What if? Emulative Simulation with World Models for Situated Reasoning

Ruiping Liu[1], Yufan Chen[1], Yuheng Zhang[2], Junwei Zheng[1,3], Kunyu Peng[1,4], Chengzhi Wu[1], Chenguang Huang[5], Di Wen[1], Jiaming Zhang[2], Kailun Yang[2,*], and Rainer Stiefelhagen[1]

[1] Karlsruhe Institute of Technology, Germany
[2] Hunan University, China
[3] ETH Zürich, Switzerland
[4] INSAIT, Sofia University "St. Kliment Ohridski", Bulgaria
[5] RAI Institute, Switzerland

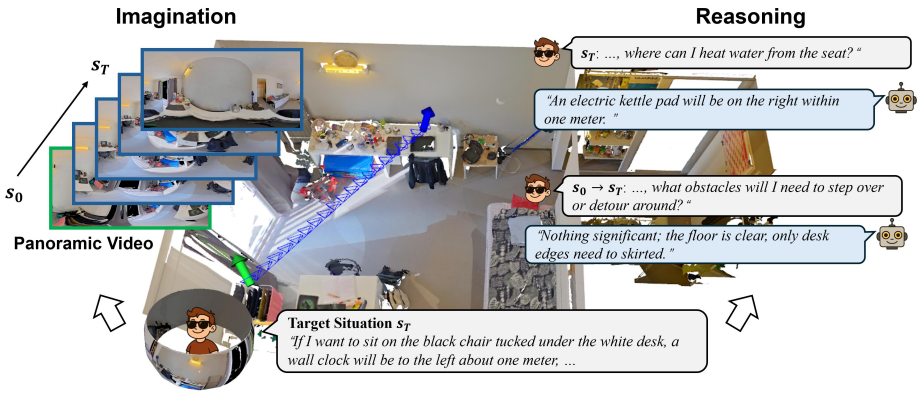


Fig. 1: Emulative simulation with WanderDream. Putting oneself in the mental shoes of the agent to imagine the visual trajectory from the current perception s_0 toward the target situation s_T, and reasoning along the imagined path to answer *"what-if"* questions. Throughout the paper, green denotes the current state, while blue represents imagination.

Abstract. Situated reasoning often relies on active exploration, yet in many real-world scenarios such exploration is infeasible due to physical constraints of robots or safety concerns of visually impaired users. Given only a limited observation, can an agent mentally simulate a future trajectory toward a target situation and answer spatial *"what-if"* questions? We introduce WanderDream, the first large-scale dataset designed for the emulative simulation of mental exploration, enabling models to reason without active exploration. WanderDream-Gen comprises 15.8K panoramic videos across 1,088 real scenes from HM3D, ScanNet++, and real-world captures, depicting imagined trajectories from current

* **Corresponding author:** kailun.yang@hnu.edu.cn; **First author:** ruiping.liu@kit.edu

viewpoints to target situations. WanderDream-QA contains 158K question–answer pairs, covering starting states, paths, and end states along each trajectory to comprehensively evaluate exploration-based reasoning. Extensive experiments with world models and MLLMs demonstrate (1) that mental exploration is essential for situated reasoning, (2) that world models achieve compelling performance on WanderDream-Gen, (3) that imagination substantially facilitates reasoning on WanderDream-QA, and (4) that WanderDream data exhibit remarkable transferability to real-world scenarios. The source code and all data are released at https://github.com/RuipingL/WanderDream.

Keywords: Emulative simulation · World model · Situated reasoning

1 Introduction

Situated reasoning [12], where inferences draw on the local context, the perceptual surroundings, and the grounding established through interaction [36], is a fundamental capability of the cognitive system in both embodied agents, *e.g.*, robots [19, 85, 115], and body-worn assistive agents, *e.g.*, wearable navigation assistants for people with visual impairments [50, 51]. However, existing situated reasoning approaches typically rely on either pre-explored scenarios [45, 55, 118] or refinement during active exploration [26, 101, 109], both of which inherently depend on physical exploration and are constrained by various real-world limitations, as illustrated in Fig. 2. Due to mechanical design differences, robots are subject to distinct physical constraints. For example, warehouse robots can only operate within flat-grid warehouse zones, must continually adapt to new obstacles [70], and are unable to navigate stairs [72] or uneven terrain [46]. In contrast, although more flexible in their body movement, visually impaired individuals could face psychological constraints. They may hesitate to explore further when they feel unsafe [59, 100], *e.g.*, when encountering obstacles that block their path [60]. Moreover, the explore-then-understand paradigm fails in dynamic environments [52], where continuous changes demand ongoing memory updates. In such cases, imagination can bridge the gap by enabling agents to understand their potential situations based on their current egocentric view without additional physical movement. World models can serve as the imagination engine [18, 39, 127] for answering these situated *"what-if"* questions.

Mental imagination is divided into two layers [58]: instrumental simulation and emulative simulation, as illustrated in Fig. 3. Instrumental simulation is task-oriented to facilitate decision making and has been widely explored with world models, *e.g.* in imagination-based navigation [3, 32, 61, 80] and action reasoning [5, 121]. In contrast, **emulative simulation** is experience-oriented and serves as the core for answering *"what-if"* questions by placing oneself in the mental shoes to explore the scene and reason along the path, yet it remains underexplored. MindJourney [108] couples a world model with an MLLM for step-wise visual imagination to answer questions about view changes. GenEx [54] introduces a dataset of forward-moving panoramic videos to reason about the

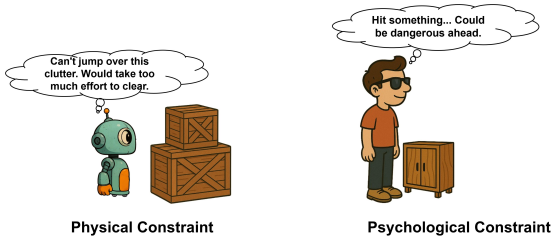


Fig. 2: Constraints of active exploration: robot embodiment limits (*e.g.*, inability to climb stairs) and visually impaired users' psychological safety barriers when encountering obstacles without tactile cues.

current state. However, there is no existing dataset that provides temporally consistent video toward target situations for imaginative video generation, together with reasoning information along the path to enable emulative simulation.

We introduce **WanderDream**, a large-scale dataset, as the first benchmark for studying emulative simulation. It comprises WanderDream-Gen, which provides panoramic trajectories from current viewpoints to imagined target situations, and WanderDream-QA, which contains question–answer pairs for evaluating reasoning along these imagined trajectories. To support human–robot collaboration, world models are expected to imagine situations from both perspectives, *e.g.*, when a robot aims to navigate to a chair while a human intends to sit on it. We collect robot navigation situations from HM3D and human action situations from ScanNet++, yielding 15.8K videos across 1,088 real scenes. For reasoning along trajectories, we design 10 QA types covering the start state, the path, and the end state, resulting in a total of 158K QA pairs.

