



Target-Bench: Can Video World Models Achieve Mapless Path Planning with Semantic Targets?

Dingrui Wang^{1,2,*}, Zhihao Liang^{1,*}, Hongyuan Ye^{1,*}, Zhexiao Sun^{1,*},
Zhaowei Lu¹, Yuchen Zhang¹, Yuyu Zhao¹, Yuan Gao¹, Marvin Seeger¹,
Finn Schäfer¹, Haotong Qin³, Wei Li^{4,5}, Luigi Palmieri², Felix Jahncke¹,
Mattia Piccinini¹, and Johannes Betz¹

¹Technical University of Munich, Munich, Germany

{dingrui.wang,zhihao.liang,hongyuan.ye,zhexiao.sun,johannes.betz}@tum.de

²Bosch Research, Stuttgart, Germany

³ETH Zurich, Zurich, Switzerland

⁴Nanjing University, Nanjing, China

⁵EABOT.AI, Shanghai, China

<https://target-bench.github.io/>

Abstract. While recent video world models can generate highly realistic videos, their ability to perform semantic reasoning and planning remains unclear and unquantified. We introduce **Target-Bench**, the first benchmark that enables comprehensive evaluation of video world models’ semantic reasoning, spatial estimation, and planning capabilities. Target-Bench provides 450 robot-collected scenarios spanning 47 semantic categories, with SLAM-based trajectories serving as motion tendency references. Our benchmark reconstructs motion from generated videos with a metric scale recovery mechanism, enabling the evaluation of planning performance with five complementary metrics that focus on target-approaching capability and directional consistency. Our evaluation result shows that the best off-the-shelf model achieves only a 0.368 overall score, revealing a significant gap between realistic visual generation and semantic reasoning in current video world models. Furthermore, we demonstrate that fine-tuning process on a relatively small real-world robot dataset can significantly improve task-level planning performance.

Keywords: World models · Semantic reasoning · Tasking benchmark

1 Introduction

Embodied AI has been advancing rapidly, while its core challenges are becoming increasingly clear. Despite major improvements in robotic hardware (e.g. actuators), the embodied intelligence remains the main bottleneck limiting robots

* Equal contribution.

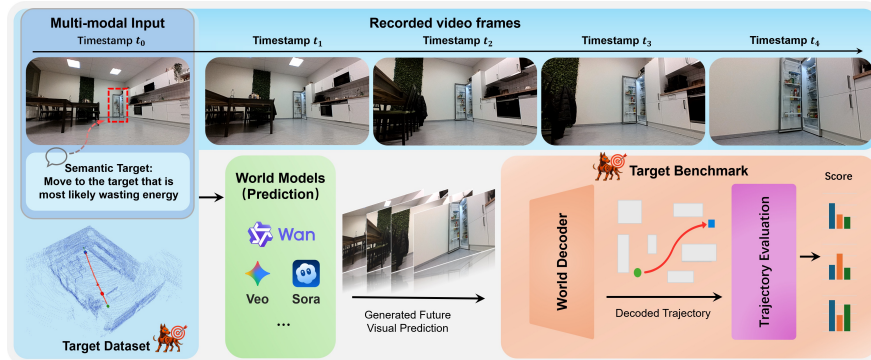


Fig. 1: A Video World Model’s planning ability is evaluated by our **benchmark**. We provide a diverse **dataset** containing annotated semantic goal, task description and motion reference to establish this novel benchmark. The generated future visual predictions are processed by our **world-decoder**, and the results are then evaluated with meticulously designed metrics that focus on tendency rather than trajectory overlap.

from realizing their full potential [27]. As a result, the robotics community is seeking to develop embodied AI systems that are more robust and generalizable.

World Models (WMs) learn to predict how the world evolves over time [12,21]. In early 2024, Sora [6], a video generation model, was the first to self-claim as a world evolution simulation model. Recent breakthroughs in video models demonstrate remarkable reasoning abilities in visual semantics and causality [6, 11, 41]. In this regard, these video models can also be regarded as world models [5]. Given an initial observation (e.g., an image frame) and a condition (e.g., a text prompt), video WMs such as Veo 3.1 [8] and Sora 2 [35] can generate future frames with photorealistic quality and high spatio-temporal consistency.

Unsurprisingly, video WMs have sparked growing interest from the robotics community [22, 31, 39, 46]. Recent approaches such as UnifoLM-WMA-0 [36] aim to integrate video WMs with low-level robot control. The underlying philosophy of these approaches is that if a model can accurately predict how the world evolves, its predictions can serve as a plan to guide the robot actions. However, a key question remains: How accurate must these predictions be to count as useful for planning? More broadly, *how can we quantitatively assess a video WM’s reasoning, task-solving, and planning ability?* Existing evaluation frameworks focus mainly on visual fidelity and spatio-temporal consistency [10, 23], while assessing mapless path planning that require strong semantic understanding remains an elusive task [27]. Bridging this gap requires a benchmark that can holistically evaluate video world models’ semantic reasoning beyond visual quality and spatio-temporal consistency.

To address this challenge, we introduce **Target-Bench**. To the best of our knowledge, it is the first benchmark for evaluating video WMs for mapless path planning toward semantic targets in unstructured real-world environments. The

semantic targets are described in natural language, and many cases are designed to include implicit semantic meanings (Fig. 1). Our contributions are as follows:

- The first systematic **benchmark** for evaluating video world models on map-less path planning with textual semantic goals, featuring a modular and replaceable world-decoder and a meticulously designed set of metrics, which makes comprehensive evaluation of video world models’ semantic reasoning, spatial estimation, and planning capabilities possible.
- An **open-source dataset** of 450 semantic-target-oriented scenarios collected with a quadruped robot, covering 47 categories across diverse indoor and outdoor environments, with ground truth trajectories and human annotations for explicitly and implicitly described targets.
- A comparative study of open-source and proprietary WMs, including the first **fine-tuning** of an open-source model on a small real-world robot dataset for path planning, showing that data-efficient adaptation can substantially boost task-level planning performance. We open-source all code and dataset.

2 Related Work

World Models. Early approaches such as World Models [12] and Dreamer [13] introduced latent state-space models, demonstrating that generative prediction has the potential to support planning and control [11]. More recently, foundation-scale methods pretrained on large video corpora have emerged: Cosmos [30] compared diffusion and autoregressive paradigms; DIAMOND [2] improved visual fidelity; V-JEPA 2 [4] enhanced efficiency by forecasting in latent space. Advances in video WMs, including Sora 2 [35], Veo 3.1 [8, 41], and Wan [38], have further improved prompt adherence and physics-aware dynamics, enabling long-horizon predictions. Interactive WMs such as Genie-1 [7], Genie-2 [32], and Genie-3 [5] provide controllable environments for use in robotics and games. Uni-tree’s recent UnifoLM-WMA-0 [36] aims to extract the reasoning capabilities of WMs and translate them into real-world actions. However, the ability of WMs to perform robot task planning in real-world scenarios is largely untested.

