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

Dingrui Wang[1,2,*] Hongyuan Ye[1,*] Zhihao Liang[1,*] Zhexiao Sun[1,*] Zhaowei Lu[1]
Yuchen Zhang[1] Yuyu Zhao[1] Yuan Gao[1] Marvin Seegert[1] Finn Schäfer[1]
Haotong Qin[3] Wei Li[4] Luigi Palmieri[2] Felix Jahncke[1]
Mattia Piccinini[1] Johannes Betz[1]

[1]TUM [2]Bosch AI Center [3]ETH [4]NJU
https://target-bench.github.io/

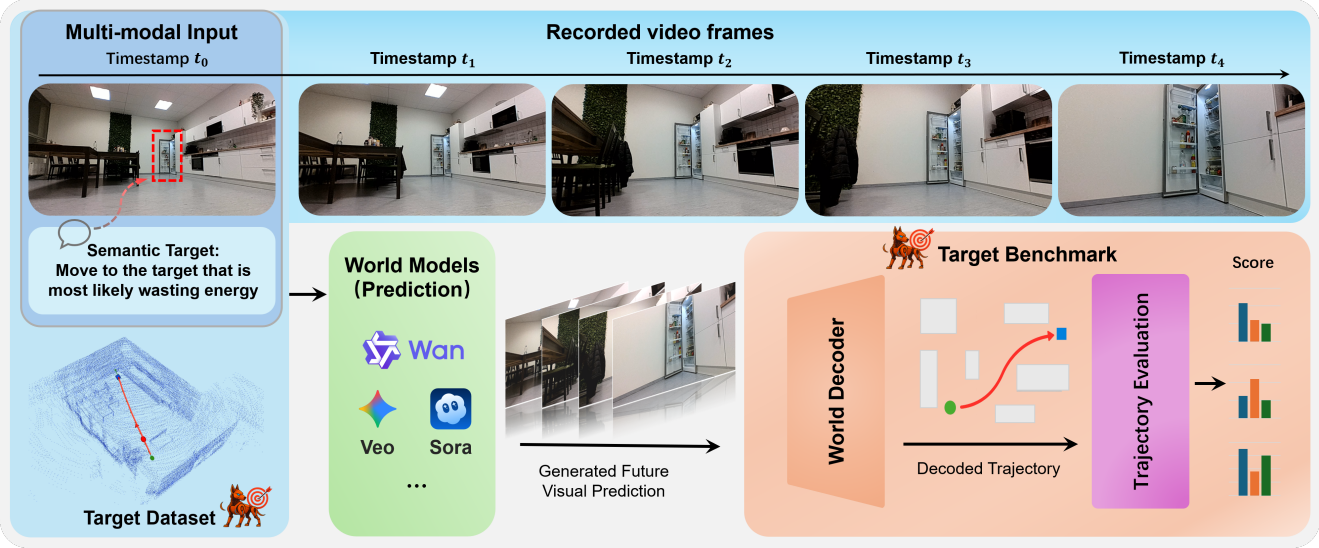


Figure 1. **Target-Bench** provides a **dataset** collected with a quadruped robot, and a **benchmark** for evaluating world models in mapless path planning toward text-specified goals with implicit semantic meaning. In Target-Bench, world models receive a camera frame and a textual prompt describing the target state, and predict a future video depicting the trajectory toward the goal. A **world decoder** then extracts the planned path from this video, which is compared against the maneuver executed by a human-operated quadruped.

Abstract

*While recent world models generate highly realistic videos, their ability to perform robot path planning remains unclear and unquantified. We introduce **Target-Bench**, the first benchmark specifically designed to evaluate world models on mapless path planning toward semantic targets in real-world environments. Target-Bench provides 450 robot-collected video sequences spanning 45 semantic categories with SLAM-based ground truth trajectories. Our evaluation pipeline recovers camera motion from generated videos and measures planning performance using five complementary metrics that quantify target-reaching capability, trajectory accuracy, and directional consistency. We evaluate state-of-the-art models including Sora 2, Veo 3.1, and the Wan series. The best off-the-shelf model (Wan2.2-Flash) achieves only 0.299 overall score, revealing significant limitations in current world models for robotic planning tasks. We show that fine-tuning an open-source 5B-parameter model on only 325 scenarios from our dataset achieves 0.345 overall score—an improvement of more than 400% over its base version and 15% higher than the best off-the-shelf model.*

∗ Equal contribution; author order settled via Mario Kart.

1. Introduction

> "If one day, we bring a humanoid robot to this conference venue, a place it has never seen before, and I can simply say 'Please bring this bottle of water to someone in the audience,' and it can smoothly walk over and do it by itself, I think that would be the robot's ChatGPT moment."
>
> — Xingxing Wang, Head of Unitree, at the World Robot Conference 2025.

Embodied AI has been advancing rapidly, while its core challenges are becoming clearer. Despite major improvements (better actuators and sensors) in robotic hardware, AI software remains the main bottleneck limiting robots from realizing their full potential [22]. As a result, the robotics and embodied AI community is seeking to develop AI systems that are more robust and more generalizable. Recent breakthroughs in World Models (WMs) [4, 7, 29, 32, 35] have drawn attention from the community. WMs learn to predict how the world evolves over time [10, 18]. Given an initial observation (e.g., an image frame) and a condition (e.g., a text prompt or an action), models such as Veo3.1 [7] and Genie3 [4] can generate future frames with high spatio-temporal consistency. More importantly, they exhibit remarkable reasoning abilities in visual semantics and causal relationships [5, 9, 35]. This potential has sparked interest within the embodied AI community in applying WMs to robotic applications [19, 26, 33, 39]. Recent approaches from Unitree, such as UnifoLM-WMA-0 [30], aim to bridge 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 WM's reasoning, task-solving, and planning ability? Existing evaluation frameworks focus mainly on reconstruction quality, from visual fidelity to physical consistency [8, 20], while assessing mapless path planning driven by semantic understanding remains an open challenge [22]. Bridging this gap requires a benchmark that evaluates not just how realistic a WM's predictions look, but whether they contain the semantic reasoning needed for robotic planning tasks.

