EWMBench: Evaluating Scene, Motion,and Semantic Quality in Embodied World Models

Yue Hu[1, 4], Siyuan Huang[1, 2], Yue Liao[1, 3], Shengcong Chen[1], Pengfei Zhou[1], Liliang Chen[1], Guanghui Ren[1], Maoqing Yao[1]

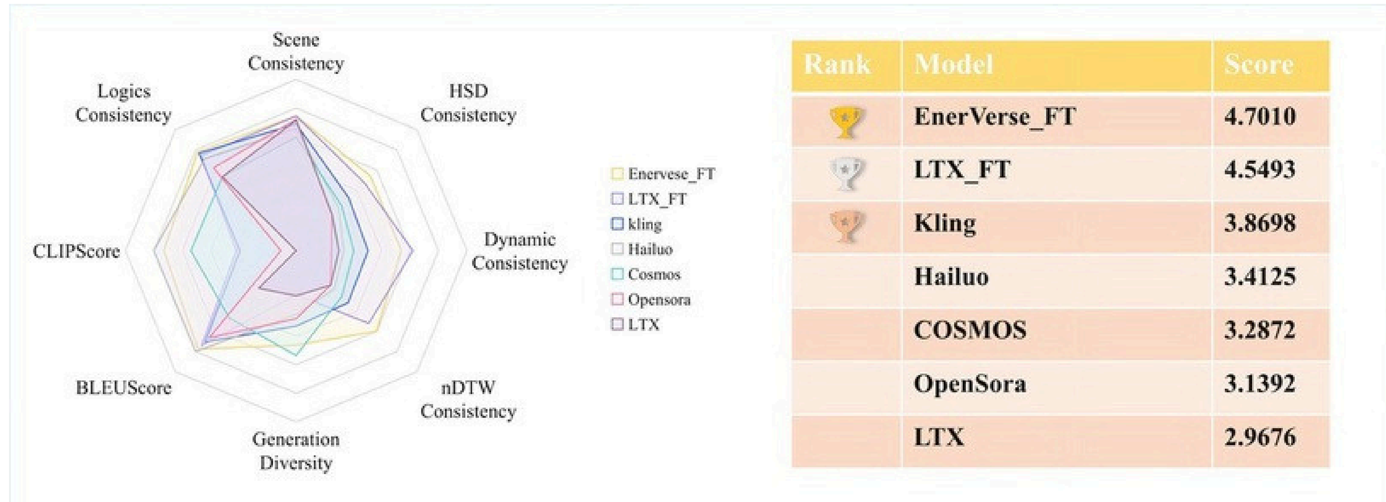
Agibot[1], Shanghai Jiao Tong University[2], National University of Singapore[3], Harbin Institute of Technology[4]

Abstract

Recent advances in creative AI have enabled the synthesis of high-fidelity images and videos conditioned on language instructions. Building on these developments, text-to-video diffusion models have evolved into embodied world models (EWMs) capable of generating physically plausible scenes from language commands, effectively bridging vision and action in embodied AI applications. This work addresses the critical challenge of evaluating EWMs beyond general perceptual metrics to ensure the generation of physically grounded and action-consistent behaviors. We propose the Embodied World Model Benchmark (EWMBench), a dedicated framework designed to evaluate EWMs based on three key aspects: visual scene consistency, motion correctness, and semantic alignment. Our approach leverages a meticulously curated dataset encompassing diverse scenes and motion patterns, alongside a comprehensive multi-dimensional evaluation toolkit, to assess and compare candidate models. The proposed benchmark not only identifies the limitations of existing video generation models in meeting the unique requirements of embodied tasks but also provides valuable insights to guide future advancements in the field. The dataset, evaluation tools will be open-sourced at : https://github.com/AgibotTech/EWMBench