Extensive experiments with various world models and MLLMs under different frameworks are conducted to verify that imagination is essential for situated reasoning, to evaluate the performance of world models on WanderDream-Gen, to examine how world-model-based imagination enhances reasoning on WanderDream-QA, and to assess the transferability of WanderDream data to the real world. The results show that although the location of the target situation can be reasoned from the current perception, imagination remains essential for situated reasoning. Moreover, world models that perform better on WanderDream-Gen tend to enable stronger reasoning on WanderDream-QA. Despite differences between WanderDream and real-world recordings, such as partial occlusions caused by the real agent, WanderDream exhibits strong transferability for both video generation and reasoning tasks.

2 Related Work

Situated scene understanding. Prior scene understanding works focus primarily on allocentric relations [29, 49, 86, 98, 128], neglecting the egocentric, situated perception of the agent. Recent efforts in situated reasoning attempt

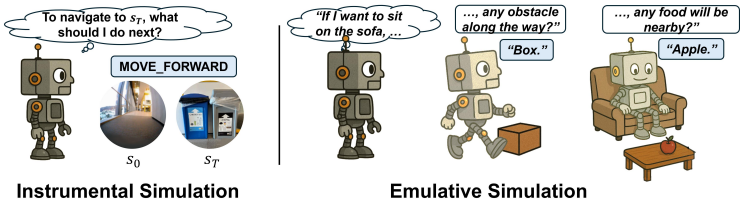


Fig. 3: Two layers of mental imagination. Task-oriented instrumental simulation (left), such as Navigation World Models [3], and experience-oriented emulative simulation (right), empowered by the proposed WanderDream.

Table 1: Datasets for situated reasoning. Visual quality: photographic *vs.* photorealistic. Task inputs and outputs include 3D point cloud, perspective video, panoramic video, multi image, panoramic image, audio, and text. Visual modalities: RGB, point cloud (pcd), depth map (D), and semantic map (S).

Dataset	Venue	Type	#Scene	#Video	#QA Types	#QA	Input	Output	Modalities
SQA3D [55]	ICLR'23		650	-	6	33.4K			RGB, pcd
MSQA [45]	NeurIPS'24		1734	-	8	251K			RGB, pcd
Situat3DChange [52]	NeurIPS'25		903	-	9	121K			RGB, pcd
Pano-AVOA [114]	ICCV'21		n/a	5.4K	2	51.7K			RGB
SAT [68]	COLM'25		22K	-	6	175K			RGB
MMSI-Bench [107]	arXiv'25		n/a	-	11	1K			RGB
OpenEQA [56]	CVPR'24		187	187	7	1.6K			RGB
VSI-Bench [104]	CVPR'25		288	288	8	5K			RGB
DSI-Bench [119]	arXiv'25		-	1K	6	1.7K			RGB
VSI-590K [106]	arXiv'25		n/a	6K	8	590.7K	/		RGB, S
WanderDream (Ours)	-		1088	15.8K	10	158.1K			RGB, D, S

to imagine situations within pre-explored static scenarios. Datasets such as SQA3D [55], MSQA [45], Spartun3D [118], and SURPRISE3D [28] support egocentric reasoning in such contexts. Corresponding methods [27, 57, 85] successfully solve these tasks, whereas HIS-GPT [120] introduces a benchmark for reasoning about virtual scenes involving a virtual human. Other datasets, including Situat3DChange [52] and SOK-Bench [78], emphasize temporal evolution and situational change. Another line of work tackles situated reasoning through video streams acquired during active exploration [53, 56, 66, 106, 113, 114, 123], either by direct perception or through representations such as code maps [104], image graphs [23, 109], online query embeddings [129], or situational scene graphs [74]. These approaches support decision-making during exploration, *e.g.*, in navigation [31]. Multi-view frameworks like AffordBot [85] and Omni-View [22] leverage previously seen frames for gaze and affordance reasoning, whereas STAR [91] and SpatialLadder [41] introduce learning curricula bridging perception and reasoning. ODI Bench [105] and CFPano [116] benchmark the understanding of panoramic imagery. In contrast, we address situated reasoning through emulative imagination with world models, allowing agents to reason without memory from active exploration, especially in inaccessible environments. A comparison of related datasets is shown in Tab. 1.

Video generation and understanding. Video generation models [4, 20] are increasingly vital for producing realistic, temporally coherent visual content in applications such as entertainment, simulation, robotics, and data augmentation. Recent advances have moved from frame-by-frame synthesis to spatio-temporally consistent generation using diffusion-based architectures [4, 20, 92, 97, 124]. Modern approaches enhance controllability [24, 33, 54, 81], enable multimodal conditioning [8, 43], and improve real-world fidelity [25], yielding high-quality, semantically aligned outputs. Unified frameworks deal with video generation and understanding separately, and have emerged in pinhole video settings [7, 75, 95], while panoramic video generation [17, 47, 82, 83, 90, 94] is gaining traction for its immersive 360° views and richer scene understanding [6, 117]. While front views are limited when facing room corners or large objects, panoramic views are widely adopted in real-world edge devices [73, 87, 89, 93]. We therefore use panoramic videos to facilitate visual imagination with a wider field of view, while also releasing single-view videos.

World models for spatial intelligence. World models aim to internalize environmental dynamics, enabling agents to predict, plan, and act through imagination rather than direct perception [13]. Early work focuses on generating coherent virtual worlds [9, 38, 42, 65, 76, 125] that support free user exploration. Following the two-layer structure of imagination [58], most current world models for real-world spatial intelligence remain within instrumental simulation, which predicts successive states needed to complete a task. Navigation World Models [3] and PathDreamer [32] synthesize future views from an initial image and a sequence of given actions. Models including 3D VLA [121], WorldVLA [5], OccLLaMA [88], DrivingGPT [10], and AdaWorld [16] integrate vision, language, and action. EgoTwin [99] predicts both egocentric views and human poses from stepwise action descriptions. Theoretical insights from Richens *et al.* [69] emphasize that robust intelligence requires causal world modeling, while benchmarks such as WAGIBench [77] highlight embodied goal inference in rea- world settings. In this work, we focus on the second layer, emulative simulation, which imagines the visual experience along the path to an *"if"* target situation and supports answering *"what-if"* questions.

