completion is judged from generated videos rather than measured through robot execution.

Metric	Perceived task completion	Visual realism
RCBD	**0.726**	0.584
LPSA	0.710	0.649
CISR	0.569	0.706
PMPA	0.651	0.622
CPDM	0.465	0.653
FPHS	0.516	**0.834**

Table 8: Results on the six HTEWorld-specific metrics. Higher is better. PAN and WEM use the Wan2.2 backbone.

Model	RCBD	LPSA	CISR	PMPA	CPDM	FPHS
WoW-7B	0.23	0.83	0.49	0.45	0.47	0.85
Cosmos-Predict2.5-2B	0.24	0.83	0.50	0.47	0.48	0.86
Cosmos-Predict2.5-14B	0.26	0.83	0.51	0.48	0.49	0.85
PAN	0.27	0.86	0.49	0.50	0.46	0.88
WEM	**0.31**	**0.87**	**0.57**	**0.54**	**0.52**	**0.89**

D.2 Human Correlation and Interpretation

We evaluate the agreement between the six HTEWorld-specific diagnostics and human judgments using 300 generated trajectories from each of seven systems, yielding 2,100 complete multi-turn rollouts. We recruit 15 volunteers, each of whom independently evaluates all 2,100 anonymized rollouts on a five-point scale along two dimensions. *Perceived task completion* measures whether the generated video appears to accomplish the instructed behavior, while *visual realism* evaluates visual coherence, temporal consistency, and physical plausibility. For each rollout, ratings are averaged across the 15 annotators, and Pearson correlation is computed between the resulting human scores and each diagnostic metric.

As shown in Table 7, all six metrics are positively correlated with both rating dimensions. RCBD shows the strongest correlation with perceived task completion, while FPHS shows the strongest correlation with visual realism. LPSA also correlates strongly with both dimensions. The different correlation patterns indicate that the six metrics capture complementary properties of multi-turn generation rather than serving as interchangeable measures.

The metrics nevertheless remain diagnostic rather than direct measures of robot task success. RCBD and PMPA depend on optical-flow estimates and cannot establish that the correct object is manipulated. LPSA, CISR, and CPDM rely on visual representations that may overlook fine-grained contact, geometry, or physical violations. CISR also depends on the number and visual similarity of reference chunks within a trajectory. FPHS focuses on reference-derived transition regions and may miss errors elsewhere in the scene. More generally, reference-based evaluation may penalize a plausible rollout that differs from the recorded trajectory. We therefore use these diagnostics together with the primary WorldArena metrics and qualitative evaluation.

D.3 Results on HTEWorld-Specific Metrics

Table 8 reports results on the six HTEWorld-specific metrics. PAN and WEM use the Wan2.2 backbone in this comparison. WEM achieves the best result on all six metrics. In particular, RCBD increases from 0.27 to 0.31, CISR from 0.49 to 0.57, PMPA from 0.50 to 0.54, and CPDM from 0.46 to 0.52 compared with PAN. LPSA and FPHS also improve from 0.86 to 0.87 and from 0.88 to 0.89, respectively. These results provide complementary evidence for improved multi-turn continuity and navigation-manipulation generation beyond the aggregate EWMScore reported in the main text.

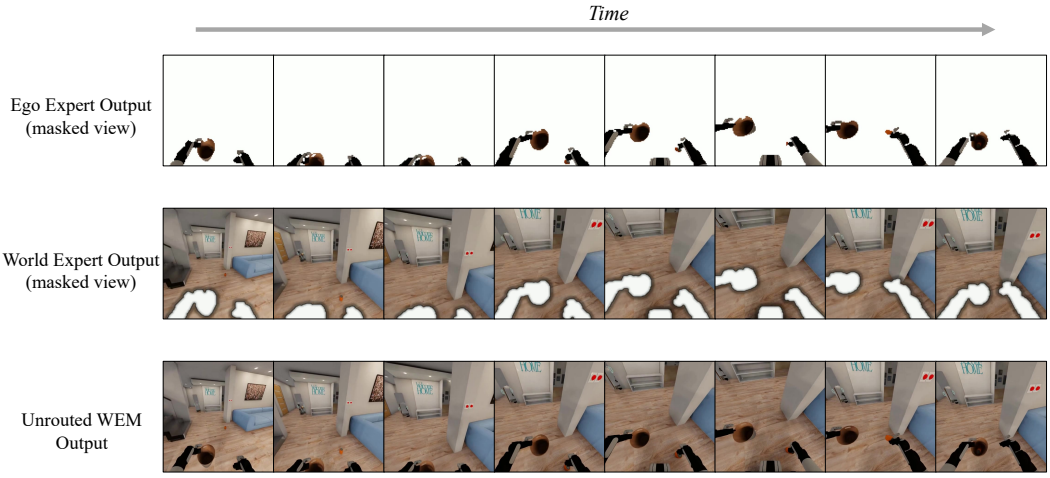


Figure 7: Masked views of the role expert outputs and the recomposed WEM output. The Ego Expert output is displayed in ego-assigned regions, and the World Expert output in world-assigned regions, using the assignment predicted by the Semantic Head. Complementary regions are left blank for visualization. The bottom row shows the output after restoring the expert outputs to their original positions. The masking is used for visualization only.