Motivation

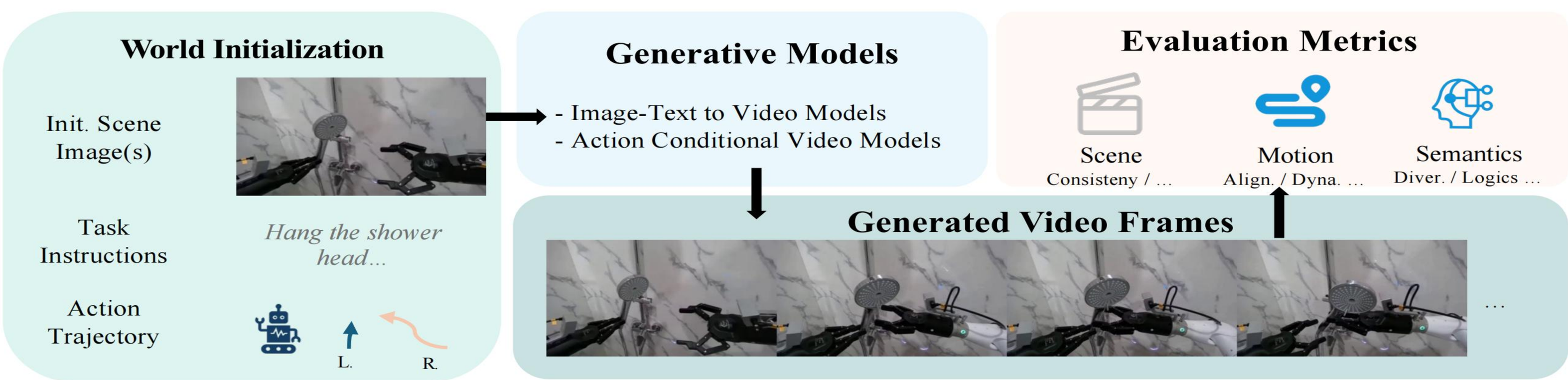
- Background: Diffusion models and large-scale data push generative AI from visual realism toward action-grounded generation.
- Challenge: It remains unclear how to evaluate whether models produce physically consistent embodied behaviors.
- Our Solution: We propose EWMBench — a unified benchmark with a diverse dataset and systematic evaluation across scene, motion, and semantics, revealing key gaps among current models.



(A) Evaluation Results of Video Generative Models (B) Comprehensive Ranking of Tested Models

Methodology

1.Overview of the EWMBENCH benchmark design



2.Evaluation Metrics

Scene Consistency

- Stability of background & objects
- Cosine similarity from **fine-tuned DINOv2** patch embeddings

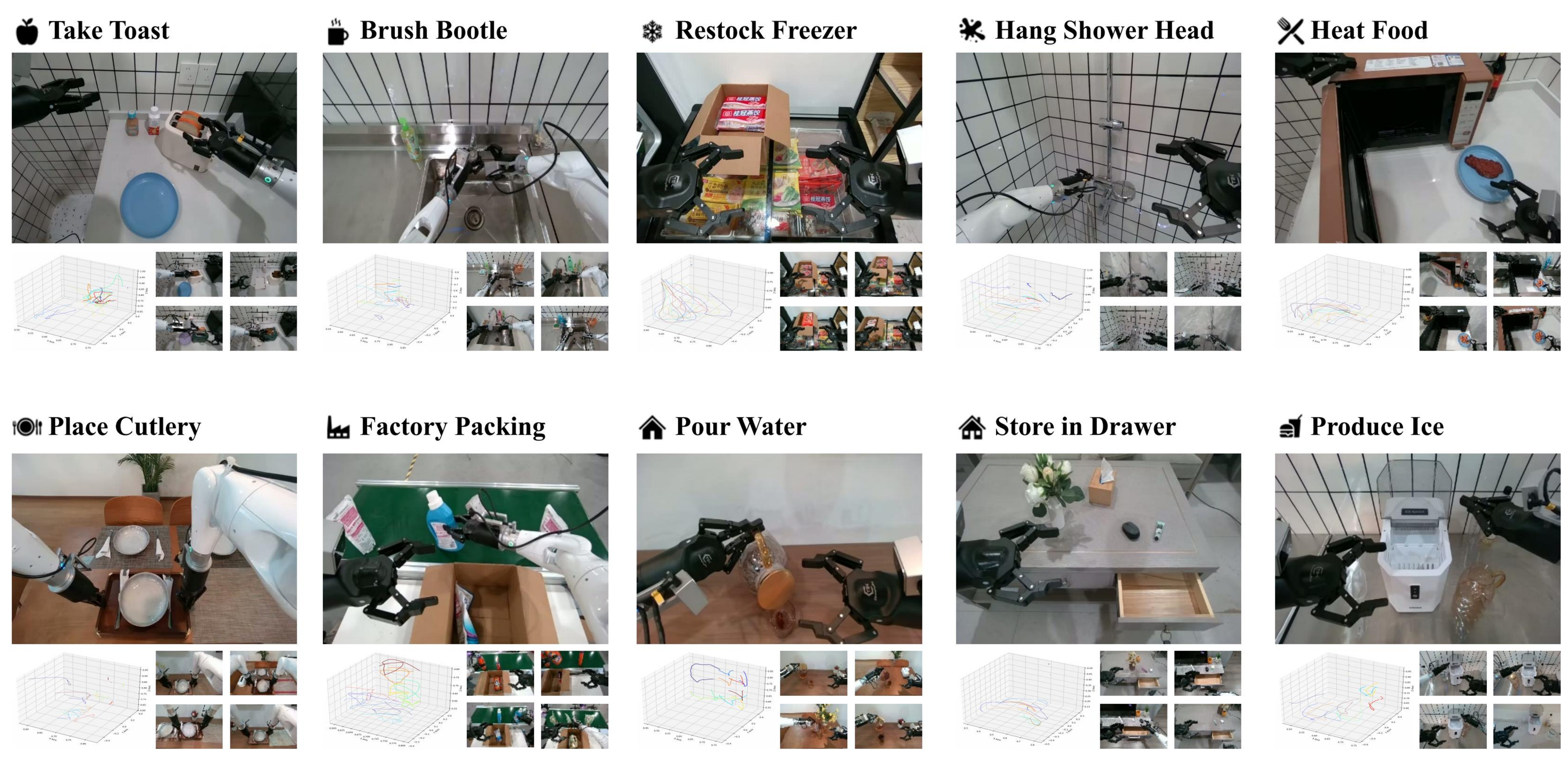
Action Motion

- EEF trajectory vs. GT
- **HSD** → spatial
- **NDTW** → temporal
- **DYN** → dynamic consistency

Semantics

- Caption-based alignment
- **BLEU Score**
- **CLIP Score**
- **Logic Consistency**