3 WanderDream

WanderDream supports emulative simulation through two components: WanderDream-Gen (Sec. 3.1) for spatial imagination from the current state to the target situation, and WanderDream-QA (Sec. 3.2) for reasoning along the trajectory. Data quality is ensured through rigorous control, and dataset statistics are provided in Sec. 3.3. A small real-world test set (Sec. 3.4) is included to assess sim-to-real transfer and to support future evaluation.

3.1 WanderDream-Gen

To build an imagination engine that understands both robotic and human situations for effective human-robot collaboration, we consider the distinct sit-

uations of each. Robots navigate toward landmarks to support further movement, whereas humans interact with nearby objects. Prior situated reasoning work [45,52] uses an anchor object to locate the situation, and Epstein *et al.* [14] report that imagined trajectories within the cognitive map follow the shortest path. Following these principles, we treat object navigation as *robotic situated path imagination* on HM3D [67] and spatial shortest path prediction as *human situated path imagination* on ScanNet++ [111].

Robotic situated path imagination. For robotic agents, we select salient landmarks from 19 classes and generate ground-truth object-navigation paths using Habitat-Sim's default shortest-path planner [71], discretized into actions: `MOVE_FORWARD` ($0.25m$), `TURN_LEFT` ($10°$), and `TURN_RIGHT` ($10°$), as illustrated in the first row of Fig. 4. The maximum path length is $5m$.

Human situated path imagination. For human situation, we adopt situation definitions from [45, 52], including: `interacting`, `standing`, and `sitting`. Given human physical flexibility, such as the ability to step over obstacles like trash cans instead of navigating around them like robots, trajectory prediction becomes more complex. Due to the small room sizes in the ScanNet++ domain, a random navigable starting point is sampled at a distance of $1.5m$∼$3m$ for each situation. When no non-traversable obstacles exist, we employ direct path interpolation for both position and orientation as ground truth (middle row of Fig. 4). To mimic the shortest path when non-traversable obstacles are present, we build a 3D Probabilistic Roadmap (PRM) and run Dijkstra's algorithm on capsule-validated edges with vertical penalties. Invalid segments are revalidated and locally repaired using micro-PRM (bottom row of Fig. 4). To mimic the distortion characteristics of real-world egocentric panoramas, we apply random head pitch variations within $\pm 30°$ at the start frame. Since the end state is imagined, the view is kept horizontally aligned for `standing` and `sitting` situations, while during `interacting`, the head is oriented toward the target object with pitch constrained within $\pm 30°$. In addition to the RGB videos, the corresponding depth maps, semantic maps, and camera poses will also be released.

3.2 WanderDream-QA

QA categories. For each video, exactly ten questions are distributed along the trajectory, which is divided into three phases: the start state (s_0) with three questions, the path phase ($s_0 \rightarrow s_T$) with four questions, and the end state (s_T) with three questions. Each question belongs to a specific reasoning category, inspired by existing situated reasoning benchmarks [30, 48, 54, 56, 68].

At the start state, we evaluate *Object Awareness*, focusing on nearby objects within a local range to help localize the agent and understand its immediate surroundings. *Navigability Reasoning* assesses whether paths are clear in the four cardinal directions, whereas *Ego-Target Orientation* captures the overall spatial relationship between the agent and the target situation.

During the path phase, *Landmark Sequencing* evaluates the order in which landmarks are encountered along the trajectory. *Spatial Estimation* measures the path length to estimate the effort needed to reach the target. *Obstacle Reasoning*

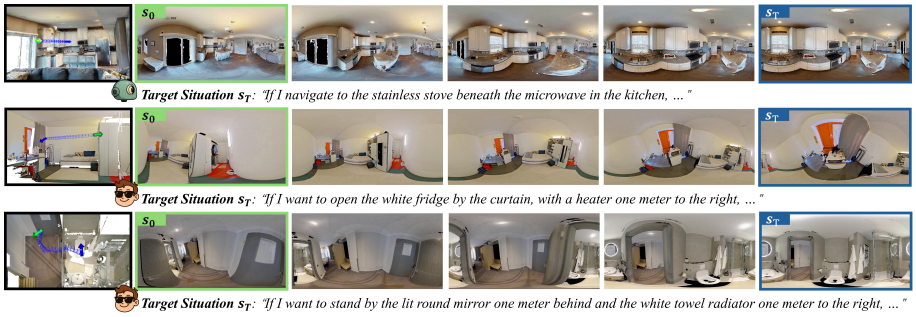


Fig. 4: WanderDream-Gen. Top row: Object navigation as robotic situated path imagination in HM3D. Middle row: Human-situated perspective with direct interpolation as the shortest path when no non-traversable obstacles are present. Bottom row: Human-situated perspective with computed shortest path accounting for non-traversable obstacles (*e.g.*, walls).

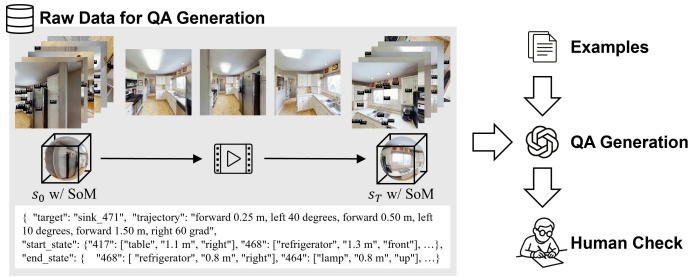


Fig. 5: WanderDream-QA generation pipeline with an example in HM3D.

identifies obstacles that need to be stepped over by humans and those block robot movement. For robots, *Route Planning* determines the necessary turning and forward movements. For humans, where path planning is less natural, we INSTEAD consider *Relative Distance Change*, indicating whether the agent is moving closer to or farther from a specific object along the route to the target.

At the end state, *Affordance* determines whether functional objects are present near the target situation. *Egocentric Spatial Relationship* describes the direction and distance of specific objects relative to the agent's final position, and *Object Proximity* compares the relative distances between pairs of objects from the agent's viewpoint.

QA generation. We follow established strategies [21, 45, 112] and use GPT-5 [64] to generate the large-scale WanderDream-QA, as shown in Fig. 5. We adopt ground-truth-annotated Set-of-Mark (SoM) [11, 103] to help GPT-5 ground object instances in the image. Instance IDs are overlaid on cubemaps of the start and end states, and front-view images along the path are provided for orientation. In addition, a JSON file supplies trajectory metadata, with direction and distance at both the start and end states, to provide structured spatial context.

Table 2: Dataset statistics of WanderDream. Trajectory length is in *meters*, and text length is in *words*. Real recordings from two scenes are not included in the table.