Semantic Navigation Datasets. Navigation datasets such as R2R [3] and RxR [19] are primarily restricted to indoor simulation environments. To address real-world complexity, datasets such as MuSoHu [29], EgoWalk [1], and SANPO [37] capture multimodal data via wearable sensors, though the latter faces heading misalignment and IMU drift. Others utilize human videos or robot demonstrations for language-conditioned policies (LeLaN [14]) and socially compliant interaction (SCAND [18], SACSoN [15]). For broader coverage, Sekai [25] and CityWalker [26] leverage web-scale videos to achieve global diversity, although trajectories are only annotated up-to-scale. Crucially, as compared in Table 1, these datasets lack explicit navigation targets, and their annotations are descriptive rather than target-oriented.

Benchmarks for World Models. Evaluation of WMs has transitioned from assessing visual quality and camera controllability (VBench [17], WorldScore [10])

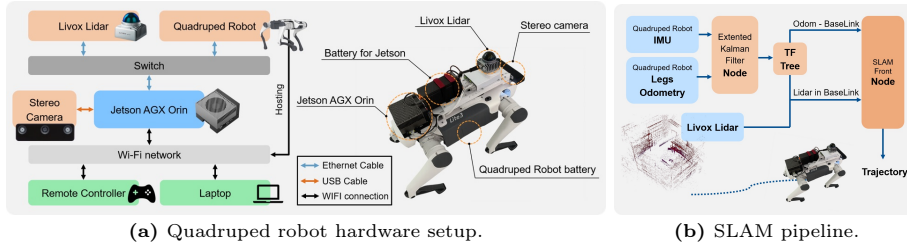


Fig. 2: Robot setup and SLAM pipeline.

and basic physics (WorldModelBench [23]) to functional utility. While frameworks like World-in-World [45] and WorldArena [34] enable closed-loop testing, they are predominantly confined to simulation. Crucially, many recent robotics-oriented suites (RBench [9], EWMBench [44], WorldEval [24]) rely on subjective Vision Language Model (VLM)-based judges for success detection, which introduces inherent judge uncertainty and limits result reproducibility. Moreover, existing benchmarks typically focus on explicit semantic instruction rather than implicit semantic reasoning. A concise comparison across these dimensions is provided in Table 2.

3 Target-Bench

As shown in Fig. 1, Target-Bench consists of two components: a Target dataset and the Target benchmark evaluation pipeline. The **Target dataset** (Fig. 2a) is collected with a quadrupe robot equipped with multi-modal sensors in diverse indoor and outdoor environments. It spans 47 semantic target categories (e.g., cars, chairs, bicycles) and includes 450 scenarios. Our **Target benchmark** comprises two stages: world-decoder and path evaluation. The world-decoder extracts camera poses from generated video to form a path, which is then compared against the ground truth path. Path evaluation metrics focus on the proximity to the target and directional consistency, jointly assessing scenario reconstruction quality and semantic navigation capability.

3.1 Target Dataset Setup

Quadrupe Robot Platform As shown in Fig. 2a, our data-collection platform is built on a DEEP Robotics Lite 3 Venture quadrupe robot. It carries a Livox Mid-360 LiDAR, an OAK-D Pro W stereo RGB camera, and an NVIDIA Jetson AGX Orin for mapping and state estimation. The LiDAR connects to the robot base via an Ethernet switch, and the camera connects to the Jetson via USB. The Jetson is powered by a dedicated onboard battery. A Wi-Fi link connects the Jetson to a remote laptop for monitoring and logging, and to a handheld controller for teleoperation. Our software stack (Fig. 2b) uses a

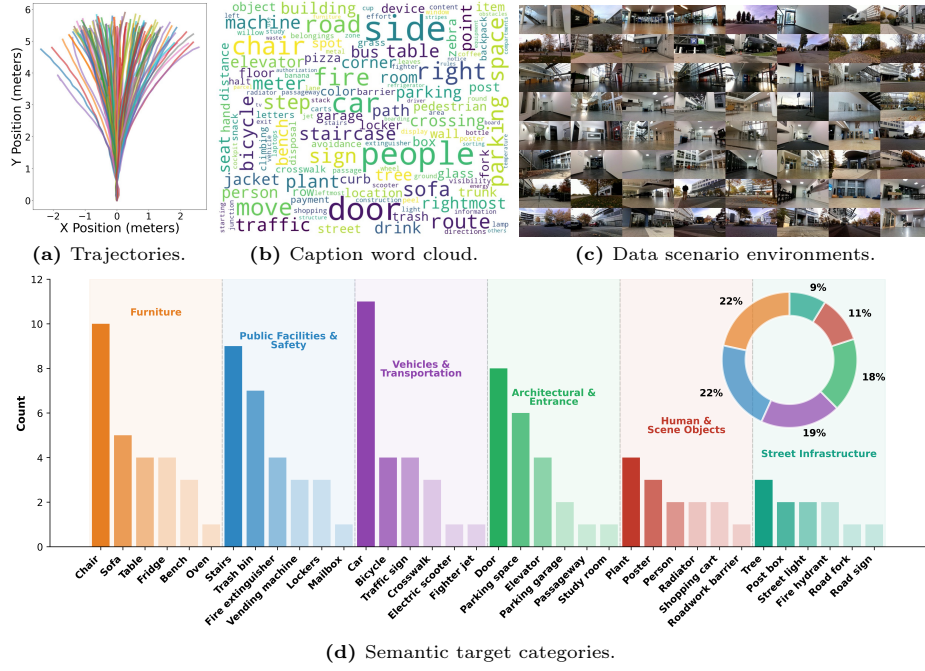


Fig. 3: Visualization of benchmark dataset structure and semantics: (a) spatial distribution of trajectories, (b) word cloud representation of caption vocabulary, (c) data collection environments, and (d) distribution of semantic target categories.

LiDAR-centric SLAM pipeline [20,33] with multi-sensor fusion. Inertial Measurement Unit (IMU) data and legged odometry are fused via an Extended Kalman Filter (EKF) to produce a stable base-frame pose, which is broadcast through the ROS TF tree to align the LiDAR and robot frames. The LiDAR point clouds are processed by the Simultaneous Localization and Mapping (SLAM) front-end with motion compensation and incremental registration guided by the fused odometry. The back-end then optimizes the trajectory and builds a global map, enabling accurate pose estimation.

Semantic Target Data Collection The data collected in this work is to support the evaluation of video world models on mapless semantic navigation tasks. The objective is to assess whether a model can infer a plausible motion trend toward a specific semantic target in the scene. Therefore, the robot dog is always initialized with the target already present in the field of view. To guarantee a clear and unambiguous motion trend, a human operator manipulates the robot using a direct goal-approach strategy, maintaining constant forward speed and minimizing unnecessary heading oscillations. This avoids exploration and results in stable and reproducible motion patterns across different scenes.