E Additional Experiments and Analyses

E.1 World-Ego Assignment and Expert Specialization

Assignment accuracy. We evaluate the world-ego masks predicted by the Semantic Head on the HTEWorld validation set using the default WEM configuration. At the final denoising step, the predicted masks achieve an Ego IoU of 70.3%, a World IoU of 91.7%, and a mean IoU of 81.0%. The predictions are already accurate at an early denoising stage, such as the second step, and improve only moderately during later steps. The lower Ego IoU reflects the greater difficulty of identifying the smaller robot-object interaction regions compared with the larger world region.

Dynamic role transitions. The semantic world-ego assignment changes with the current interaction rather than assigning a fixed role to each object category. An object belongs to the world before manipulation, becomes part of the ego while it is manipulated or carried, and can return to the world after the interaction. For example, in a cleanup trajectory, a can is initially part of the surrounding scene and is reassigned to the ego after the robot picks it up. The assignment therefore follows the interaction state of an object rather than whether its pixels are moving.

Role expert specialization. Fig. 7 visualizes the regions associated with the World and Ego Experts together with the recomposed WEM output. The Ego Expert view contains the robot and currently manipulated objects, while the World Expert view contains the remaining scene. The assignments also evolve as objects enter or leave the interaction. Since these views are obtained by masking the expert outputs according to the predicted assignment, they visualize which regions are associated with each expert rather than independently proving expert specialization. The quantitative generator ablations in the main text provide the corresponding evidence for the benefit of specialized generation parameters.

E.2 Comparison with Mask2IV

Mask2IV (Li et al., 2026) uses object masks to guide interaction-centric video generation. We compare WEM with its Text-conditioned Trajectory Generation (TT-Gen) variant on WorldArena. Mask2IV is fine-tuned on the same WorldArena training split as WEM, with both methods using masks from the same SAM 3 annotation pipeline.

Table 9: Additional comparison with Mask2IV on WorldArena, evaluated by us. Both methods use text instructions; Mask2IV additionally receives an initial object mask at inference.

Method	EWMScore ↑
Mask2IV TT-Gen	52.36
WEM (Wan2.2)	**58.10**

The two methods differ in how masks are used during generation. Mask2IV takes a SAM 3-derived initial object mask as an additional inference condition and predicts a mask trajectory to guide a shared video generator. WEM requires no mask at inference; instead, it predicts the world-ego assignment during denoising and routes tokens to separate World and Ego Experts. Because the methods differ in both architecture and inference conditions, this serves as an additional benchmark comparison rather than a controlled isolation of token routing; the controlled SR-MoE ablations are provided in Sec. 5.3.

As shown in Table 9, WEM achieves an EWMScore of 58.10, outperforming Mask2IV TT-Gen by 5.74 points. Since WorldArena focuses on manipulation, this result further shows that the advantage of WEM is not limited to trajectories containing navigation.

F Limitations and Future Work

Long-horizon error accumulation. WEM improves long-horizon scene consistency and instruction following, but autoregressive generation still accumulates errors over successive rounds. Incorrect object states or scene geometry in one generated chunk can affect subsequent predictions, and information from early rounds may gradually be lost as the trajectory becomes longer. The current world and ego states summarize the available context but do not provide persistent memory over arbitrarily long horizons. Incorporating more explicit long-term memory or hierarchical representations of task progress is a promising direction.

Real-world hybrid evaluation. DROID allows us to evaluate WEM on real-world tabletop manipulation videos, but it does not contain long-horizon trajectories with interleaved navigation and manipulation. The lack of corresponding real-world data therefore limits evaluation of the complete hybrid setting outside simulation. Extending the evaluation to real trajectories combining large viewpoint changes, navigation, manipulation, sensing noise, and diverse physical interactions would provide a more comprehensive test of long-horizon embodied generation.

Downstream scope. Our experiments evaluate embodied video generation rather than closed-loop robot control. Improvements in EWMScore, HTEWorld-specific diagnostics, and perceived task completion therefore do not directly establish improved planning or policy performance. Evaluating whether the generated world and ego evolution can benefit downstream planning, policy learning, or real-robot control remains an important direction for future work.