To address this challenge, we introduce **Target-Bench**. To our knowledge, it is the first benchmark evaluating WMs for mapless path planning toward semantic targets in unstructured real-world environments. The target state is specified by user text and can have implicit semantic meaning (Fig. 1). Our contributions are:

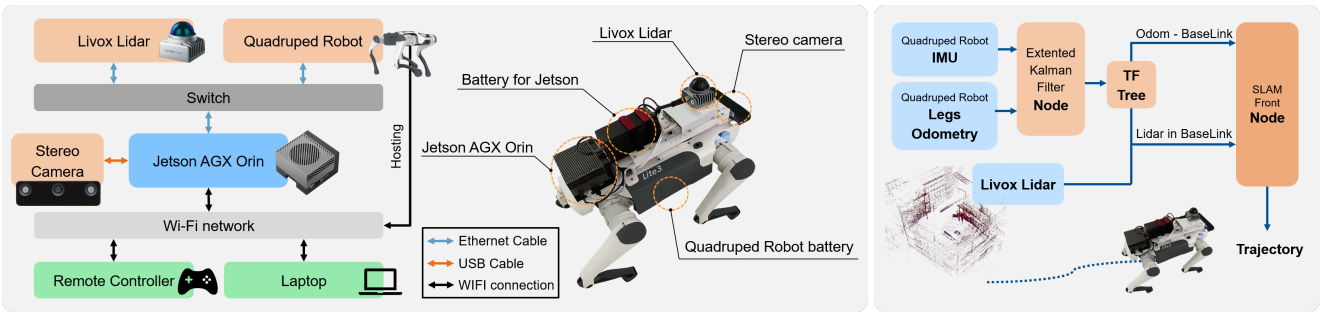
- An **open-source dataset** of 450 videos (112,500 frames) collected with a quadruped robot, covering 45 semantic target categories in diverse indoor and outdoor environments, with ground truth trajectories and human annotations for explicitly- and implicitly-stated targets.
- The first systematic **evaluation pipeline** for WMs in mapless path planning with textual semantic goals.
- A **world decoder** that extracts camera motion from generated videos, comparing spatial reconstruction methods and introducing a new scale recovery technique.
- A comparative study of open-source and proprietary WMs, including the first **fine-tuning** of an open-source model on a small real-world dataset for path planning, demonstrating improved generalization to unseen environments over proprietary models.

2. Related Work

World Models. Early approaches such as World Models [10] and Dreamer [11] introduced latent state-space models, demonstrating that generative prediction has the potential to support planning and control [9]. More recently, foundation-scale methods pretrained on large video corpora have emerged: Cosmos [25] compared diffusion and autoregressive paradigms; DIAMOND [2] improved visual fidelity; V-JEPA 2 [3] enhanced efficiency by forecasting in latent space. Advances in video generation, including Sora 2 [29], Veo 3.1 [7, 35], and Wan [32], have further improved prompt adherence and physics-aware dynamics, enabling long-horizon predictions. Interactive models such as Genie-1 [6], Genie-2 [27], and Genie-3 [4] provide controllable environments for use in robotics and games. Unitree's recent UnifoLM-WMA-0 [30] 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.

Benchmarks for World Models. Recent works have proposed benchmarks to evaluate generative WMs, often relying on Vision Language Model (VLM)-based judges to assess the quality and consistency of generated scenes. VBench [15] targets 16 fine-grained video quality dimensions, while WorldModelBench [20] focuses on physics consistency (e.g., Newtonian motion). Compared to VBench, WorldScore [8] emphasizes controllability and dynamics in next-scene prediction. World-in-World [38] enables closed-loop evaluation but uses WMs as simulators and does not test their planning abilities. Overall, existing benchmarks primarily assess spatio-temporal or physical consistency, and do not consider high-level robot planning. Recent advances in spatio-temporal scene reconstruction [14, 34, 36] further highlight the limits of using only VLM-based scoring for WM evaluation. We extend this literature by assessing not only spatio-temporal coherence but also path planning abilities.

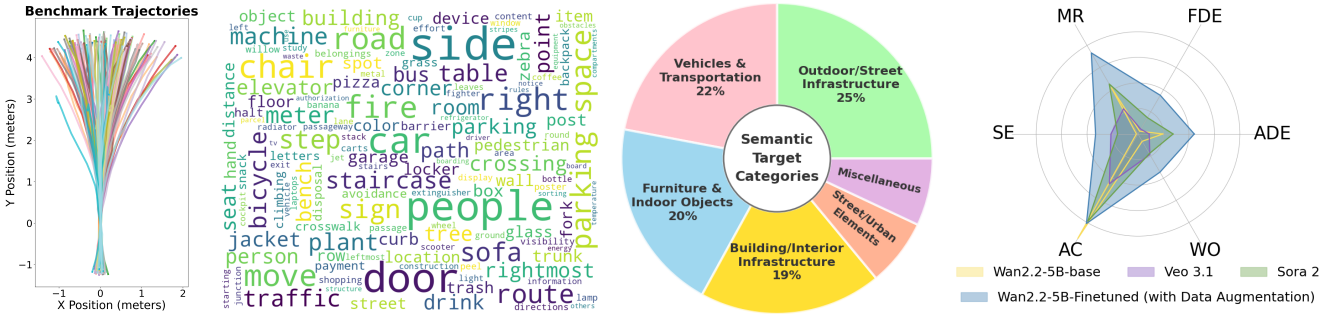
Semantic Navigation Datasets. Several datasets address semantic robot navigation, with social compliance in real-



(a) Quadruped robot hardware setup and communication network.

(b) SLAM pipeline for state estimation.

Figure 2. Robot setup and SLAM pipeline.



(a) All trajectories.

(b) Word cloud of captions.

(c) Semantic target categories.

(d) Target-Bench Evaluation Results.

Figure 3. Visualization of dataset structure and semantics.

world scenarios. EgoWalk [1] provides egocentric data with stereo images, odometry, and traversability labels, but suffers from heading misalignment and IMU drift. LeLaN [13] leverages unlabeled human walking videos to train language-conditioned policies, while SCAND [16] offers socially compliant robot demonstrations. MuSoHu [24] captures multimodal wearable data for natural interactions, SACSoN [12] focuses on office navigation with human–robot interaction, and SANPO [31] targets outdoor navigation. CityWalker [21] enables mapless urban navigation amid traffic and crowds. However, these datasets lack explicit navigation targets, and their VLM-based annotations are not target-oriented.

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 quadruped robot equipped with multi-modal sensors in diverse indoor and outdoor environments. It spans 45 semantic target categories (e.g., doors, chairs, trees) 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

3.1.1. Quadruped Robot Platform

As shown in Fig. 2a, our data-collection platform is built on a DEEP Robotics Lite 3 Venture quadruped. 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 LiDAR-centric SLAM pipeline [17, 28] 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.

3.1.2. Semantic Target Data Collection

Each data sample in the Target dataset consists of four components: a video sequence, a ground truth trajectory, a se-

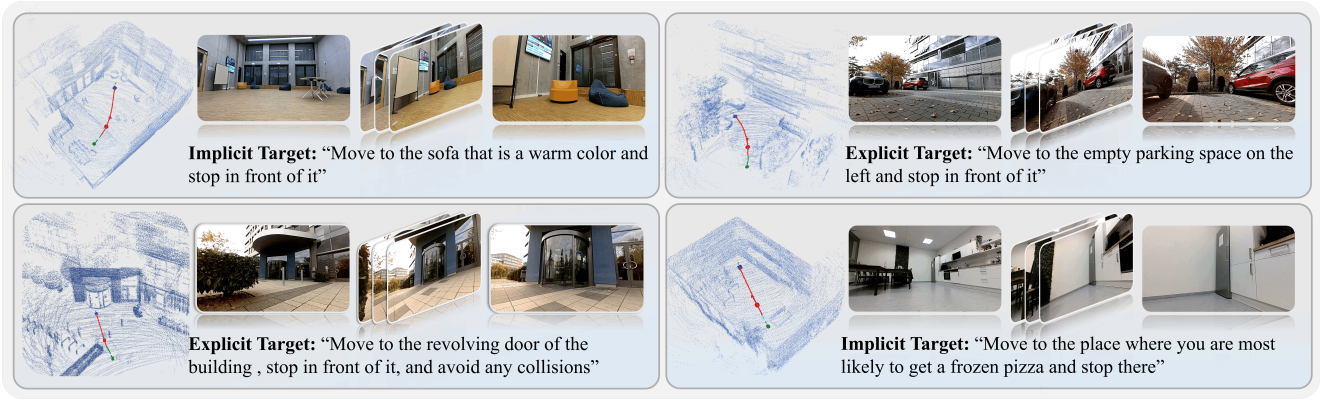


Figure 4. Data sample visualization.

mantic target and a point cloud map. Video frames are captured at 25Hz, yielding approximately 10 seconds of continuous observation per sample. The semantic targets are annotated and pre-selected by human experts to ensure diversity and relevance for navigation tasks. Fig. 3a shows that our trajectories span diverse directions and movement patterns, providing balanced and realistic navigation coverage. Fig. 3c presents the semantic target distribution with a variety of scenarios. This mix captures both indoor and outdoor environments with common static and dynamic objects. Fig. 3b highlights the dataset's rich caption vocabulary. The targets are 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). Overall, the whole dataset is divided into a train split (325 scenarios) and a benchmark (evaluation) split (125 scenarios).

3.2. Target Benchmark Architecture

The Target benchmark architecture (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 compares predicted trajectories against SLAM-verified ground truth using a comprehensive suite of metrics.

3.2.1. 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 [34], SpaTracker [36], and ViPE [14]. Each method processes video frames to recover camera poses, but differs in their approach and output characteristics.

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 $\mathcal{I} = \{I_1, I_2, \ldots, I_S\}$, 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 matrix, $\mathbf{q}_s \in \mathbb{R}^4$ 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 $\mathbf{E}_s = [\mathbf{R}(\mathbf{q}_s) \mid \mathbf{T}_s] \in \mathbb{R}^{3\times4}$, where $\mathbf{R}(\mathbf{q}_s)$ converts the quaternion to a rotation matrix.

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 $\mathbf{Q} \in \mathbb{R}^{N\times3}$ on the first frame, the tracker predicts their 3D trajectories and optimizes camera poses through bundle adjustment:

$$\{\mathbf{E}_s^*\} = \arg\min_{\{\mathbf{E}_s\}} \sum_{s=1}^{S}\sum_{n=1}^{N} \rho\left(\|\pi(\mathbf{K}_s\mathbf{E}_s\mathbf{X}_n) - \mathbf{x}_{n,s}\|^2\right) \tag{1}$$

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 $\mathbf{C}_{2W,s} \in \mathbb{R}^{4\times4}$ is the full camera-to-world transform. As in VGGT, SpaTracker outputs are at arbitrary scale.

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, removing the need for scale recovery. The final path is a sequence of SE(3) transformations $\{\mathbf{T}_s\}_{s=1}^S$, where each $\mathbf{T}_s \in \mathbb{R}^{4\times4}$ encodes rotation and translation in meters.

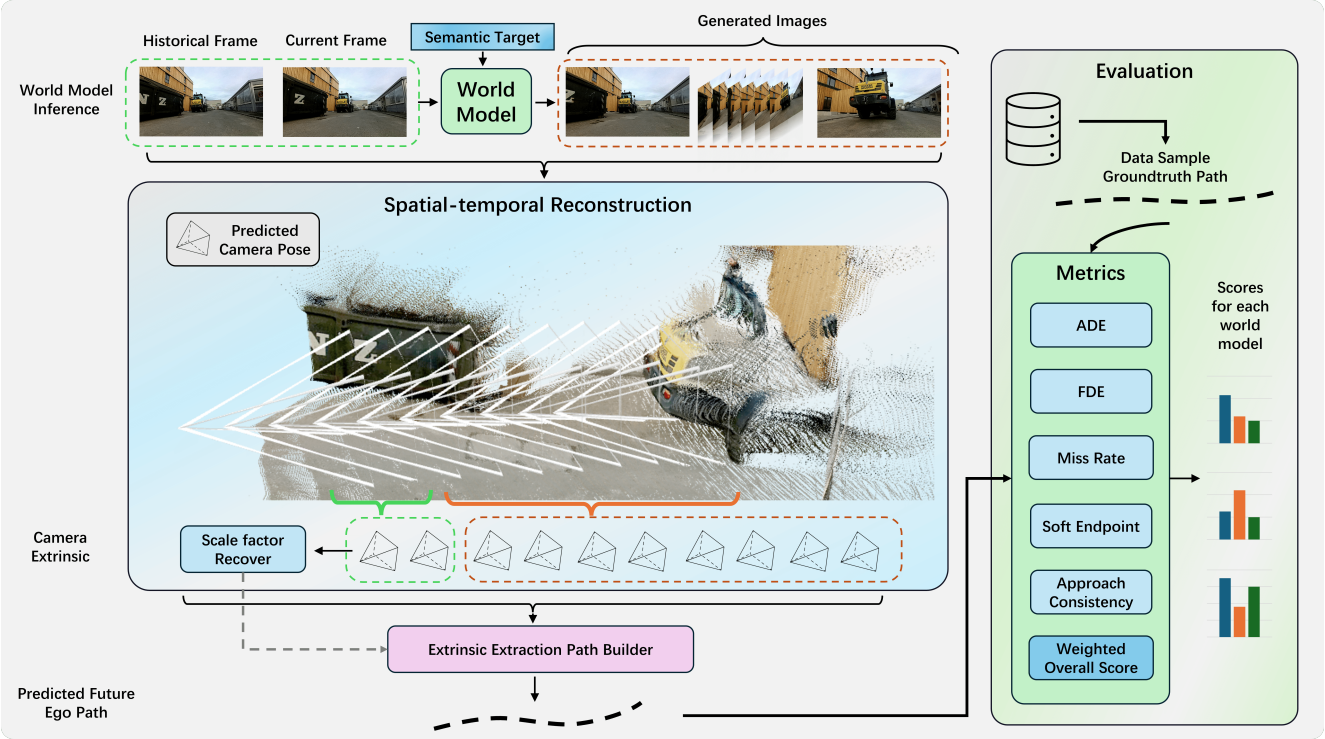


Figure 5. Target Benchmark Architecture.

Scale Factor Recovery. Monocular methods such as VGGT and SpaTracker estimate camera motion only up to an unknown global scale. We restore metric consistency at the segment level by anchoring predictions to a single scalar scale factor λ derived from ground truth displacement. Let $\mathbf{E}_1, \mathbf{E}_k \in \mathbb{R}^{3\times4}$ be the predicted extrinsic matrices for the first and the k-th frame. We extract the translation vectors $\mathbf{t}_1$ and $\mathbf{t}_k$ as the fourth column of $\mathbf{E}_1$ and $\mathbf{E}_k$. The predicted displacement is $d_{\text{pred}} = \|\mathbf{t}_k - \mathbf{t}_1\|_2$, and the scale factor is $\lambda = d_{\text{real}}/d_{\text{pred}}$, while the real displacement is d_{real}. We then rescale all predicted translations uniformly, $\mathbf{t}_s^{\text{scaled}} = \lambda \mathbf{t}_s$ for $s = 1, \ldots, S$. This lightweight, data-anchored normalization preserves relative geometry while lifting trajectories to meters, enabling direct and fair comparison against ground truth paths. Note that ViPE inherently produces metric-scale outputs via sensor fusion and therefore does not require this step.

3.2.2. Path Evaluation

We evaluate predicted trajectories on the test split of the Target dataset 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 $\mathbf{s}^{\text{GT}} = \{s_1, s_2, \ldots, s_T\}$ and prediction $\hat{\mathbf{s}} = \{\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_T\}$, where $s_t, \hat{s}_t \in \mathbb{R}^2$ and T is the total 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: 2.0 m): $\text{MR} = \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}\left[\|\hat{s}_t - s_t\|_2 > \tau\right] \cdot 100$

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

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

where $\sigma = 0.6$ m controls 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 each a variable radius:

$$\sigma_i = \sigma_{\min} + (\sigma_{\max} - \sigma_{\min}) \exp\left(-\frac{(p_i - 0.5)^2}{2\beta^2}\right) \tag{3}$$

where $p_i = i/(M-1)$ is the normalized progress, $\sigma_{\min} = 0.15$ m, $\sigma_{\max} = 0.5$ m, and $\beta = 0.25$. A predicted point $\hat{s}_j$

is covered if $\min_i \|\hat{s}_j - s_i^{\mathrm{GT}}\|_2 \le \sigma_i$. The AC score is:

$$\mathrm{AC} = \begin{cases} 1, & N_{\mathrm{c}} = N_{\mathrm{p}} \\ \exp\left(-\gamma \cdot \dfrac{N_{\mathrm{p}} - N_{\mathrm{c}}}{N_{\mathrm{p}}}\right), & \text{otherwise} \end{cases} \tag{4}$$

where N_{p} is the total number of predicted points, N_{c} is the number covered by the corridor, and $\gamma = 5$.

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

$$S_{\mathrm{overall}} = w_{\mathrm{ADE}} \cdot \exp\left(-\frac{\mathrm{ADE}}{\tau_{\mathrm{ADE}}}\right) + w_{\mathrm{FDE}} \cdot \exp\left(-\frac{\mathrm{FDE}}{\tau_{\mathrm{FDE}}}\right) + w_{\mathrm{MR}} \cdot \left(1 - \frac{\mathrm{MR}}{100}\right) + w_{\mathrm{SE,AC}} \cdot \mathrm{SE} \cdot \mathrm{AC} \tag{5}$$

with default weights: $w_{\mathrm{ADE}} = 0.05$, $w_{\mathrm{FDE}} = 0.10$, $w_{\mathrm{MR}} = 0.10$, $w_{\mathrm{SE,AC}} = 0.65$, and scale parameters $\tau_{\mathrm{ADE}} = \tau_{\mathrm{FDE}} = 1.0$ m.

4. Experiments

Our proposed evaluation framework uses five key metrics detailed in Sec. 3.2.2. The analysis includes several state-of-the-art world models, including Sora 2 [29], Veo 3.1 [7], 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) [32]. 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 [23], with 8× NVIDIA A800 80 GB GPUs. For inference of the fine-tuned 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 DELL Alienware Aurora R15 workstation equipped with an NVIDIA RTX 4090 GPU.

4.2. Evaluation for off-the-shelf Models

Table 1 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 0.299. Specifically, it obtains the lowest errors in FDE (1.362m), ADE (1.005m), and MR (38.75%), while achieving the highest SE (0.292). Figure 6 visualizes the performance comparison across all models. The ground truth video achieves the highest scores across all metrics, and is still affected by reconstruction errors like the other models.

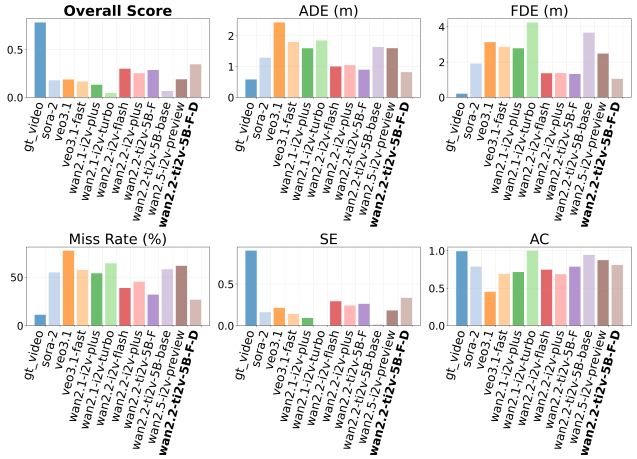


Figure 6. World model performance comparison with VGGT as world decoder's spatio-temporal reconstruction tool.

Displacement Errors: ADE values range from 1.0m to 2.4m across different models. Veo 3.1 exhibits the highest displacement error (2.432m), while Wan2.2-Flash and Wan2.2-Plus achieve the best accuracy (around 1.0m). For FDE, errors show wider variation, with Wan2.1-Turbo displaying the highest FDE (4.243m).

Reliability Metrics: The Miss Rate varies significantly from 38.75% to 77.25%. Veo 3.1 shows the highest miss rate (77.25%), indicating poor trajectory quality, while Wan2.2-Flash achieves the best (38.75%). Most models score below 0.3 on the Soft Endpoint metric, suggesting challenges in reaching target endpoints accurately.

Consistency: Wan2.1-Turbo achieves the highest Approach Consistency (1.000), indicating perfect directional alignment, while Veo 3.1 shows the poorest performance (0.450). Sora 2 demonstrates good consistency at 0.788.

Explicit vs. Implicit Targets: As shown in Table 2, performance on implicitly defined targets closely matches that on explicit targets, with small, model-specific fluctuations. This indicates that current WMs can understand semantic goals even when they are not explicitly defined.

4.3. Fine-tuned Models

For WM fine-tuning, we use 325 scenarios from our dataset and sample 121 frames from each video. To expand the dataset size, we apply a shifting-frame augmentation: for each video, we generate four clips by evenly sampling 121 frames starting from different offsets, preserving similar temporal coverage. We choose to fine-tune the open-sourced Wan2.2-TI2V-5B model under two settings: without and with data augmentation. Table 1 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 0.066 to 0.287. The augmented ver-

Spatial Method	World Model	Open Source	Video Length (s) & Resolution	Metrics FDE ↓	ADE ↓	MR ↓	SE ↑	AC ↑	WO ↑
VGGT [34]	gt_video	-	8 - 720p	0.203	0.580	11.08	0.901	0.993	0.783
	Sora 2	✗	8 - 720p	1.912	1.289	54.84	0.160	0.788	0.178
	Veo 3.1	✗	8 - 720p	3.125	2.432	77.25	0.212	0.450	0.187
	Veo 3.1-fast	✗	8 - 720p	2.863	1.798	57.42	0.140	0.691	0.165
	Wan2.5-Preview†	✗	10 - 720p	2.478	1.596	61.69	0.182	0.873	0.188
	Wan2.2-Plus†	✗	5 - 1080p	1.377	1.044	45.25	0.240	0.686	0.252
	Wan2.2-Flash†	✗	5 - 720p	1.362	1.005	38.75	0.292	0.746	0.299
	Wan2.1-Plus†	✗	5 - 720p	2.782	1.594	53.90	0.090	0.715	0.131
	Wan2.1-Turbo†	✗	5 - 720p	4.243	1.850	64.21	0.000	**1.000**	0.046
	Wan2.2-5B-base	✓	8 - 720p	3.666	1.636	58.11	0.012	0.944	0.066
	Wan2.2-5B-FT§	✓	8 - 720p	1.320	0.897	31.91	0.261	0.787	0.287
	Wan2.2-5B-FT-DA¶	✓	8 - 720p	**1.050**	**0.816**	**26.71**	**0.333**	0.810	**0.345**

Table 1. Evaluation results of Final Displacement Error (FDE), Average Displacement Error (ADE) and Miss Rate (MR), Soft Endpoint (SE), Approach Consistency (AC) and Weighted Overall (WO). † *Image-to-Video (I2V) models,* § *fine-tuned,* ¶ *fine-tuned with data augmentation.*

Data	World Model	Metrics FDE ↓	ADE ↓	MR ↓	SE ↑	AC ↑	WO ↑
Explicit Target	gt_video	0.210	0.586	10.71	0.892	0.994	0.778
	Sora 2	1.979	1.381	56.93	0.165	0.740	0.177
	Veo 3.1	3.275	2.598	80.69	0.172	0.419	0.155
	Veo 3.1-f	3.225	1.928	57.37	0.136	0.685	0.162
	Wan2.5*	2.449	1.656	61.70	0.182	**0.826**	0.189
	Wan2.2-P†	**1.413**	1.028	42.25	0.242	0.724	0.258
	Wan2.2-Fl†	1.418	**1.006**	**35.69**	**0.295**	0.776	**0.305**
Implicit Target	gt_video	0.193	0.573	11.61	0.913	0.992	0.791
	Sora 2	1.819	1.160	51.91	0.154	0.856	0.179
	Veo 3.1	2.914	2.200	72.43	0.269	0.493	0.231
	Veo 3.1-f	2.355	1.616	57.49	0.144	0.700	0.170
	Wan2.5*	2.519	1.512	61.67	0.181	**0.941**	0.188
	Wan2.2-P†	1.327	1.067	49.46	0.238	0.634	0.242
	Wan2.2-Fl†	**1.283**	**1.002**	**43.05**	**0.288**	0.706	**0.290**

Table 2. Evaluation results with explicit and implicit semantic targets. * *Wan2.5-I2V-Preview,* † *Image-to-Video (I2V) models.*

sion (`Wan2.2-5B-FT-DA`) outperforms the base model by more than 400% and achieves the best overall score.

4.4. Ablation Study

Comparison Among Reconstruction Tools. Fig. 7, Table 1 and Table 3 compare the performance of three spatio-temporal reconstruction tools: VGGT, SpaTracker, and ViPE. VGGT achieves the best results, with the ground truth video obtaining a weighted overall score of 0.783 and the best-performing off-the-shelf world model (Wan2.2-Flash) achieving 0.299. SpaTracker shows slightly lower performance, while ViPE produces the weakest results. The high score achieved by VGGT on ground truth videos (0.783)

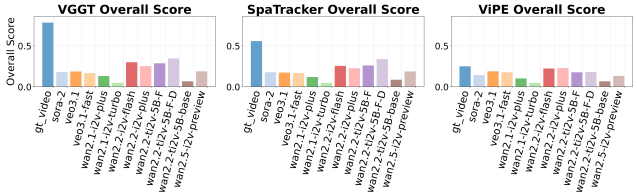


Figure 7. Overall score comparison between different spatio-temporal reconstruction tools. Detailed evaluation results with SpaTracker or ViPE can be found in the **appendix**.

confirms that decoded trajectories align well with ground truth trajectories, validating our evaluation approach.

4.5. Sensitivity to the Planning Horizon Length

To assess how the path planning horizon influences WM performance, Table 4 reports results for two WMs evaluated at three horizons: 8 s, 6 s, and 4 s. Reducing the horizon from 6 s to 4 s yields an 8% WO improvement for `Wan2.2-I2V-Flash`, while shortening it from 8 s to 6 s increases the weighted score of `Wan2.2-I2V-Plus` by more than 15%. Overall, the weighted score consistently improves as the horizon decreases, suggesting that WMs are more reliable when planning on shorter temporal windows.

5. Discussion

Benchmark Validity and Performance Ceiling. Ground truth videos achieve a score of 0.783 when decoded with VGGT (Table 1), establishing an upper bound for the WO score. The gap from a perfect score stems from inherent reconstruction limitations: (1) monocular scale recovery introduces systematic errors, and (2) feed-forward pose estimation struggles with motion blur. Critically, this ceiling is **not** a flaw: it proves that our pipeline correctly distinguishes high-quality inputs (0.783 for ground truth) from poor ones (0.066-0.299 for base models). As reconstruc-

World Model	SpaTracker [36]						ViPE [14]					
	FDE↓	ADE↓	MR↓	SE↑	AC↑	WO↑	FDE↓	ADE↓	MR↓	SE↑	AC↑	WO↑
gt_video	0.526	0.600	17.19	0.59	0.95	0.558	1.345	0.845	30.22	0.19	0.89	0.249
Sora 2	1.913	1.257	51.33	0.14	0.82	0.177	2.147	1.245	51.04	0.09	0.85	0.142
Veo 3.1	2.607	2.133	78.49	0.20	0.42	0.172	2.916	1.945	55.07	0.16	0.62	0.188
Veo 3.1-fast	3.402	2.150	62.89	0.15	0.59	0.166	2.601	1.715	58.76	0.15	0.62	0.175
Wan2.5-Preview†	2.379	1.510	61.93	0.18	0.88	0.187	2.686	1.608	59.52	0.10	0.83	0.132
Wan2.2-Plus†	1.409	1.631	73.19	0.25	0.68	0.225	**1.621**	**0.976**	**36.65**	**0.18**	0.80	**0.227**
Wan2.2-Flash†	1.407	1.572	71.06	0.28	0.71	0.254	1.687	1.006	37.42	0.17	0.81	0.222
Wan2.1-Plus†	2.810	2.110	63.08	0.08	0.67	0.118	2.858	1.510	53.70	0.04	0.76	0.102
Wan2.1-Turbo†	4.219	1.833	63.45	0.00	**1.00**	0.047	4.256	1.856	64.35	0.00	**1.00**	0.046
Wan2.2-TI2V-5B	3.661	1.666	57.05	0.03	0.89	0.085	3.773	1.705	60.27	0.01	0.91	0.066
Wan2.2-5B-FT§	1.342	0.951	38.53	0.23	0.74	0.259	1.980	1.079	42.76	0.12	0.80	0.176
Wan2.2-5B-FT-DA¶	**1.055**	**0.875**	**33.99**	**0.33**	0.75	**0.338**	1.841	1.023	39.01	0.11	0.83	0.180

Table 3. Evaluation results of SpaTracker and ViPE. † *Image-to-Video (I2V) models,* § *Wan2.2-TI2V-5B fine-tuned model,* ¶ *Wan2.2-TI2V-5B model fine-tuned with data augmentation.*

Model	Horizon	Metrics					WO ↑
		FDE ↓	ADE ↓	MR ↓	SE ↑	AC ↑	
2.2-Flash*	8s	1.362	1.005	38.75	0.292	0.746	0.299
	6s	1.393	0.762	23.76	0.261	0.579	0.290
	4s	1.278	0.713	23.18	0.290	0.574	0.314
2.2-Plus†	8s	1.377	1.044	45.25	0.240	0.686	0.252
	6s	1.390	0.789	25.36	0.263	0.584	0.291
	4s	1.400	0.783	25.91	0.265	0.534	0.292

Table 4. Evaluation results with different path planning horizons, with VGGT. * *Wan2.2-I2V-Flash,* † *Wan2.2-I2V-Plus.*

tion methods improve, this ceiling can be raised by substituting VGGT with better alternatives, making Target-Bench forward-compatible with future advances.

Current World Model Performance. The best off-the-shelf WM, Wan2.2-Flash, achieves only 0.299 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 challenging scenarios with implicit semantic targets. Since we evaluate one-time inference without replanning, WMs show promising potential for improvement.

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 reconstruction and low path accuracy. This holistic assessment fundamentally differs from existing benchmarks that isolate individual dimensions (visual quality, physics consistency) and miss their integration. **Our benchmark shows that perceptual realism does not guarantee path planning utility**, addressing a critical blind spot in current evaluation practices.

The Power of Targeted Domain Adaptation. Fine-tuning with merely 325 real-world robot scenarios produces remarkable gains: 423% improvement over the base model, and surpassing all off-the-shelf models. This result challenges conventional wisdom that larger models with more pre-training data always outperform smaller specialized models. Instead, our findings demonstrate that **high-quality domain-specific data matter more than scale**, for robotic planning tasks. The rapid adaptation (335% gain even without data augmentation) indicates that WMs have strong potential for spatial reasoning and require only targeted exposure to robot-specific tasks, suggesting pathways for deploying WMs in robotics.

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 [4, 37] 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.

6. Conclusion and Future Work

We introduced Target-Bench for evaluating 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 commercial 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 commercial WMs and improves over its base version by more than 400%. 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.

References

[1] Timur Akhtyamov, Mohamad Al Mdfaa, Javier Antonio Ramirez, Sergey Bakulin, German Devchich, Denis Fatykhov, Alexander Mazurov, Kristina Zipa, Malik Mohrat, Pavel Kolesnik, et al. Egowalk: A multimodal dataset for robot navigation in the wild. *arXiv preprint arXiv:2505.21282*, 2025. 3

[2] Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. In *Advances in Neural Information Processing Systems*, pages 58757–58791. Curran Associates, Inc., 2024. 2

[3] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, and et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning, 2025. 2

[4] Philip J. Ball, Jakob Bauer, Frank Belletti, Bethanie Brownfield, Ariel Ephrat, Shlomi Fruchter, Agrim Gupta, Kristian Holsheimer, Aleksander Holynski, Jiri Hron, Christos Kaplanis, Marjorie Limont, Matt McGill, Yanko Oliveira, Jack Parker-Holder, Frank Perbet, Guy Scully, Jeremy Shar, Stephen Spencer, Omer Tov, Ruben Villegas, Emma Wang, Jessica Yung, and et al. Genie 3: A new frontier for world models. 2025. 2, 8

[5] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 2

[6] Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *Proceedings of the 41st International Conference on Machine Learning*, pages 4603–4623. PMLR, 2024. 2

[7] Google DeepMind. Veo 3: Advanced controllable video generation with physics-aware dynamics, 2025. 2, 6

[8] Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. *arXiv preprint arXiv:2504.00983*, 2025. 2

[9] Yuan Gao, Mattia Piccinini, Yuchen Zhang, Dingrui Wang, Korbinian Moller, Roberto Brusnicki, Baha Zarrouki, Alessio Gambi, Jan Frederik Totz, Kai Storms, et al. Foundation models in autonomous driving: A survey on scenario generation and scenario analysis. *arXiv preprint arXiv:2506.11526*, 2025. 2