3.Diversity of World Initialization Dataset

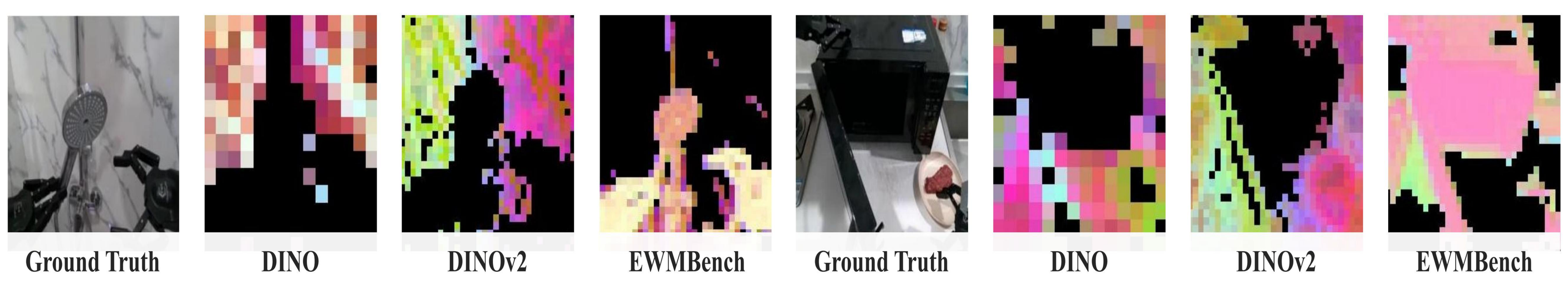


Experiment Result

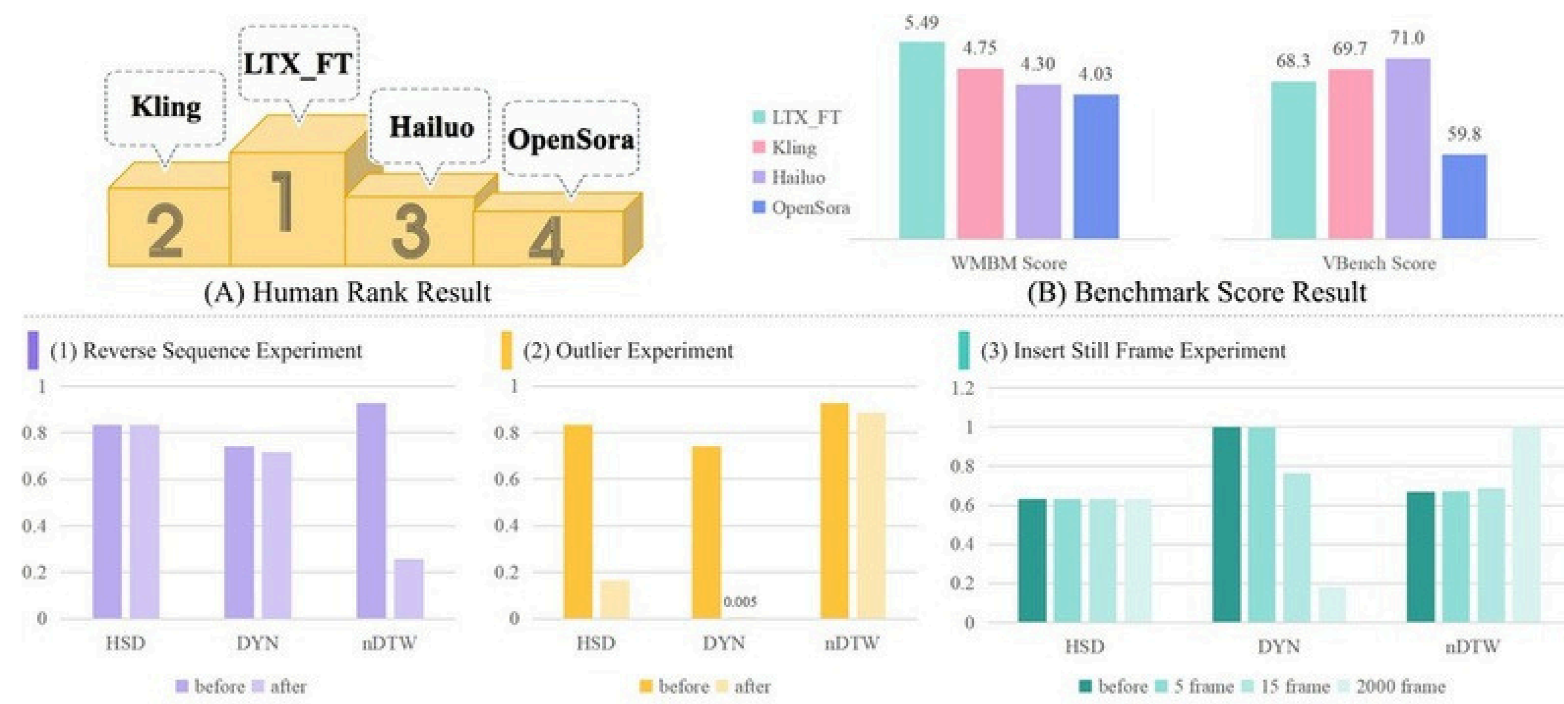
1.Evaluation results categorized into task scene, action motion, and semantics

Type	Model	Scene	Motion				Semantics					Overall
		SceneC	HSD	Dyn	nDTW	Sum.	Diversity	BLEU	CLIP	Logics	Sum.	
Dom.	EnerVerse_FT[17]	0.9427	**0.5356**	0.5363	**0.5957**	**1.6676**	0.0691	**0.1800**	0.8638	**0.9778**	**2.0907**	**4.7010**
	LTX_FT	**0.9436**	0.4758	**0.6197**	0.5208	1.6163	0.0162	0.1740	0.8548	0.9444	1.9894	4.5493
Comm.	Kling[22]	0.8888	0.3231	0.3047	0.3162	0.9440	0.0493	0.1675	0.8535	0.9667	2.0370	3.8698
	Hailuo[14]	0.8577	0.2229	0.1344	0.1789	0.5362	0.0370	0.1848	**0.8857**	0.9111	2.0186	3.4125
Open.	COSMOS[1]	0.7963	0.2500	0.2052	0.2533	0.7085	**0.0803**	0.1230	0.8458	0.7333	1.7824	3.2872
	OpenSora[29]	0.9210	0.1548	0.0474	0.1420	0.3442	0.0415	0.1598	0.8505	0.8222	1.8739	3.1392
	LTX[13]	0.9156	0.1575	0.1002	0.1425	0.4002	0.0174	0.0687	0.8324	0.7333	1.6518	2.9676

2.Feature Map Comparison within the Scene Consistency Metric



3.(A) Aggregated human rankings of model predictions. (B) Comparison of rankings produced by EWMBench and VBench, highlighting EWMBenchs's closer alignment with human judgments.(C) Complementarity of trajectory metrics.



(A) Human Rank Result (B) Benchmark Score Result

(C) Experimental results of trajectory consistency complementarity