Subset	WanderDream-Gen					WanderDream-QA		
	Train scenes	Train situation videos	Val. scenes	Val. situation videos	Avg. trajectory length	Avg. situation length	Avg. question length	Avg. answer length
ScanNet++	856	8274	50	478	2.19 ± 0.58	22.29 ± 2.61	$8,81 \pm 1.52$	12.19 ± 3.93
HM3D	144	5632	36	1404	2.62 ± 0.84	14.98 ± 1.99	8.96 ± 1.70	13.37 ± 3.48
Overall	1000	13906	86	1882	2.38 ± 0.74	19.03 ± 4.33	8.88 ± 1.60	12.72 ± 3.78

3.3 Data Quality Control and Statistics

For WanderDream-Gen, we initially sample 40 situations per HM3D scene and 10 per ScanNet++ scene, followed by filtering. Following ReferIt3D [1], we exclude from anchor selection any category with more than six distractors of the same-category within a region in both subsets. In addition, we discard ScanNet++ scenes that are too small for imagination. Each orientation of the target situation in ScanNet++ is double-checked by annotators, with reorientation applied to specific cases, such as adjusting the orientation of the situation *'sitting on a piano chair'* to face the *piano*. Situations whose anchor object is not visible in the semantic map of the final frame are automatically filtered out.

For WanderDream-QA, human annotators first corrected erroneous answers, *e.g.*, those with instance IDs or redundant phrases. Then, 80 situations with 800 QA pairs were sampled and rated for situation quality, question quality, and answer accuracy on a Likert scale (1 to 5), yielding averages of 4.83±0.38, 4.88±0.35, and 4.73±0.47, respectively, indicating high textual quality. The annotators were independent of this work and will be acknowledged.

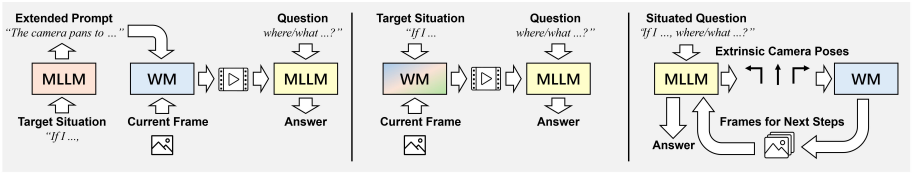
The dataset statistics are shown in Tab. 2. We fix the video length to 21 frames following the $4N+1$ convention [34, 79]. The equirectangular panoramas have a resolution of 1024×2048.

3.4 Real-World Test Set

To evaluate the sim-to-real transferability of WanderDream data and to support future studies on emulative simulation, we recruited a human explorer who wore a panoramic head-mounted camera to record 26 videos in real environments, including an office and an apartment with living room and kitchen areas. To focus on the role of imagination in facilitating reasoning, we generated 7 questions per trajectory, covering the path $s_0 \rightarrow s_T$ and end states s_T, resulting in 182 QA pairs in total, which were subsequently refined by human annotators. The scale is comparable to the real-world test set in SAT [68] (150 QA pairs) for a related task. Videos are also resampled to 21 frames.

4 Frameworks and Metrics for Emulative Simulation

Sequential frameworks for trajectory simulation. As no open-source unified model can take a question and an image and jointly output a video sequence



(a) Sequential Framework with Prompt Extension for World Model.

(b) Sequential Framework Using World Model with WanderDream Knowledge.

(c) Closed-Loop Framework with Step-wise Camera Control.

Fig. 6: Frameworks for emulative simulation. (a) and (b) are the sequential frameworks we use to imagine a consistent view trajectory. (c) is the closed-loop framework of MindJourney [108], which generates novel views step by step. Yellow MLLM modules are for reasoning, while the pink one is for prompt extension.

and an answer, we design a sequential framework that combines a world model and an MLLM for imagination and reasoning. We adopt leading world models in V-Bench-2.0 [122], including HunyuanVideo [34], CogVideoX [110], and Wan [79], but their foreground-centered training causes zero-shot queries for target situations to produce static frames. To enable controlled camera motion, we use MLLM-driven prompt-extension scripts to decompose target trajectories into action sequences (Fig. 6a), a strategy widely used in both open-source [79, 110] and commercial video generation models [37, 62]. We also fine-tune models on WanderDream (Fig. 6b). For reasoning, we use Qwen3-VL-32B [102] and LLaVA-OneVision-1.5-8B [2], which perform competitively on Video-MME [15]. These sequential frameworks achieve implicit camera pose control toward the target situation and answer questions along the trajectory.

Closed-loop framework for step-wise simulation. In contrast to our sequential framework that imagines the entire trajectory at once, MindJourney [108] performs emulative simulation per question through explicit step-wise camera control. An MLLM interprets each situated question and proposes successive camera actions in a closed loop (Fig. 6c). Because it imagines per question via test-time scaling rather than per situation, we evaluate it only on the real-world test set. Its imagination module uses the front-view novel-view synthesis model SVC [126], which fails in occluded or corner regions. To adapt it to our panoramic setting, we decompose panoramas into directional views and let the framework select the most informative view for reasoning.

Metrics. We carefully select video generation metrics to evaluate imagination in WanderDream-Gen: FVD for trajectory coherence, End-FID for target-state prediction accuracy, and Spherical SSIM and LPIPS for geometric and perceptual consistency between generated videos and ground truth. Since our work focuses on imagination rather than navigation, we do not leverage traditional navigation metrics (*e.g.*, path-length- or -weighted metrics). While the shortest-path assumption is suitable for cognitive map modeling, it does not necessarily reflect subjective real-world trajectories, for which no definitive ground truth exists.

To evaluate the long-form questions on WanderDream-QA, we adopt an LLM-as-a-judge [40] evaluation to assess factual correctness. Following prior

works [45, 52, 112], we use a GPT-based scoring framework to uniformly rate model responses:

$$C = \frac{1}{N} \sum_{i=1}^{N} \frac{s_i - 1}{4} \times 100\%, \tag{1}$$

where C is the overall correctness over N samples, and $s_i \in [1, 5]$ is the GPT-assigned rating given the question, ground-truth answer, and model response. Human evaluation over 800 QA pairs closely aligns with GPT, with a very strong Spearman correlation of 0.9722. Additional ablations supporting the rationale for the metric are provided in the supplementary material.

5 Experiments

We first address emulative simulation, where agents imagine the mental journey toward a target situation and reason about *"what-if"* questions along the path. Our experiments investigate: (1) whether answering *"what-if"* questions requires imagination; (2) how world models perform in imagining view trajectories toward target situations on WanderDream-Gen; (3) how imagination facilitates reasoning on WanderDream-QA; and (4) the transferability of WanderDream to real-world data.

5.1 Implementation Details

For fine-tuning to inject WanderDream knowledge (Fig. 6b), we train videos at a resolution of 384×768, limited by available computational resources. Some models require fixed aspect ratios, resulting in outputs smaller than 384×768. For consistent evaluation on WanderDream-Gen, all videos are resized to 256×512. The generated videos used for reasoning are sampled every five frames ($s_{\Delta 5}$) to reduce biases introduced by different video-processing pipelines in MLLMs. The world models (Sec. 4) are fine-tuned on both subsets of WanderDream-Gen for 8 epochs using LoRA. For CogVideoX, LoRA is insufficient due to its image–preprocessing pipeline, so we apply supervised fine-tuning (SFT) for 10 epochs instead. As CogVideoX uses an $8N{+}1$ frame scheme with $N{=}3$, we remove four redundant frames during evaluation on WanderDream-Gen to maintain temporal alignment. All other training settings follow their original implementations and are detailed in the supplementary material.

For all frameworks, we use Qwen3-VL as the reasoning module unless otherwise stated. MLLMs are used in a few-shot setting to answer situated questions.

5.2 Results

Necessity of imagination for *"what-if"* reasoning. To investigate this, we input different sets of frames from the ground-truth videos to MLLMs: (1) only the current frame s_0; (2) two frames representing the current and target states, s_0, s_T; (3) frames sampled at an interval of 5, resulting in five frames in

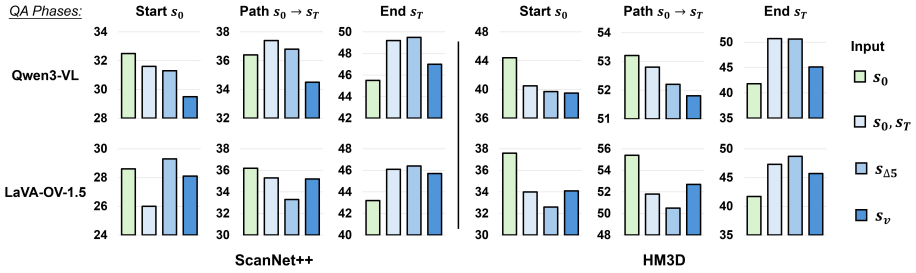


Fig. 7: Results of MLLMs under different video input settings for answering questions across the phases of WanderDream-QA, used to verify the necessity of imagination for reasoning along the trajectory toward the target situation.

Table 3: Results on WanderDream-Gen. * indicates results may be affected by the removal of redundant frames.

Model	Setting	ScanNet++				HM3D			
		FVD ↓	End-FID ↓	S-SSIM ↑	LPIPS ↓	FVD ↓	End-FID ↓	S-SSIM ↑	LPIPS ↓
Wan2.2	+ PE	18.73	87.02	0.41 ± 0.10	0.57 ± 0.06	17.62	46.86	0.34 ± 0.09	0.56 ± 0.06
CogVideoX1.5	+PE	34.48*	85.57	0.43 ± 0.09	0.55 ± 0.04	38.41*	65.24	0.33 ± 0.08	0.56 ± 0.05
HunyuanVideo	+LoRA	13.42	82.82	**0.47** ± 0.10	0.52 ± 0.06	14.50	62.01	**0.43** ± 0.09	0.49 ± 0.06
Wan2.1	+LoRA	**7.90**	70.56	**0.47** ± 0.10	**0.50** ± 0.06	**5.96**	40.20	0.39 ± 0.09	**0.47** ± 0.05
Wan2.2	+LoRA	9.67	77.50	0.45 ± 0.09	0.51 ± 0.07	6.19	**40.13**	0.38 ± 0.08	**0.47** ± 0.06
CogVideoX1.5	+SFT	16.96*	**61.49**	0.43 ± 0.09	0.52 ± 0.05	22.16*	42.74	0.34 ± 0.08	0.53 ± 0.05

total ($s_{\Delta 5}$); and (4) the full video file s_v, which may be influenced by varying frame-sampling strategies across different MLLMs.

The average WanderDream-QA scores across three trajectory phases on both subsets are shown in Fig. 7. MLLMs often suffer from contextual interference [44] when multiple frames add irrelevant or redundant visual information that distracts them from key cues. Consequently, for start-state questions, using only the current frame s_0 should give the highest accuracy, whereas for end-state questions, the combination s_0, s_T is expected to perform best.

The first assumption holds. With our short trajectories, s_0 *allows the MLLMs understand the surroundings and plan toward the target, while it cannot interpret the end state.* Notably, the second assumption does not hold. The model with $s_{\Delta 5}$ performs on-par with or even better than the s_0, s_T input when reasoning about the end state, which suggests that *intermediate imagined frames also strengthen the understanding of the final situation.* Overall, the importance of imagination increases along the trajectory. **Performance of world models on WanderDream-Gen.** Tab. 3 summarizes the performance of world models that incorporate target situations either through prompt extension or through fine-tuning (LoRA or SFT) on the WanderDream-Gen validation set. Wan2.1 achieves the best overall performance, with strong end-state estimation (End-FID). CogVideoX1.5 and Wan2.2, after fine-tuning on WanderDream, provide the leading end-state estimation on HM3D and ScanNet++, respectively.

Table 4: Results on WanderDream-QA using ScanNet++ captured videos and human-perspective QA types. The first row shows the input with only the start-state frame s_0, without imagination.

Model	Setting	Start s_o				Path $s_0 \rightarrow s_T$					End s_T			
		OA	NR	EO	Avg.	LS	SE	OR	DC	Avg.	OP	Aff.	ESR	Avg.
-	-	**20.4**	**45.1**	**32.0**	**32.5**	**23.7**	53.0	**30.1**	38.7	36.4	60.1	39.1	37.4	45.5
Wan2.2	+PE	20.0	42.8	30.2	31.0	22.8	53.8	29.2	39.2	36.2	60.0	39.9	36.8	45.6
CogVideoX1.5	+PE	**20.4**	42.1	31.5	31.7	23.2	**57.0**	29.3	38.5	37.0	60.1	41.0	36.6	45.9
Wan2.2	+LoRA	19.4	40.6	30.7	30.2	23.0	54.0	28.7	41.0	36.7	60.2	40.3	37.3	45.9
Wan2.1	+LoRA	18.1	42.0	29.1	29.7	23.1	56.0	28.3	39.3	36.7	**60.8**	38.0	**39.7**	46.2
HuyuanVideo	+LoRA	19.2	41.8	29.8	30.3	22.3	54.6	29.7	40.4	36.8	59.4	39.4	39.2	46.0
CogVideoX1.5	+SFT	19.5	42.5	29.9	30.6	23.5	56.0	29.6	**41.4**	**37.6**	60.5	**42.6**	39.5	**47.5**

Table 5: Results on WanderDream-QA using HM3D-captured videos and robot-perspective QA types. The first row shows the input with only the start-state frame s_0, without imagination.