[10] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems 31*, pages 2451–2463. Curran Associates, Inc., 2018. 2

[11] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020. 2

[12] Noriaki Hirose, Dhruv Shah, Ajay Sridhar, and Sergey Levine. Sacson: Scalable autonomous control for social navigation. *IEEE Robotics and Automation Letters*, 9(1):49–56, 2023. 3

[13] Noriaki Hirose, Catherine Glossop, Ajay Sridhar, Dhruv Shah, Oier Mees, and Sergey Levine. Lelan: Learning a language-conditioned navigation policy from in-the-wild video. In *Proceedings of The 8th Conference on Robot Learning*, pages 666–688. PMLR, 2025. 3

[14] Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixe, and Sanja Fidler. Vipe: Video pose engine for 3d geometric perception. In *NVIDIA Research Whitepapers arXiv:2508.10934*, 2025. 2, 4, 8

[15] 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

[16] Haresh Karnan, Anirudh Nair, Xuesu Xiao, Garrett Warnell, Sören Pirk, Alexander Toshev, Justin Hart, Joydeep Biswas, and Peter Stone. Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation. *IEEE Robotics and Automation Letters*, 7 (4):11807–11814, 2022. 3

[17] Rainer Kümmerle, Giorgio Grisetti, Hauke Strasdat, Kurt Konolige, and Wolfram Burgard. G2o: A general framework for graph optimization. In *2011 IEEE International Conference on Robotics and Automation*, pages 3607–3613, 2011. 3

[18] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022. 2

[19] Chenhao Li, Andreas Krause, and Marco Hutter. Robotic world model: A neural network simulator for robust policy optimization in robotics. *arXiv preprint arXiv:2501.10100*, 2025. 2

[20] 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. *CoRR*, abs/2502.20694, 2025. 2

[21] Xinhao Liu, Jintong Li, Yicheng Jiang, Niranjan Sujay, Zhicheng Yang, Juexiao Zhang, John Abanes, Jing Zhang, and Chen Feng. Citywalker: Learning embodied urban navigation from web-scale videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6875–6885, 2025. 3

[22] Xiaoxiao Long, Qingrui Zhao, Kaiwen Zhang, Zihao Zhang, Dingrui Wang, Yumeng Liu, Zhengjie Shu, Yi Lu, Shouzheng Wang, Xinzhe Wei, et al. A survey: Learning embodied intelligence from physical simulators and world models. *arXiv preprint arXiv:2507.00917*, 2025. 2

[23] modelscope. Diffsynth-studio: examples/wanvideo, 2025. Accessed: 2025-11-14. 6

[24] Duc M Nguyen, Mohammad Nazeri, Amirreza Payandeh, Aniket Datar, and Xuesu Xiao. 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)*, pages 7442–7447. IEEE, 2023. 3

[25] NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, and et al. Cosmos: World foundation model platform for physical ai. *CoRR*, abs/2501.03575, 2025. 2

[26] Jing-Cheng Pang, Nan Tang, Kaiyuan Li, Yuting Tang, Xin-Qiang Cai, Zhen-Yu Zhang, Gang Niu, Masashi Sugiyama, and Yang Yu. Learning view-invariant world models for visual robotic manipulation. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[27] Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, Stephen Spencer, Jessica Yung, Michael Dennis, Sultan Kenjeyev, Shangbang Long, Vlad Mnih, Harris Chan, Maxime Gazeau, Bonnie Li, Fabio Pardo, Luyu Wang, Lei Zhang, Frederic Besse, Tim Harley, Anna Mitenkova, Jane Wang, Jeff Clune, Demis Hassabis, Raia Hadsell, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 2: A large-scale foundation world model. 2024. 2

[28] Aleksandr Segal, Dirk Haehnel, and Sebastian Thrun. Generalized-ICP. In *Robotics: Science and Systems*, 2009. 3

[29] The OpenAI Sora Team. 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. 2, 6

[30] Unitree. Unifolm-wma-0: A world-model-action (wma) framework under unifolm family, 2025. 2

[31] Sagar M Waghmare, Kimberly Wilber, Dave Hawkey, Xuan Yang, Matthew Wilson, Stephanie Debats, Cattalyya Nuengsigkapian, Astuti Sharma, Lars Pandikow, Huisheng Wang, et al. Sanpo: A scene understanding, accessibility and human navigation dataset. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7866–7875. IEEE, 2025. 3

[32] 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. *CoRR*, abs/2503.20314, 2025. 2, 6

[33] Dingrui Wang, Zhexiao Sun, Zhouheng Li, Cheng Wang, Youlun Peng, Hongyuan Ye, Baha Zarrouki, Wei Li, Mattia Piccinini, Lei Xie, et al. Enhancing physical consistency in lightweight world models. *arXiv preprint arXiv:2509.12437*, 2025. 2

[34] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5294–5306, 2025. 2, 4, 7

[35] Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025. 2

[36] Yuxi Xiao, Jianyuan Wang, Nan Xue, Nikita Karaev, Iurii Makarov, Bingyi Kang, Xin Zhu, Hujun Bao, Yujun Shen, and Xiaowei Zhou. Spatialtrackerv2: 3d point tracking made easy. In *ICCV*, 2025. 2, 4, 8

[37] Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. *arXiv preprint arXiv:2506.03141*, 2025. 8

[38] Jiahan Zhang, Muqing Jiang, Nanru Dai, Taiming Lu, Arda Uzunoglu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal M Patel, Paul Pu Liang, et al. World-in-world: World models in a closed-loop world. *arXiv preprint arXiv:2510.18135*, 2025. 2

[39] Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. *arXiv preprint arXiv:2504.02792*, 2025. 2