Table 1: Comparison with existing navigation datasets. Our dataset is collected via target-oriented teleoperation of a real quadruped robot in real-world environments, providing video sequences paired with semantic target prompts, point cloud maps, and metric-scale trajectories with kinematic consistency.

Dataset	Source Type	Metric Traj.	Semantic Target	Point Cloud	Kinem. Match
R2R [3]	Simulation	✗	✓	✗	✗
RxR [19]	Simulation	✗	✓	✗	✗
LeLaN [14]	Real-Rob./Human	✗	✓	✗	○
MuSoHu [29]	Real-Human	✓	✗	✓	✗
SACSoN [15]	Real-Robot	✓	✗	✗	✓
SANPO [37]	Sim./Real	✓	✗	✗	✗
Sekai [25]	Sim./Real	✗	✗	✗	✗
CityWalker [26]	Real-Rob./Human	○	✗	✗	○
SCAND [18]	Real-Robot	✓	✗	✓	✓
EgoWalk [1]	Real-Human	✓	✗	✗	✗
Ours	Real-Robot	✓	✓	✓	✓

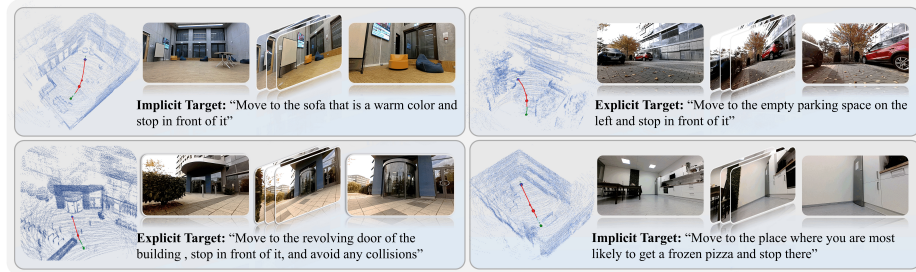


Fig. 4: Data Sample Visualization.

Each processed data sample in the Target dataset consists of four components: a video sequence, a ground truth trajectory, a semantic target specified in the form of a prompt, and a point cloud map. Video frames are captured at 25Hz, yielding approximately 10 seconds of continuous observation per sample. The first 2 seconds are used for scale factor computation in Sec. 3.2, the latter 8 seconds are used as future frames for model inference and training. Each prompt is a concise imperative command expressed in natural language, ensuring that the navigation outcome can be judged with respect to a single well-defined goal in the scene. The semantic targets are pre-selected by human experts to ensure diversity and relevance for navigation tasks. Each target is defined as either *explicit* or *implicit*. Explicit targets specify the object name directly, whereas implicit targets describe the object through its attributes or functions embedded in the prompt context (Fig. 4). Among the 450 collected samples, 232 are indoors and 218 are outdoors, 72 have implicit targets and 378 have explicit targets. 125 samples are curated as the benchmark set based on scene/target diversity and a balanced distribution of left/forward/right motion trends, while the remain-

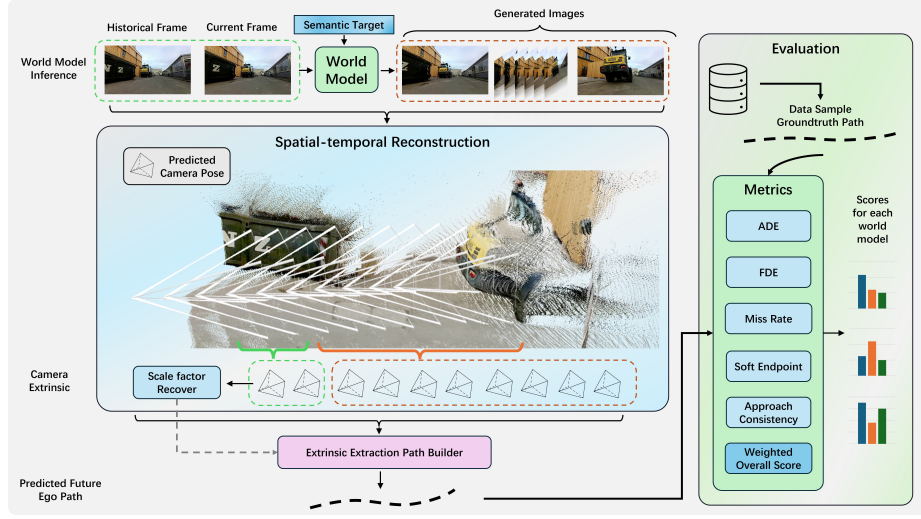


Fig. 5: Target Benchmark Architecture.

ing 325 samples are used exclusively as training set. Fig. 3 demonstrates the benchmark dataset structure.

3.2 Target Benchmark Architecture

The Target benchmark (Fig. 5) consists of two main components designed to quantify world model planning capabilities. First, the **world-decoder** extracts camera trajectories from generated videos using state-of-the-art 3D reconstruction methods. Second, the **path evaluation** module computes a score focusing on approach tendency using a comprehensive suite of metrics.

World-Decoder

Spatio-temporal Reconstruction. As shown in Fig. 5, to extract camera trajectories from WM’s generated videos, we employ three state-of-the-art 3D reconstruction methods: VGGT [40], SpaTracker [42], and ViPE [16]. Each method processes video frames to reconstruct camera poses.

VGGT. Visual Geometry Grounded Transformer (VGGT) is a feed-forward transformer that directly predicts camera poses, depth maps, and point clouds from multi-view images. Given a sequence of S images $\mathcal{I} = \{I_1, I_2, \dots, I_S\}$, VGGT first encodes them through a vision transformer to obtain aggregated tokens. A camera head then predicts a pose encoding $\mathbf{p}_s = [\mathbf{T}_s; \mathbf{q}_s; \mathbf{fov}_s] \in \mathbb{R}^9$ for each frame I_s , where $\mathbf{T}_s \in \mathbb{R}^3$ is the camera translation vector, $\mathbf{q}_s \in \mathbb{R}^4$ is the rotation quaternion, and $\mathbf{fov}_s \in \mathbb{R}^2$ represents the horizontal and vertical field of view. This encoding is converted to standard camera parameters $\mathbf{E}_s = [\mathbf{R}(\mathbf{q}_s) \mid \mathbf{T}_s] \in \mathbb{R}^{3 \times 4}$, where $\mathbf{R}(\mathbf{q}_s)$ converts the quaternion to a rotation matrix.

Table 2: Comparison of Video World Model Benchmarks. Target-Bench provides invariant, task-driven evaluation of world-model planning in real-world robot scenes, and includes tasks that require implicit semantic reasoning, without relying on subjective model-based scoring, thus reducing judge uncertainty.