Model	Setting	Start s_o				Path $s_o \rightarrow s_T$					End s_T			
		OA	NR	ED	Avg.	LS	SE	OR	RP	Avg.	OP	Aff.	ESR	Avg.
-	-	**27.4**	**63.2**	42.5	**44.4**	76.0	53.5	**50.8**	32.5	53.2	54.8	44.5	26.0	41.8
Wan2.2	+PE	24.6	60.8	**44.5**	43.3	76.6	56.1	44.7	34.9	53.1	54.0	43.9	26.9	41.6
CogVideoX1.5	+PE	26.7	60.9	44.4	44.0	77.6	57.2	45.1	33.5	**53.4**	55.1	44.4	25.5	41.7
Wan2.2	+LoRA	20.3	55.8	42.1	39.4	77.8	57.9	41.3	35.3	53.1	56.5	45.2	**29.1**	43.6
Wan2.1	+LoRA	20.8	57.7	43.3	40.6	76.8	**58.2**	42.6	**35.9**	**53.4**	56.7	45.0	28.2	43.3
HunyuanVideo	+LoRA	22.3	59.3	44.4	42.0	**79.0**	55.4	44.1	33.6	53.0	56.6	43.0	28.0	42.5
CogVideoX1.5	+SFT	20.8	57.7	41.9	40.1	76.4	57.8	41.6	34.1	52.5	**57.6**	**45.4**	**29.1**	**44.0**

Wan2.2 with camera control via prompt extension performs particularly well on HM3D, delivering superior video quality and end-state prediction without prior knowledge, outperforming fine-tuned CogVideoX1.5 in FVD (17.62 compared with 22.16) and HunyuanVideo in End-FID (46.86 compared with 62.01).

Impact of imagination on WanderDream-QA reasoning. Tab. 4 and Tab. 5 show the reasoning results on WanderDream-QA. The videos generated by the world models are uniformly sampled to construct $s_{\Delta 5}$, while retaining the start and end frames. Although the sole start frame input s_0 remains dominant for start-state reasoning due to contextual interference, *we observe that models with higher video generation quality also provide stronger support for reasoning along the trajectory.* CogVideoX1.5 trained with SFT, which produces accurate end states on the ScanNet++ subset (61.49 End-FID) and competitive results on HM3D (42.74 End-FID), achieves the highest path reasoning score (37.6) and end-state reasoning score (47.5) on ScanNet++, as well as strong end-state reasoning (44.0) on HM3D. Wan2.1 trained with LoRA, which maintains consistently high video quality across all metrics, obtains the highest path reasoning score (53.4) on HM3D and the second-highest end-state reasoning score (46.2) on ScanNet++. Wan2.2 also demonstrates strong generation and reasoning performance at the end state on HM3D.

Table 6: Sim-to-real results on real-world test set. The panorama is decomposed into directional views to adapt to MindJourney.

Framework (World Model)	Video Generation				QA
	FVD ↓	End-FID ↓	S-SSIM ↑	LPIPS ↓	
-	-	-	-	-	38.5
MindJourney [108] (SVC)	-	-	-	-	31.9
Sequential (Wan2.2 +PE)	41.65	178.61	0.36 ± 0.06	0.57 ± 0.08	38.8
Sequential (Wan2.1 +LoRA)	**27.49**	**175.98**	**0.38** ± 0.05	**0.54** ± 0.04	**43.0**

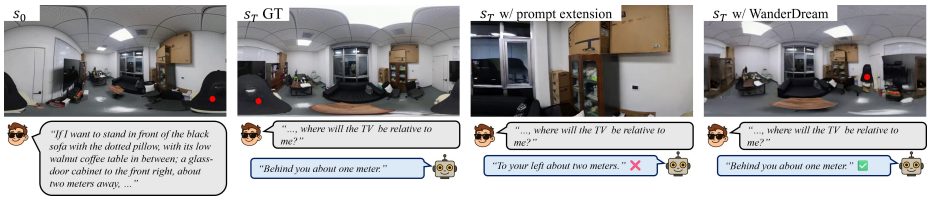


Fig. 8: Sim-to-real qualitative results. Red dots (•) indicate the position of the explorer and the generated artifacts caused by it.

Transferability of WanderDream to real-world data. We use Wan2.1 fine-tuned with LoRA and Wan2.2 with prompt-extended camera control for the sim-to-real video generation and QA. MindJourney, a closed-loop framework that imagines and reasons step by step without consistent video generation, is evaluated only for QA.

The results in Tab. 6 show that Wan2.1, fine-tuned on WanderDream, surpasses Wan2.2 with prompt extension in both video generation and QA. Although the gains in End-FID, S-SSIM, and LPIPS are marginal, fine-tuning brings a clear improvement in overall video quality (FVD) together with a +4.2% increase in QA accuracy. Surprisingly, even though real human motion in captured videos does not follow the shortest path, training on imagined shortest-path trajectories in WanderDream still produces a notable FVD improvement that better mimics video dynamics. The real-world FVD is worse than on the WanderDream validation set. We attribute this to suboptimal real-world trajectories and varying velocity along the trajectory. MindJourney performs even worse than the MLLM that receives only the start panorama.

Fig. 8 shows a qualitative example. The fine-tuned model sometimes misinterprets agents, such as *generating the explorer's hat as an object on the door*, but still produces plausible layouts and correct answers. In contrast, prompt extension helps to plan roughly correct directions, *e.g.*, *toward the window with the cabinet on the right*, but generates front-view images instead of panoramas and places the camera in the wrong position, which leads to incorrect spatial understanding. Despite agent occlusions and discrepancies between imagined trajectories and real exploration, WanderDream shows promising transferability for imagination and reasoning on our real-world test set, while larger-scale validation across more diverse environments is left for future work.

6 Limitations and Future Work

Failure Analysis: Video Generation for Reasoning. While more qualitative results are provided in the supplementary, Fig. 9 illustrates representative failure cases. Incorrect anchor localization leads to erroneous spatial relations. Under severe occlusion, models may still infer global layouts, but fine-grained details are lost, degrading performance on Affordance and Egocentric Spatial Relationship questions, which require awareness of specific objects.

Latency in Foundation Models. WanderDream imagines trajectories consistently, rather than answering per-question prompts [108], so it could reduce latency when many queries are associated with a single trajectory. However, video-generation foundation models still incur substantial inference latency: we report per-trajectory runtimes of Wan 1.3B (35s), Wan 5B (53s), HunyuanVideo (211s), and CogVideo (283s). While efficiency is not the claim of this work, more efficient world models could further strengthen the emulative simulation setting.

Reliability of Real-World Evaluation and Shortest-Path Supervision. Our real-world set contains 182 QA pairs, exceeding the 150 in SAT [68], a widely used view-comparison benchmark. We deliberately prioritize quality over scale: from a large pool of collected videos, we apply strict filtering (*e.g.*, stable lighting, clear anchor objects) to build a reliable sim-to-real testbed, while the large-scale validation set serves as our primary benchmark. For training, since ground truth for imagined trajectories is inherently unavailable, spatial shortest paths offer a controlled and reproducible supervision signal. This choice is also supported by classic mental-imagery findings: the mental scanning experiments of Kosslyn *et al.* [35] show that imagined trajectories preserve metric spatial structure such as straight-line distances, consistent with treating shortest paths as a proxy for imagined exploration. It is further justified empirically: as shown in Tab. 6, training on WanderDream shortest-path data aligns substantially better with real-world paths than prompt extension (FVD 27.49 *vs.* 41.65), demonstrating effective sim-to-real transfer.

Toward Unified Video-and-Text Modeling. WanderDream's emulative simulation takes the current egocentric view, a target-situation description, and a question as input, and outputs an imagined video and a corresponding answer. Existing end-to-end models [84,95,96] typically use a unified decoder to generate either video or text, but not both. We therefore adopt a GPT-style tool-calling pipeline [63] and provide all pipeline and evaluation prompts in the supplementary material. We encourage future work toward unified end-to-end solutions.

7 Conclusion

In this work, we address emulative simulation, where a virtual agent mentally explores the environment and reasons along the way. We introduce WanderDream, which consists of WanderDream-Gen for training and evaluating world models in imagining paths to target situations, and WanderDream-QA for assessing MLLMs in reasoning along imagined trajectories. Experiments show that imagination facilitates reasoning, confirming the correlation between world-model

Fig. 9: Representative failure cases: anchor confusion, occlusion-induced detail loss, and long-range spatial compression.

imagination and MLLM reasoning, and on our curated real-world test set the dataset exhibits sim-to-real transferability under occlusions and nonideal trajectories, though broader generality remains to be verified on larger and more diverse real-world data. We hope this work inspires further progress in imagination for exploring inaccessible real-world areas, and in enabling models to interpret human commands in virtual worlds, visualize imagined trajectories, and reason accordingly.

Acknowledgments

We thank Weicheng Dai, Jingqi Zhang, and Zirui Wang for joining the human annotation and evaluation. This work was supported in part by the Ministry of Science, Research and the Arts of Baden-Württemberg (MWK) through the Cooperative Graduate School Accessibility through AI-based Assistive Technology (KATE) under Grant BW6-03, in part by funding from the pilot program Core-Informatics of the Helmholtz Association (HGF), in part by Karlsruhe House of Young Scientists (KHYS), and in part by the Helmholtz Association Initiative and Networking Fund on the HAICORE@KIT and HOREKA@KIT partition. This work was supported in part by the National Natural Science Foundation of China (Grant No. 62473139, Grant No. 62503166), in part by the Hunan Provincial Research and Development Project (Grant No. 2025QK3019, Grant No. 2026QK3018), in part by the State Key Laboratory of Autonomous Intelligent Unmanned Systems (the opening project number ZZKF2025-2-10), in part by the Yuelushan Industrial Innovation Center, and in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - SFB 1574 - 471687386. This research was partially funded by the Ministry of Education and Science of Bulgaria (support for INSAIT, part of the Bulgarian National Roadmap for Research Infrastructure).

References

1. Achlioptas, P., Abdelreheem, A., Xia, F., Elhoseiny, M., Guibas, L.J.: ReferIt3D: Neural listeners for fine-grained 3D object identification in real-world scenes. In: ECCV (2020) 8
2. An, X., Xie, Y., Yang, K., Zhang, W., Zhao, X., Cheng, Z., Wang, Y., Xu, S., Chen, C., Wu, C., Tan, H., Li, C., Yang, J., Yu, J., Wang, X., Qin, B., Wang, Y., Yan, Z., Feng, Z., Liu, Z., Li, B., Deng, J.: LLaVA-OneVision-1.5: Fully open framework for democratized multimodal training. arXiv preprint arXiv:2509.23661 (2025) 9
3. Bar, A., Zhou, G., Tran, D., Darrell, T., LeCun, Y.: Navigation world models. In: CVPR (2025) 2, 4, 5
4. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V., Rombach, R.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023) 5
5. Cen, J., Yu, C., Yuan, H., Jiang, Y., Huang, S., Guo, J., Li, X., Song, Y., Luo, H., Wang, F., Zhao, D., Chen, H.: WorldVLA: Towards autoregressive action world model. arXiv preprint arXiv:2506.21539 (2025) 2, 5
6. Chen, H., Hou, Y., Qu, C., Testini, I., Hong, X., Jiao, J.: 360+x: A panoptic multi-modal scene understanding dataset. In: CVPR (2024) 5
7. Chen, L., Wei, X., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Bin, L., Tang, Z., Yuan, L., Qiao, Y., Lin, D., Zhao, F., Wang, J.: ShareGPT4Video: Improving video understanding and generation with better captions. In: NeurIPS (2024) 5
8. Chen, L., Ma, T., Liu, J., Li, B., Chen, Z., Liu, L., He, X., Li, G., He, Q., Wu, Z.: HuMo: Human-centric video generation via collaborative multi-modal conditioning. arXiv preprint arXiv:2509.08519 (2025) 5
9. Chen, L., Zhou, Z., Zhao, M., Wang, Y., Zhang, G., Huang, W., Sun, H., Wen, J.R., Li, C.: FlexWorld: Progressively expanding 3D scenes for flexible-view exploration. In: NeurIPS (2025) 5
10. Chen, Y., Wang, Y., Zhang, Z.: DrivingGPT: Unifying driving world modeling and planning with multi-modal autoregressive transformers. In: ICCV (2025) 5
11. Cheng, A.C., Fu, Y., Chen, Y., Liu, Z., Li, X., Radhakrishnan, S., Han, S., Lu, Y., Kautz, J., Molchanov, P., Yin, H., Wang, X., Liu, S.: 3D aware region prompted vision language model. arXiv preprint arXiv:2509.13317 (2025) 7
12. Clancey, W.J.: Situated Cognition: On Human Knowledge and Computer Representations. Cambridge University Press, Cambridge, UK (1997) 2
13. Ding, J., Zhang, Y., Shang, Y., Zhang, Y., Zong, Z., Feng, J., Yuan, Y., Su, H., Li, N., Sukiennik, N., Xu, F., Li, Y.: Understanding world or predicting future? A comprehensive survey of world models. ACM Computing Surveys (2025) 5
14. Epstein, R.A., Patai, E.Z., Julian, J.B., Spiers, H.J.: The cognitive map in humans: spatial navigation and beyond. Nature Neuroscience (2017) 6
15. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., Chen, P., Li, Y., Lin, S., Zhao, S., Li, K., Xu, T., Zheng, X., Chen, E., Shan, C., He, R., Sun, X.: Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In: CVPR (2025) 9
16. Gao, S., Zhou, S., Du, Y., Zhang, J., Gan, C.: AdaWorld: Learning adaptable world models with latent actions. In: ICML (2025) 5

17. Gui, D., Guo, X., Zhou, W., Lu, Y.: Image as a world: Generating interactive world from single image via panoramic video generation. In: NeurIPS (2025) 5
18. Ha, D., Schmidhuber, J.: Recurrent world models facilitate policy evolution. In: NeurIPS (2018) 2
19. Hao, Y., Yang, F., Fang, N., Liu, Y.S.: EMBOSR: Embodied spatial reasoning for enhanced situated question answering in 3D scenes. In: IROS (2024) 2
20. Ho, J., Salimans, T., Gritsenko, A.A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. In: NeurIPS (2022) 5
21. Hong, Y., Zhen, H., Chen, P., Zheng, S., Du, Y., Chen, Z., Gan, C.: 3d-llm: Injecting the 3d world into large language models. In: NeurIPS (2023) 7
22. Hu, J., Zhao, S., Chen, Q.G., Qiu, X., Liu, J., Xu, Z., Luo, W., Zhang, K., Lu, Y.: Omni-View: Unlocking how generation facilitates understanding in unified 3D model based on multiview images. arXiv preprint arXiv:2511.07222 (2025) 4
23. Hu, M., Huang, Z., Wang, T., Pang, J., Lin, D., Zheng, N., Xu, R.: ChangingGrounding: 3D visual grounding in changing scenes. arXiv preprint arXiv:2510.14965 (2025) 4
24. Hu, Y., Luo, C., Chen, Z.: Make it move: Controllable image-to-video generation with text descriptions. In: CVPR (2022) 5
25. Hu, Y., Wang, L., Liu, X., Chen, L.H., Guo, Y., Shi, Y., Liu, C., Rao, A., Wang, Z., Xiong, H.: Simulating the real world: A unified survey of multimodal generative models. arXiv preprint arXiv:2503.04641 (2025) 5
26. Huang, C., Yan, S., Burgard, W.: BYE: Build your encoder with one sequence of exploration data for long-term dynamic scene understanding. IEEE Robotics and Automation Letters (2025) 2
27. Huang, J., Yong, S., Ma, X., Linghu, X., Li, P., Wang, Y., Li, Q., Zhu, S.C., Jia, B., Huang, S.: An embodied generalist agent in 3D world. In: ICML (2024) 4
28. Huang, J., Li, Z., Zhang, H., Chen, R., He, X., Guo, Y., Wang, W., Liu, T., Gong, M.: SURPRISE3D: A dataset for spatial understanding and reasoning in complex 3D scenes. arXiv preprint arXiv:2507.07781 (2025) 4
29. Jia, B., Chen, Y., Yu, H., Wang, Y., Niu, X., Liu, T., Li, Q., Huang, S.: SceneVerse: Scaling 3D vision-language learning for grounded scene understanding. In: ECCV (2024) 3
30. Jia, M., Qi, Z., Zhang, S., Zhang, W., Yu, X., He, J., Wang, H., Yi, L.: OmniSpatial: Towards comprehensive spatial reasoning benchmark for vision language models. arXiv preprint arXiv:2506.03135 (2025) 6
31. Jin, Q., Wu, Y., Chen, C.: PanoNav: Mapless zero-shot object navigation with panoramic scene parsing and dynamic memory. In: AAAI (2026) 4
32. Koh, J.Y., Lee, H., Yang, Y., Baldridge, J., Anderson, P.: Pathdreamer: A world model for indoor navigation. In: ICCV (2021) 2, 5
33. Köksal, A., Ak, K.E., Sun, Y., Rajan, D., Lim, J.H.: Controllable video generation with text-based instructions. IEEE Transactions on Multimedia (2024) 5
34. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., Wu, K., Lin, Q., Yuan, J., Long, Y., Wang, A., Wang, A., Li, C., Huang, D., Yang, F., Tan, H., Wang, H., Song, J., Bai, J., Wu, J., Xue, J., Wang, J., Wang, K., Liu, M., Li, P., Li, S., Wang, W., Yu, W., Deng, X., Li, Y., Chen, Y., Cui, Y., Peng, Y., Yu, Z., He, Z., Xu, Z., Zhou, Z., Xu, Z., Tao, Y., Lu, Q., Liu, S., Zhou, D., Wang, H., Yang, Y., Wang, D., Liu, Y., Jiang, J., Zhong, C.: HunyuanVideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024) 8, 9

35. Kosslyn, S.M., Ball, T.M., Reiser, B.J.: Visual images preserve metric spatial information: Evidence from studies of image scanning. Journal of Experimental Psychology: Human Perception and Performance **4**(1), 47–60 (1978). https://doi.org/10.1037/0096-1523.4.1.47 14
36. Krishnaswamy, N., Pustejovsky, J.: Grounding meaning representation for situated reasoning. In: Alonso, M.A., Wei, Z. (eds.) Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Tutorial Abstracts. pp. 22–27. Association for Computational Linguistics, Taipei (Nov 2022). https://doi.org/10.18653/v1/2022.aacl-tutorials.4, https://aclanthology.org/2022.aacl-tutorials.4/ 2
37. Labs, P.: Pika: Ideas to video. https://pika.art (2024), accessed: 2024 9
38. Labs, W.: Marble: A multimodal 3D world model. https://www.worldlabs.ai (2025), accessed: 2025-11-12 5
39. LeCun, Y.: A path towards autonomous machine intelligence. OpenReview (2022), https://openreview.net/pdf?id=BZ5a1r-kVsf, accessed: 2025-11-12 2
40. Li, D., Jiang, B., Huang, L., Beigi, A., Zhao, C., Tan, Z., Bhattacharjee, A., Jiang, Y., Chen, C., Wu, T., Shu, K., Cheng, L., Liu, H.: From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In: EMNLP (2025) 9
41. Li, H., Li, D., Wang, Z., Yan, Y., Wu, H., Zhang, W., Shen, Y., Lu, W., Xiao, J., Zhuang, Y.: SpatialLadder: Progressive training for spatial reasoning in vision-language models. arXiv preprint arXiv:2510.08531 (2025) 4
42. Li, S., Yang, C., Fang, J., Yi, T., Lu, J., Cen, J., Xie, L., Shen, W., