WorldSimBench: Towards Video Generation Models as World Simulators

Yiran Qin[1,2*], Zhelun Shi[3*], Jiwen Yu[4], Xijun Wang[2], Enshen Zhou[3], Lijun Li[2],
Zhenfei Yin[2‡], Xihui Liu[4], Lu Sheng[3], Jing Shao[2†], Lei Bai[2†], Wanli Ouyang[2], Ruimao Zhang[1†]
[1]The Chinese University of Hong Kong, Shenzhen [2] Shanghai Artificial Intelligence Laboratory
[3]Beihang University [4]The University of Hong Kong

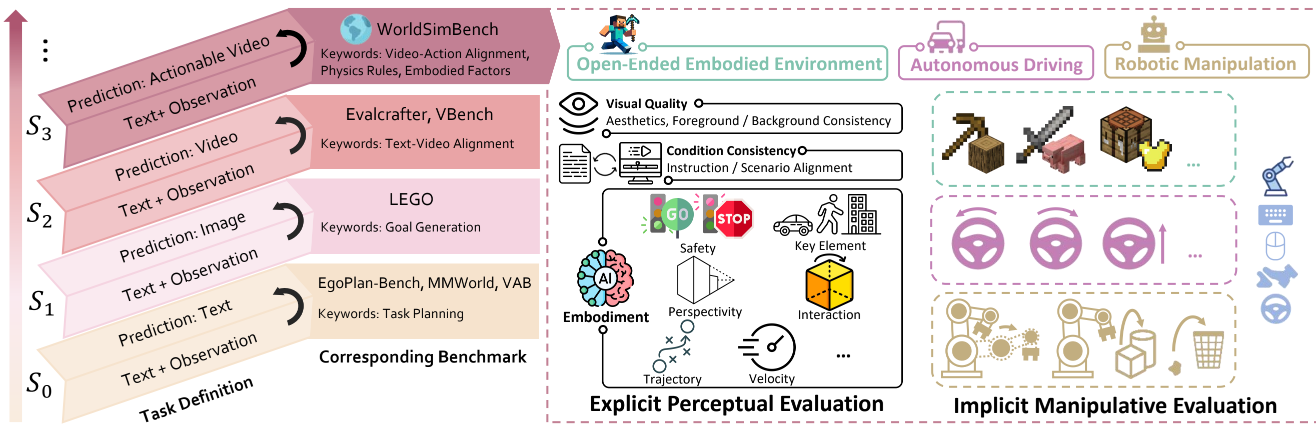

Figure 1. **Overview of the hierarchical capabilities of the Predictive Models.** Models at higher stages demonstrate more advanced capabilities. We take the initial step in evaluating Predictive Generative Models up to the S_3 stage, known as World Simulators, by introducing a parallel evaluation framework, WorldSimBench. WorldSimBench assesses the models both Explicit Perceptual Evaluation and Implicit Manipulative Evaluation, focusing on video generation and action transformation across three critical embodied scenarios.

Abstract

Recent advancements in predictive models have demonstrated exceptional capabilities in predicting the future state of objects and scenes. However, the lack of categorization based on inherent characteristics continues to hinder the progress of predictive model development. Additionally, existing benchmarks are unable to effectively evaluate higher-capability, highly embodied predictive models from an embodied perspective. In this work, we classify the functionalities of predictive models into a hierarchy and take the first step in evaluating World Simulators by proposing a dual evaluation framework called WorldSimBench. WorldSimBench includes ***Explicit Perceptual Evaluation*** *and* ***Implicit Manipulative Evaluation****, encompassing human preference assessments from the visual perspective and action-level evaluations in embodied tasks, covering three representative embodied scenarios: Open-Ended Embodied Environment, Autonomous Driving, and Robot Manipulation. In the Explicit Perceptual Evaluation, we introduce the HF-Embodied Dataset, a video assessment dataset based on fine-grained human feedback, which we use to train a Human Preference Evaluator that aligns with human perception and explicitly assesses the visual fidelity of World Simulators. In the Implicit Manipulative Evaluation, we assess the video-action consistency of World Simulators by evaluating whether the generated situation-aware video can be accurately translated into the correct control signals in dynamic environments. Our comprehensive evaluation offers key insights that can drive further innovation in video generation models, positioning World Simulators as a pivotal advancement toward embodied artificial intelligence.*

1. Extended Abstract

Before taking action, humans make predictions based on their objectives and observations of the current environment. These predictions manifest in various forms, *e.g.*, textual planning, visual imagination of future scene changes, or even subconscious planning at the action level. With the development of generative models, agents driven by these models are exhibiting predictive capabilities that enable them to complete embodied tasks by making human-like predictions, *e.g.*, high-level planning [3, 9], image-based guidance [1, 8], or future video prediction to drive actions [4, 5]). We refer to these models as **Predictive Models**. Recently, these models have been widely applied across various domains spanning

Table 1. **Comparisons between existing Predictive Model benchmarks.** Interactive Environment refers to the interaction with the simulation environment during the prediction phase. Task-Level Interaction denotes that each task interacts once, whereas Action-Level Interaction represents the frequency of interactions that occur through the generation of actions for control purposes.

Benchmark	Input Modality	Output Modality	Based Method	Stage	Interactive Env.	Evaluation Strategy
AgentBench [10]	Text	Text	LLM	S_0	Task-Level	Human Judgement
EgoPlan-Bench [2]	Text & Images	Text	MLLM	S_0	N/A	Multi-choice
MMWorld [6]	Text & Images	Text	MLLM	S_0	N/A	GPT Judgement
VAB [11]	Text & Images	Text	MLLM	S_0	Task-Level	Human Judgement
LEGO [8]	Text & Images	Image	IGM	S_1	Task-Level	Feature Similarity
VBench [7]	Text	Video	VGM	S_2	N/A	Feature Similarity
EvalCrafter [12]	Text & Images	Video	VGM	S_2	N/A	Feature Similarity
WorldSimBench	Text & Images	Actionable Video	VGM	S_3	Action-Level	Human Preference Evaluator Embodied Metric

from developing agents to solve inference tasks to leveraging predictions for driving robots to perform specific actions.

Nevertheless, the rich application scenarios and diverse model designs make predictive models a broad family. However, without categorizing them based on their inherent characteristics, the advancement of predictive model development remains limited. This leads to our first question: *Can we establish a reasonable hierarchical system for Predictive Models based on their output modality?* With a well-defined categorization, we can better target the evaluation of Predictive Models from different perspectives in diverse embodied environments, ensuring that their strengths and weaknesses are adequately assessed. In the literature, existing evaluations have typically focused on task planning capabilities by assessing text outputs or evaluating visual outputs from an aesthetic perspective. However, such approaches significantly limit the evaluation of highly embodied Predictive Models, as embodied scenarios are more concerned with physical properties (*e.g.*, perspective consistency, object breakability), which these methods fail to effectively assess. This brings us to our second question: *Can we conduct a more detailed evaluation of highly embodied Predictive Models from an embodied perspective?*

To answer the first question, we categorize the functionalities of Predictive Models into a hierarchy from S_0 to S_3, defined by the model's capabilities and output modality, accompanied by corresponding evaluation benchmarks as illustrated in Fig. 1. Models are classified based on the output modality in their output modalities. From lower to higher stages, the models are capable of generating: text, images, videos, and actionable videos (*i.e.*, the videos that can be translated into actions). It is worth noting that Predictive Models at S_3 capable of generating actionable videos integrate robust 3D scene understanding and physical rule priors to provide precise guidance for generating executable actions. These models are closely aligned with the recently proposed concept of World Simulators [13].

To answer the second question, we review the related benchmarks, as listed in Tab. 1. Evaluations on models in S_0 that generate text primarily focus on assessing task planning capabilities, while S_1 and S_2 assessments on visual output measure aesthetic quality through feature similarity analyses with ground truth data. With clearly defined evaluation dimensions and extensive annotated datasets, both types of assessments can be effectively conducted. However, evaluating World Simulators introduces complexities due to the intricate physical definitions involved. Additionally, conventional evaluation methods are inadequate for assessing the actionablilty of the generated videos, as there is no definite ground truth for actionable videos towards completing a specific embodied task. These factors pose significant challenges to the evaluation of World Simulators.

We argue that an evaluation aligned with human perception could provide a more intuitive and accurate reflection of the characteristics of the synthesized videos, including their adherence to physical rules. Besides, the actionability can be assessed through a closed-loop manner in simulations deployed with a unified video-to-action policy network. Considering these aspects, we take the very first step in evaluating World Simulators by proposing a dual evaluation framework called WorldSimBench. As shown in Fig. 1, WorldSimBench assesses World Simulators through two complementary approaches: **Explicit Perceptual Evaluation**, which focuses on the Visual Quality, Condition consistency, and Embodiment of the generated content, and **Implicit Manipulative Evaluation**, which measures the World Simulator's performance through the conversion of video into control signals. We present three representative embodied scenarios: Open-Ended Embodied Environment (OE), Autonomous Driving (AD), and Robot Manipulation (RM), to thoroughly evaluate the capability of World Simulators in generating and representing scenario-specific attributes.

In the Explicit Perceptual Evaluation, we first define evaluation criteria which is used to construct a comprehensive set of prompts specific to each scenario. The prompt lists are then used by various video generation models to produce a large number of video clips. Following extensive human feedback and annotation, these video clips are compiled into the HF-Embodied dataset which consists of a total of 35,701 tuples with multi-dimensional scores and fine-grained hu-

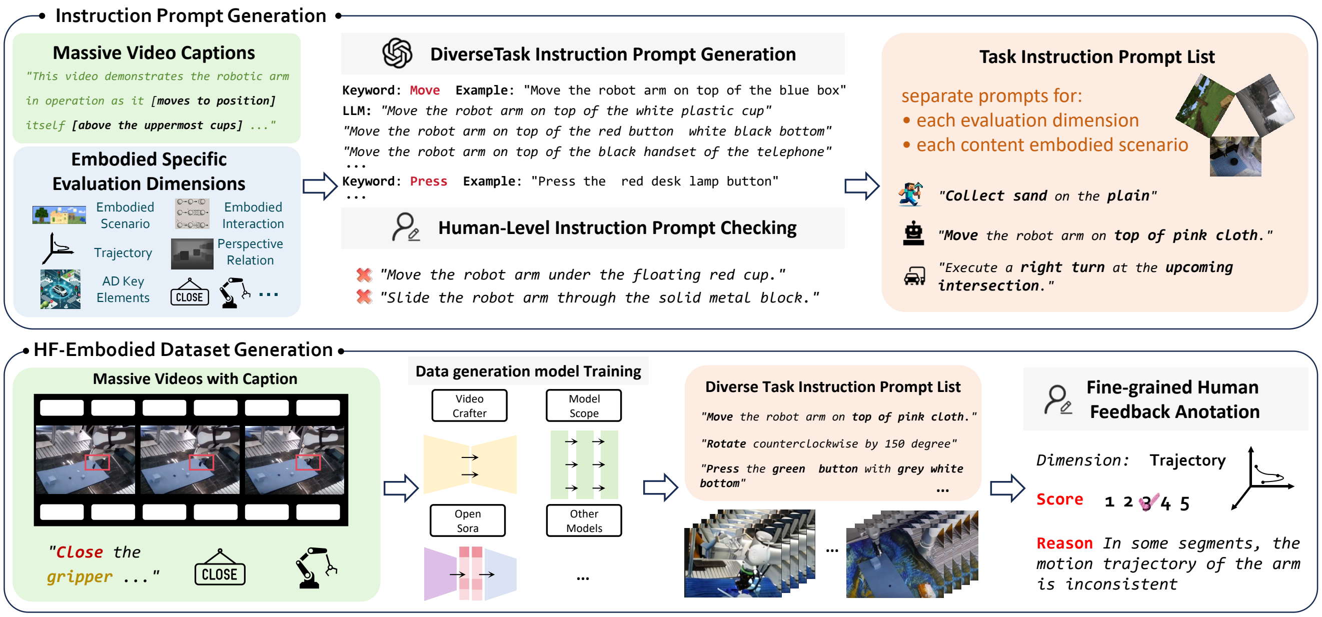

Figure 2. Overview of **Explicit Perceptual Evaluation**. (Top) **Instruction Prompt Generation.** We use a large collection of video captions from the internet and our predefined embodied evaluation dimensions. These are expanded using GPT and manually verified to create a corresponding Task Instruction Prompt List for data generation and evaluation. (Bottom) **HF-Embodied Dataset Generation.** Massive internet-sourced embodied videos with captions are used to train data generation models. Fine-grained Human Feedback Annotation is then applied to the embodied videos according to the corresponding Task Instruction Prompt List, covering multiple embodied dimensions.

man feedback. Additionally, we train Human Preference Evaluator, using the HF-Embodied dataset to assess World Simulators at the perceptual level, offering a robust evaluation of both their visual fidelity and contextual accuracy. For the Implicit Manipulative Evaluation, we deploy three simulation environments for the three embodied scenarios respectively. These environments are used to collect data and train inverse dynamic or goal-based video-to-action models capable of mapping future videos to actions. In each of these embodied scenarios, the World Simulator is tasked with generating situation-aware videos in real-time, based on current observations and provided text instructions. These generated videos are then converted into actions using the pre-trained video-to-action models. The effectiveness of the World Simulator is implicitly evaluated by measuring the performance of the tasks, using relevant metrics to reflect the quality and accuracy of the generated video.

In summary, the main contributions are as follows: (1)We categorize the functionalities of Predictive Models into a hierarchy, defined by the model's capabilities and output modality, to advance research and development in the field and take the very first step in evaluating World Simulators. (2)We propose a dual evaluation framework called WorldSimBench, through Explicit Perceptual Evaluation and Implicit Manipulative Evaluation, we conducted a comprehensive evaluation of the World Simulator's capabilities from an embodied perspective, focusing on both the visual and action levels. (3)We conducted extensive testing across multiple models and performed a thorough analysis of the experimental results. Our findings highlight the strengths and limitations of current World Simulators and provide actionable insights for improving future video generation models. (4)We developed HF-Embodied Dataset, which includes fine-grained human feedback across three scenarios and 20 dimensions, with a total of 35,701entries. This dataset, containing both human ratings and the reasons behind them, not only enables the evaluation of World Simulators but also provides broader applications (*e.g.*,alignment) for future video generation models.

2. WorldSimBench Construction

WorldSimBench evaluates the embodied capabilities of World Simulators across two distinct levels. The **Explicit Perceptual Evaluation** in Fig. 2 assesses the simulators based on human-perceived quality across different embodied scenarios, while the **Implicit Manipulative Evaluation** in Fig. 3 implicitly evaluates the simulators' capabilities by converting the generated videos into control signals and observing their performance in various closed-loop embodied tasks.

The evaluation of World Simulators encompasses three critical embodied scenarios: Open-Ended Embodied Environment (OE), Autonomous Driving (AD), and Robot Manipulation (RM). Minecraft serves as a popular testbed for OE, providing a challenging platform for agents to handle complex, unstructured tasks. In the context of AD, especially in outdoor settings, ensuring the stability and robustness of

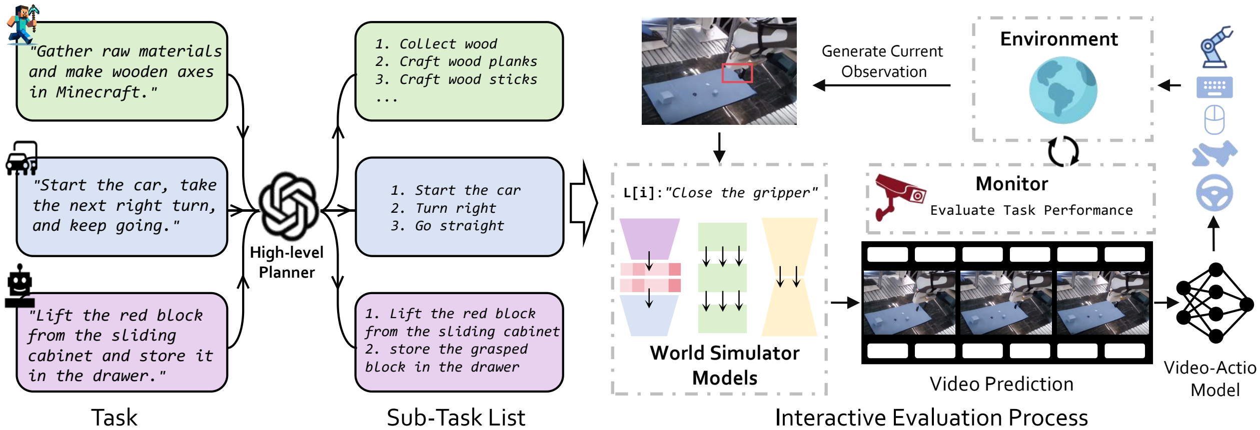

Figure 3. Overview of **Implicit Manipulative Evaluation**. Embodied tasks in different scenarios are decomposed into executable sub-tasks. The video generation model generates corresponding predicted videos based on the current instructions and real-time observations. Using a pre-trained IDM or a goal-based policy, the agent executes the generated sequence of actions. After a fixed timestep, the predicted video is refreshed by sampling again from the video generation model, and this process repeats. Finally, the success rates of various embodied tasks are obtained through monitors in the simulation environment.

the agent's actions is crucial, making it an essential domain for assessing a World Simulator's capability in dynamic and uncertain environments. RM, a core task in embodied intelligence, demands precise and adaptive control, testing the world simulator's ability to generate actionable predictions that align with physical interactions. Together, these scenarios provide a comprehensive benchmark for evaluating the effectiveness of World Simulators across a range of real-world tasks.

References

[1] Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pretrained image-editing diffusion models. *arXiv preprint arXiv:2310.10639*, 2023. 1

[2] Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. Egoplan-bench: Benchmarking egocentric embodied planning with multimodal large language models. *arXiv preprint arXiv:2312.06722*, 2023. 2

[3] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 1

[4] Yilun Du, Mengjiao Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B Tenenbaum, et al. Video language planning. *arXiv preprint arXiv:2310.10625*, 2023. 1

[5] Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[6] Xuehai He, Weixi Feng, Kaizhi Zheng, Yujie Lu, Wanrong Zhu, Jiachen Li, Yue Fan, Jianfeng Wang, Linjie Li, Zhengyuan Yang, et al. Mmworld: Towards multi-discipline multi-faceted world model evaluation in videos. *arXiv preprint arXiv:2406.08407*, 2024. 2

[7] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 2

[8] Bolin Lai, Xiaoliang Dai, Lawrence Chen, Guan Pang, James M Rehg, and Miao Liu. Lego: Learning egocentric action frame generation via visual instruction tuning. *arXiv preprint arXiv:2312.03849*, 2023. 1, 2

[9] Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18061–18070, 2024. 1

[10] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*, 2023. 2

[11] Xiao Liu, Tianjie Zhang, Yu Gu, Iat Long Iong, Yifan Xu, Xixuan Song, Shudan Zhang, Hanyu Lai, Xinyi Liu, Hanlin Zhao, et al. Visualagentbench: Towards large multimodal models as visual foundation agents. *arXiv preprint arXiv:2408.06327*, 2024. 2

[12] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22139–22149, 2024. 2

[13] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023. 2