Benchmark	Task-Driven	Real-Rob. Scene	Invari. Eval.	Visual Consist.	Spatial Estim.	Implicit Reasoning
VBench [17]	✗	✗	✗	✓	✗	✗
WorldModelBench [23]	✓	○	✗	✗	✗	✗
EWMBench [44]	✓	✓	✗	✓	✓	✗
WorldEval [24]	✓	✓	✗	✗	✗	✗
WorldScore [10]	✗	✗	○	✓	✓	✗
World-in-World [45]	✓	✗	✗	✓	✗	✗
WorldArena [34]	✓	✗	○	✓	✓	✗
RBench [9]	✓	✓	✗	✓	✗	✗
Target-Bench	✓	✓	✓	✓	✓	✓

SpaTracker. SpaTracker extends VGGT with tracking for better temporal consistency. It adopts a two-stage design: (1) VGGT4Track predicts initial poses $\{\mathbf{E}_s^{(0)}\}_{s=1}^S$ and depths $\{D_s\}_{s=1}^S$; (2) a tracking module refines these poses via point correspondences across frames. Given query points $\mathbf{Q} \in \mathbb{R}^{N \times 3}$ on the first frame, the tracker optimizes camera poses through bundle adjustment: $\{\mathbf{E}_s^*\} = \arg \min_{\{\mathbf{E}_s\}} \sum_{s=1}^S \sum_{n=1}^N \rho(\|\pi(\mathbf{K}_s \mathbf{E}_s \mathbf{X}_n) - \mathbf{x}_{n,s}\|^2)$, where $\mathbf{X}_n$ is the 3D position of point n , $\pi(\cdot)$ the projection, $\mathbf{x}_{n,s}$ the tracked 2D point, and $\rho(\cdot)$ a robust loss. The result is a refined camera-to-world trajectory $\{\mathbf{C}_{2W,s}\}_{s=1}^S$, where each $\mathbf{C}_{2W,s} \in \mathbb{R}^{4 \times 4}$ is the full camera-to-world transform.

ViPE. ViPE is a SLAM-based visual-inertial pipeline providing metric-scale poses by fusing visual and inertial data. Given RGB frames and IMU inputs, it performs incremental pose estimation with loop closure. Unlike VGGT and SpaTracker, ViPE directly outputs paths in metric units via sensor fusion without requiring scale recovery. The final path is a sequence of SE(3) transformations $\{\mathbf{T}_s\}_{s=1}^S$, where each $\mathbf{T}_s \in \mathbb{R}^{4 \times 4}$ encodes rotation and translation in meters.

Scale Factor Recovery. Monocular methods such as VGGT and SpaTracker estimate camera motion only up to an unknown scale. We restore metric consistency at the segment level by anchoring predictions to a single scalar scale factor λ derived from real displacement d_{real} using the first and last ones from the scale factor recovery frames that result in two predicted extrinsic matrices $\mathbf{E}_f, \mathbf{E}_l \in \mathbb{R}^{3 \times 4}$ which are corresponding to the two predicted camera poses enclosed by green dotted box in Fig. 5. We extract the translation vectors $\mathbf{t}_f$ and $\mathbf{t}_l$ as the fourth column of $\mathbf{E}_f$ and $\mathbf{E}_l$. The predicted displacement is $d_{\text{pred}} = \|\mathbf{t}_f - \mathbf{t}_l\|_2$, and the scale factor is $\lambda = d_{\text{real}}/d_{\text{pred}}$. Then for all other predicted future frames, we rescale their predicted translation vectors together, $\mathbf{t}_s^{\text{scaled}} = \lambda \mathbf{t}_s$ for $s = 1, \dots, S$ with S be the total number of all frames. This lightweight mechanism preserves relative geometry while lifting trajectories to metric values, enabling quantitative comparison.

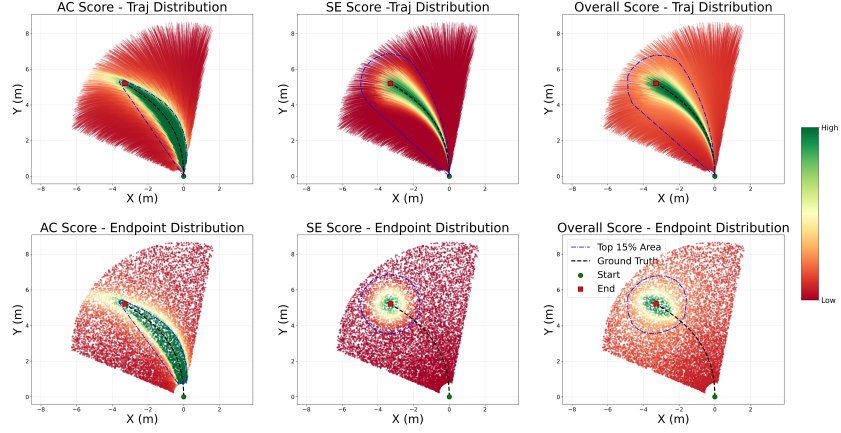


Fig. 6: Illustration of metrics AC and SE with trajectory and endpoint score distribution heatmaps. Plots visualize scores over a fixed set of candidate trajectories. Trajectories with good Approach Consistency can have a bad Endpoint (Far away from the target). Trajectories with a good Endpoint can have a bad Approach Consistency (Not within the oval corridor). Trajectories with a good Overall Score have both good Approach Consistency and Endpoint.

Path Evaluation

Metrics. We evaluate the planning results of video world models using a comprehensive suite of metrics that assess accuracy, goal-reaching capability, and path consistency. All metrics assume 2D trajectories represented as sequences of positions, with ground truth $\mathbf{s}^{\text{GT}} = \{s_1, s_2, \dots, s_T\}$ and prediction $\hat{\mathbf{s}} = \{\hat{s}_1, \hat{s}_2, \dots, \hat{s}_T\}$, where $s_t, \hat{s}_t \in \mathbb{R}^2$ and T is the number of time steps.

Average Displacement Error (ADE). Measures the average L2 distance between predicted and ground truth positions across all timesteps: $\text{ADE} = \frac{1}{T} \sum_{t=1}^T \|\hat{s}_t - s_t\|_2$.

Final Displacement Error (FDE). Evaluates the distance between final positions: $\text{FDE} = \|\hat{s}_T - s_T\|_2$.

Miss Rate (MR). Computes the percentage of predicted points that deviate beyond a threshold τ (default: 1.0 m): $\text{MR} = \frac{1}{T} \sum_{t=1}^T \mathbb{I}[\|\hat{s}_t - s_t\|_2 > \tau] \cdot 100$

Soft Endpoint (SE). Uses a Gaussian penalty to measure the proximity of the endpoint to the target:

$$\text{SE} = \exp\left(-\frac{\|\hat{s}_T - s_T\|_2^2}{2\sigma^2}\right) \quad (1)$$

where $\sigma = 0.5$ m is the tolerance. $\text{SE} \in [0, 1]$ where 1 indicates perfect alignment.

Approach Consistency (AC). Evaluates whether the predicted trajectory stays within a progress-dependent corridor around the ground truth path. We

uniformly sample $M = 20$ reference points along the ground truth trajectory and assign the corridor radius at the i -th reference point as σ_i :

$$\sigma_i = \sigma_{\min} + (\sigma_{\max} - \sigma_{\min}) \exp\left(-\frac{(p_i - 0.5)^2}{2\beta^2}\right) \quad (2)$$

where $p_i = i/(M - 1)$ is the normalized progress, $\sigma_{\min} = 0.5$ m and $\sigma_{\max} = 1.5$ m are the minimum and maximum corridor radius, and $\beta = 0.25$. A predicted point $\hat{s}_j$ is covered if $\min_i \|\hat{s}_j - s_i^{\text{GT}}\|_2 \leq \sigma_i$. The AC score is given as:

$$\text{AC} = \begin{cases} 1, & N_c = N_p \\ \exp\left(-\gamma \cdot \frac{N_p - N_c}{N_p}\right), & \text{otherwise} \end{cases} \quad (3)$$

where N_p is the total number of predicted points, N_c is the number covered by the corridor, and $\gamma = 5$. Importantly, this metric design supports non-uniqueness of valid path. The ground truth path is treated as a rough reference rather than a single correct solution, conditioned on the initial heading. As shown in Fig. 6, various trajectories with similar motional tendency can achieve the same level of high score, which means the score is tendency based, instead of trajectory based.

Weighted Overall (WO) Score. Aggregates all metrics into a unified score $\in [0, 1]$ (the higher the better):

$$\begin{aligned} S_{\text{overall}} = & w_{\text{ADE}} \cdot \exp\left(-\frac{\text{ADE}}{\tau_{\text{ADE}}}\right) + w_{\text{FDE}} \cdot \exp\left(-\frac{\text{FDE}}{\tau_{\text{FDE}}}\right) \\ & + w_{\text{MR}} \cdot \left(1 - \frac{\text{MR}}{100}\right) + w_{\text{SE,AC}} \cdot \text{SE} \cdot \text{AC} \end{aligned} \quad (4)$$

with default weights: $w_{\text{ADE}} = 0.075$, $w_{\text{FDE}} = 0.125$, $w_{\text{MR}} = 0.125$, $w_{\text{SE,AC}} = 0.675$, and scale parameters $\tau_{\text{ADE}} = \tau_{\text{FDE}} = 1.0$ m. The weight design prioritizes task success; therefore, a dominant weight is assigned to SE·AC to emphasize endpoint proximity and approach consistency, while keeping ADE, FDE, and MR lightweight to avoid over-emphasizing point-wise trajectory overlapping.

4 Experiments

Our proposed evaluation framework uses five key metrics detailed in Sec. 3.2. The analysis includes several off-the-shelf state-of-the-art video world models, including Sora 2 [35], Veo 3.1 [8], Veo 3.1-fast, and multiple variants of the Wan series (Wan2.5-I2V-Preview, Wan2.2-Plus, Wan2.2-Flash, Wan2.1-Plus, and Wan2.1-Turbo) [38]. In addition, we evaluate the fine-tuned Wan2.2-TI2V-5B models.

4.1 Implementation Details

For model fine-tuning we use the LoRA training framework offered by DiffSynth-Studio [28], with $8 \times$ NVIDIA A800 80 GB GPUs. For inference of the fine-tuned

Table 3: Evaluation results with VGGT. [†] *Image-to-Video (I2V) models*, [§] *fine-tuned*, [¶] *fine-tuned with data augmentation*.

Spatial Method	World Model	Open Source	Video Length (s) & Resolution	Metrics					WO $\uparrow$
				FDE $\downarrow$	ADE $\downarrow$	MR $\downarrow$	SE $\uparrow$	AC $\uparrow$	
VGGT [40]	gt_video	-	8 - 720p	0.231	0.120	0.39	0.879	0.978	0.882
	Sora 2	$\times$	8 - 720p	1.915	1.079	41.43	0.161	0.794	0.232
	Veo 3.1	$\times$	8 - 720p	3.129	2.006	51.19	0.212	0.448	0.250
	Veo 3.1-fast	$\times$	8 - 720p	2.860	1.598	45.68	0.139	0.689	0.212
	Wan2.5-Preview [†]	$\times$	10 - 720p	2.476	1.459	48.01	0.182	0.869	0.238
	Wan2.2-Plus [†]	$\times$	5 - 1080p	1.377	0.771	25.89	0.240	0.699	0.324
	Wan2.2-Flash [†]	$\times$	5 - 720p	1.364	0.753	24.49	0.292	0.745	0.368
	Wan2.1-Plus [†]	$\times$	5 - 720p	2.784	1.547	50.20	0.090	0.688	0.163
	Wan2.1-Turbo [†]	$\times$	5 - 720p	4.243	2.169	77.27	0.000	1.000	0.040
	Wan2.2-5B-base	$\checkmark$	8 - 720p	3.667	1.876	68.99	0.013	0.916	0.068
	Wan2.2-5B-FT [§]	$\checkmark$	8 - 720p	1.318	0.635	19.39	0.262	0.797	0.357
	Wan2.2-5B-FT-DA [¶]	$\checkmark$	8 - 720p	1.046	0.481	11.39	0.333	0.829	0.428

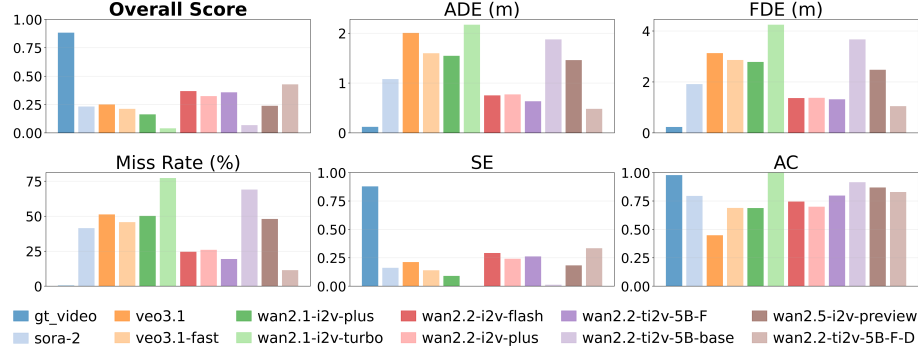


Fig. 7: World model performance comparison with VGGT as world-decoder’s spatio-temporal reconstruction tool.

models on the Target Benchmark we utilize one single NVIDIA RTX PRO 6000 Blackwell 96 GB GPU, while closed-source models are accessed directly through their official APIs. All generated videos are produced at 720p or 1080p resolution with durations ranging from 5 to 10 seconds. All evaluation experiments are conducted on a workstation equipped with an NVIDIA RTX 4090 GPU.

4.2 Evaluation for off-the-shelf Models

Table 3 shows the evaluation results using VGGT as the spatio-temporal reconstruction tool. Among all evaluated off-the-shelf models, Wan2.2-Flash achieves the best overall performance with a weighted overall score of 0.368. Specifically, it obtains the lowest errors in FDE (1.364m), ADE (0.753m), and MR (24.49%), while achieving the highest SE (0.292). Figure 7 visualizes the performance comparison across all models.

Table 4: Evaluation results with explicit and implicit semantic targets.

Data	World Model	Metrics					WO $\uparrow$
		FDE $\downarrow$	ADE $\downarrow$	MR $\downarrow$	SE $\uparrow$	AC $\uparrow$	
Explicit Target	gt_video	0.266	0.140	0.67	0.848	0.962	0.854
	Sora 2	1.984	1.163	44.37	0.166	0.758	0.229
	Veo 3.1	3.285	2.176	54.15	0.171	0.428	0.216
	Veo 3.1-fast	3.224	1.744	48.04	0.136	0.687	0.204
	Wan2.5-I2V-Preview	2.448	1.487	47.47	0.182	0.828	0.240
	Wan2.2-I2V-Plus	1.412	0.778	26.40	0.242	0.742	0.328
	Wan2.2-I2V-Flash	1.418	0.767	24.12	0.296	0.773	0.372
Implicit Target	gt_video	0.182	0.091	0.00	0.923	1.000	0.921
	Sora 2	1.819	0.960	37.30	0.154	0.845	0.235
	Veo 3.1	2.909	1.768	47.03	0.270	0.477	0.298
	Veo 3.1-fast	2.350	1.394	42.35	0.144	0.691	0.222
	Wan2.5-I2V-Preview	2.516	1.421	48.77	0.182	0.925	0.236
	Wan2.2-I2V-Plus	1.327	0.761	25.18	0.238	0.639	0.319
	Wan2.2-I2V-Flash	1.288	0.732	25.02	0.286	0.705	0.362

Displacement Errors: ADE values range from 0.7m to 2.1m across different models. Wan2.1-Turbo exhibits the highest displacement error (2.169m), while Wan2.2-Flash and Wan2.2-Plus achieve the best accuracy (around 0.75m). For FDE, errors show wider variation, with Wan2.1-Turbo displaying the worst FDE.

Reliability Metrics: The MR varies significantly from 24.49% to 77.27%. Wan2.1-Turbo shows the highest MR (77.27%), indicating poor trajectory quality, while Wan2.2-Flash achieves the best (24.49%). Most models score below 0.3 on the SE metric, suggesting challenges in reaching target endpoints accurately.

Consistency: Wan2.1-Turbo achieves the highest Approach Consistency (1.000), indicating no directional misalignment, while Veo 3.1 shows the poorest performance (0.448). Sora 2 demonstrates good consistency at 0.794.

Explicit vs. Implicit Targets: As shown in Table 4, performance on implicitly defined targets closely matches that on explicit targets, with small, model-specific fluctuations. This indicates that current video WMs can understand semantic goals even when they are not explicitly defined.

4.3 Fine-tuned Models

We use the train split from our dataset for fine-tuning the open-sourced Wan2.2-TI2V-5B model. The text prompt and the first video frame are taken as inputs. We evenly sample 121 frames from each video to form a ground truth frame sequence. The model is fine-tuned under two settings: without and with data augmentation. For data augmentation, we apply a shifting-frame strategy: for each video of a scenario, we generate four clips of 121 frames starting from different offsets, resulting in a four times large training set while preserving similar temporal coverage and reusing the same captions of each sample. Table 3

Table 5: Evaluation results of SpaTracker and ViPE. [†] *I2V models*, [§] *Wan2.2-TI2V-5B fine-tuned model*, [¶] *Wan2.2-TI2V-5B model fine-tuned with data augmentation*.

World Model	SpaTracker [42]						ViPE [16]					
	FDE↓	ADE↓	MR↓	SE↑	AC↑	WO↑	FDE↓	ADE↓	MR↓	SE↑	AC↑	WO↑
gt_video	0.526	0.600	17.19	0.59	0.95	0.623	1.345	0.845	30.22	0.19	0.89	0.291
Sora 2	1.913	1.257	51.33	0.14	0.82	0.208	2.147	1.245	51.04	0.09	0.85	0.170
Veo 3.1	2.607	2.133	78.49	0.20	0.42	0.194	2.916	1.945	55.07	0.16	0.62	0.218
Veo 3.1-fast	3.402	2.150	62.89	0.15	0.59	0.191	2.601	1.715	58.76	0.15	0.62	0.202
Wan2.5-Preview [†]	2.379	1.510	61.93	0.18	0.88	0.213	2.686	1.608	59.52	0.10	0.83	0.156
Wan2.2-Plus [†]	1.409	1.631	73.19	0.25	0.68	0.253	1.621	0.976	36.65	0.18	0.80	0.265
Wan2.2-Flash [†]	1.407	1.572	71.06	0.28	0.71	0.284	1.687	1.006	37.42	0.17	0.81	0.260
Wan2.1-Plus [†]	2.810	2.110	63.08	0.08	0.67	0.141	2.858	1.510	53.70	0.04	0.76	0.125
Wan2.1-Turbo [†]	4.219	1.833	63.45	0.00	1.00	0.061	4.256	1.856	64.35	0.00	1.00	0.059
Wan2.2-TI2V-5B	3.661	1.666	57.05	0.03	0.89	0.104	3.773	1.705	60.27	0.01	0.91	0.083
Wan2.2-5B-FT [§]	1.342	0.951	38.53	0.23	0.74	0.300	1.980	1.079	42.76	0.12	0.80	0.208
Wan2.2-5B-FT-DA [¶]	1.055	0.875	33.99	0.33	0.75	0.385	1.841	1.023	39.01	0.11	0.83	0.214

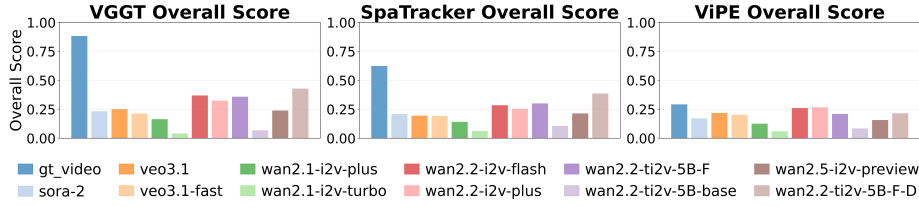


Fig. 8: Overall score comparison between different spatio-temporal reconstruction tools. Detailed evaluation results can be found in the **appendix**.

evaluates all models on unseen data using VGGT for path reconstruction. The fine-tuned Wan2.2-5B (Wan2.2-5B-FT) improves its score from 0.068 to 0.357. The augmented version (Wan2.2-5B-FT-DA) outperforms the base model with a even better score of 0.428.

4.4 Ablation Study

Comparison Among Reconstruction Tools. Fig. 8, Table 3 and Table 5 compare the performance of three spatio-temporal reconstruction tools: VGGT, SpaTracker, and ViPE. VGGT achieves the best weighted overall score of 0.882 with the ground truth videos. SpaTracker shows lower performance, while ViPE produces the weakest results. The high score achieved by VGGT on ground truth videos confirms that decoded trajectories align well with ground truth trajectories, validating our evaluation approach. Notably, this component is replaceable in our framework as temporal reconstruction methods continue to advance.

5 Discussion

World Models Performance. The best off-the-shelf video WM, **Wan2.2-Flash**, achieves only 0.368 overall score. This indicates a significant gap between current world model capabilities and reliable path planning. However, qualitative inspection of generated videos reveals that most models correctly understand the semantic target and show plausible motion tendencies. Additionally, our benchmark includes scenarios with implicit semantic targets, which makes our benchmark challenging for video world models to harvest good scores.

Holistic Evaluation: Beyond Visual Quality. Target-Bench intrinsically evaluates three capabilities simultaneously: (1) spatio-temporal consistency, (2) semantic reasoning, and (3) geometric path planning accuracy. Blurred frames, temporal discontinuities, or spatial warping will directly result in poor 3D reconstruction which causes low path accuracy. The holistic assessment of our work fundamentally differs from existing benchmarks that isolate individual dimensions (visual quality, physics consistency) and miss their integration.

Real-World Navigation. Given an observation and semantic goal, WMs predict future frames depicting motion toward the target, and our world-decoder extracts a path for the robot to execute. This has potential for future use in closed-loop robot navigation in unstructured, mapless environments. Recent work [5,43] shows WMs can explore new spaces while retaining information in latent memory. Combined with our decoder, robots could follow semantic goals using only visual predictions. As shown in the website¹, we also show some preliminary results in terms of deployment of video world model with world-decoder on a quadruped robot regarding real-world navigation tasks.

6 Conclusion and Future Work

We introduced Target-Bench for evaluating video world models on mapless path planning toward semantic targets. Our evaluation pipeline recovers planned paths from generated videos and measures planning performance against SLAM-based ground-truth robot paths. The best off-the-shelf video WM achieves a relatively low score. However, fine-tuning an open-source 5B-parameter model on only 325 scenarios from our dataset surpasses the best off-the-shelf video WMs. This suggests that WMs can effectively learn navigation tasks from limited real-world data, showing promising potential for robot path planning. Future work could extend the current framework towards closed-loop re-planning on the robot, and studying how latent memory supports complex receding-horizon navigation tasks. Better reconstruction tools are expected to further improve path planning accuracy, facilitating real-world deployment of WMs in robotics.

¹ <https://target-bench.github.io/>

References

1. Akhtyamov, T., Mdfaa, M.A., Ramirez, J.A., Bakulin, S., Devchich, G., Fatykhov, D., Mazurov, A., Zipa, K., Mohrat, M., Kolesnik, P., et al.: Egowalk: A multimodal dataset for robot navigation in the wild. arXiv preprint arXiv:2505.21282 (2025)
2. Alonso, E., Jelley, A., Micheli, V., Kanervisto, A., Storkey, A., Pearce, T., Fleuret, F.: Diffusion for world modeling: Visual details matter in atari. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) *Advances in Neural Information Processing Systems*. vol. 37, pp. 58757–58791. Curran Associates, Inc. (2024). <https://doi.org/10.52202/079017-1873>
3. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., Van Den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 3674–3683 (2018)
4. Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., et al.: V-jepa 2: Self-supervised video models enable understanding, prediction and planning (2025)
5. Ball, P.J., Bauer, J., Belletti, F., Brownfield, B., Ephrat, A., Fruchter, S., Gupta, A., Holsheimer, K., Holynski, A., Hron, J., Kaplanis, C., Limont, M., McGill, M., Oliveira, Y., Parker-Holder, J., Perbet, F., Scully, G., Shar, J., Spencer, S., Tov, O., Villegas, R., Wang, E., Yung, J., et al.: Genie 3: A new frontier for world models (2025)
6. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., Ramesh, A.: Video generation models as world simulators (2024)
7. Bruce, J., Dennis, M., Edwards, A., Parker-Holder, J., Shi, Y., Hughes, E., Lai, M., Mavalankar, A., Steigerwald, R., Apps, C., Aytar, Y., Bechtle, S., Behbahani, F., Chan, S., Heess, N., Gonzalez, L., Osindero, S., Ozair, S., Reed, S., Zhang, J., Zolna, K., Clune, J., de Freitas, N., Singh, S., Rocktäschel, T.: Genie: Generative interactive environments. In: Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F. (eds.) *Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research*, vol. 235, pp. 4603–4623. PMLR (21–27 Jul 2024)
8. DeepMind, G.: Veo 3: Advanced controllable video generation with physics-aware dynamics (2025)
9. Deng, Y., Pan, Z., Zhang, H., Li, X., Hu, R., Ding, Y., Zou, Y., Zeng, Y., Zhou, D.: Rethinking video generation model for the embodied world. arXiv preprint arXiv:2601.15282 (2026)
10. Duan, H., Yu, H.X., Chen, S., Fei-Fei, L., Wu, J.: Worldscore: A unified evaluation benchmark for world generation. arXiv preprint arXiv:2504.00983 (2025)
11. Gao, Y., Piccinini, M., Zhang, Y., Wang, D., Moller, K., Brusnicki, R., Zarrouki, B., Gambi, A., Totz, J.F., Storms, K., et al.: Foundation models in autonomous driving: A survey on scenario generation and scenario analysis. arXiv preprint arXiv:2506.11526 (2025)
12. Ha, D., Schmidhuber, J.: Recurrent world models facilitate policy evolution. In: *Advances in Neural Information Processing Systems* 31, pp. 2451–2463. Curran Associates, Inc. (2018)
13. Hafner, D., Lillicrap, T., Ba, J., Norouzi, M.: Dream to control: Learning behaviors by latent imagination. In: *International Conference on Learning Representations* (2020)

14. Hirose, N., Glossop, C., Sridhar, A., Shah, D., Mees, O., Levine, S.: Lelan: Learning a language-conditioned navigation policy from in-the-wild video. In: Agrawal, P., Kroemer, O., Burgard, W. (eds.) *Proceedings of The 8th Conference on Robot Learning. Proceedings of Machine Learning Research*, vol. 270, pp. 666–688. PMLR (06–09 Nov 2025)
15. Hirose, N., Shah, D., Sridhar, A., Levine, S.: Sacson: Scalable autonomous control for social navigation. *IEEE Robotics and Automation Letters* **9**(1), 49–56 (2023)
16. Huang, J., Zhou, Q., Rabeti, H., Korovko, A., Ling, H., Ren, X., Shen, T., Gao, J., Slepichev, D., Lin, C.H., Ren, J., Xie, K., Biswas, J., Leal-Taixe, L., Fidler, S.: Vipe: Video pose engine for 3d geometric perception. In: *NVIDIA Research Whitepapers* arXiv:2508.10934 (2025)
17. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 21807–21818 (2024)
18. Karnan, H., Nair, A., Xiao, X., Warnell, G., Pirk, S., Toshev, A., Hart, J., Biswas, J., Stone, P.: Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation. *IEEE Robotics and Automation Letters* **7**(4), 11807–11814 (2022)
19. Ku, A., Anderson, P., Patel, R., Ie, E., Baldridge, J.: Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 4392–4412 (2020)
20. Kümmerle, R., Grisetti, G., Strasdat, H., Konolige, K., Burgard, W.: G2o: A general framework for graph optimization. In: *2011 IEEE International Conference on Robotics and Automation*. pp. 3607–3613 (2011). <https://doi.org/10.1109/ICRA.2011.5979949>
21. LeCun, Y.: A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review* **62**(1), 1–62 (2022)
22. Li, C., Krause, A., Hutter, M.: Robotic world model: A neural network simulator for robust policy optimization in robotics. *arXiv preprint arXiv:2501.10100* (2025)
23. Li, D., Fang, Y., Chen, Y., Yang, S., Cao, S., Wong, J., Luo, M., Wang, X., Yin, H., Gonzalez, J.E., et al.: Worldmodelbench: Judging video generation models as world models. *CoRR* **abs/2502.20694** (February 2025)
24. Li, Y., Zhu, Y., Wen, J., Shen, C., Xu, Y.: Worldeval: World model as real-world robot policies evaluator. *arXiv preprint arXiv:2505.19017* (2025)
25. Li, Z., Li, C., Mao, X., Lin, S., Li, M., Zhao, S., Xu, Z., Li, X., Feng, Y., Sun, J., et al.: Sekai: A video dataset towards world exploration. *arXiv preprint arXiv:2506.15675* (2025)
26. Liu, X., Li, J., Jiang, Y., Sujay, N., Yang, Z., Zhang, J., Abanes, J., Zhang, J., Feng, C.: Citywalker: Learning embodied urban navigation from web-scale videos. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 6875–6885 (2025)
27. Long, X., Zhao, Q., Zhang, K., Zhang, Z., Wang, D., Liu, Y., Shu, Z., Lu, Y., Wang, S., Wei, X., et al.: A survey: Learning embodied intelligence from physical simulators and world models. *arXiv preprint arXiv:2507.00917* (2025)
28. modelscope: Diffsynth-studio: examples/wanvideo (2025), accessed: 2025-11-14
29. Nguyen, D.M., Nazeri, M., Payandeh, A., Datar, A., Xiao, X.: Toward human-like social robot navigation: A large-scale, multi-modal, social human navigation dataset. In: *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. pp. 7442–7447. IEEE (2023)

30. NVIDIA, Agarwal, N., Ali, A., Bala, M., Balaji, Y., et al.: Cosmos: World foundation model platform for physical ai. CoRR **abs/2501.03575** (January 2025)
31. Pang, J.C., Tang, N., Li, K., Tang, Y., Cai, X.Q., Zhang, Z.Y., Niu, G., Sugiyama, M., Yu, Y.: Learning view-invariant world models for visual robotic manipulation. In: The Thirteenth International Conference on Learning Representations (2025)
32. Parker-Holder, J., Ball, P., Bruce, J., Dasagi, V., Holsheimer, K., Kaplanis, C., Moufarek, A., Scully, G., Shar, J., Shi, J., Spencer, S., Yung, J., Dennis, M., Kenjeyev, S., Long, S., Mnih, V., Chan, H., Gazeau, M., Li, B., Pardo, F., Wang, L., Zhang, L., Besse, F., Harley, T., Mitenkova, A., Wang, J., Clune, J., Hassabis, D., Hadsell, R., Bolton, A., Singh, S., Rocktäschel, T.: Genie 2: A large-scale foundation world model (2024)
33. Segal, A., Haehnel, D., Thrun, S.: Generalized-ICP. In: Robotics: Science and Systems (2009). <https://doi.org/10.15607/RSS.2009.V.021>
34. Shang, Y., Li, Z., Ma, Y., Su, W., Jin, X., Wang, Z., Jin, L., Zhang, X., Tang, Y., Su, H., et al.: Worldarena: A unified benchmark for evaluating perception and functional utility of embodied world models. arXiv preprint arXiv:2602.08971 (2026)
35. Team, T.O.S.: Sora 2 is here: Our latest video generation model is more physically accurate, realistic, and more controllable than prior systems. it also features synchronized dialogue and sound effects. create with it in the new sora app. (2025), accessed: 2025-11-09
36. Unitree: Unifolm-wma-0: A world-model-action (wma) framework under unifolm family (2025)
37. Waghmare, S.M., Wilber, K., Hawkey, D., Yang, X., Wilson, M., Debats, S., Nungsigkapijan, C., Sharma, A., Pandikow, L., Wang, H., et al.: Sanpo: A scene understanding, accessibility and human navigation dataset. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 7866–7875. IEEE (2025)
38. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. CoRR **abs/2503.20314** (March 2025)
39. Wang, D., Sun, Z., Li, Z., Wang, C., Peng, Y., Ye, H., Zarrouki, B., Li, W., Piccinini, M., Xie, L., et al.: Enhancing physical consistency in lightweight world models. arXiv preprint arXiv:2509.12437 (2025)
40. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5294–5306 (June 2025)
41. Wiedemer, T., Li, Y., Vicol, P., Gu, S.S., Matarese, N., Swersky, K., Kim, B., Jaini, P., Geirhos, R.: Video models are zero-shot learners and reasoners. arXiv preprint arXiv:2509.20328 (2025)
42. Xiao, Y., Wang, J., Xue, N., Karaev, N., Makarov, I., Kang, B., Zhu, X., Bao, H., Shen, Y., Zhou, X.: Spatialtrackerv2: 3d point tracking made easy. In: ICCV (2025)
43. Yu, J., Bai, J., Qin, Y., Liu, Q., Wang, X., Wan, P., Zhang, D., Liu, X.: Context as memory: Scene-consistent interactive long video generation with memory retrieval. arXiv preprint arXiv:2506.03141 (2025)
44. Yue, H., Huang, S., Liao, Y., Chen, S., Zhou, P., Chen, L., Yao, M., Ren, G.: Ewmbench: Evaluating scene, motion, and semantic quality in embodied world models. arXiv preprint arXiv:2505.09694 (2025)

- 45. Zhang, J., Jiang, M., Dai, N., Lu, T., Uzunoglu, A., Zhang, S., Wei, Y., Wang, J., Patel, V.M., Liang, P.P., et al.: World-in-world: World models in a closed-loop world. arXiv preprint arXiv:2510.18135 (2025)
- 46. Zhu, C., Yu, R., Feng, S., Burchfiel, B., Shah, P., Gupta, A.: Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. arXiv preprint arXiv:2504.02792 (2025)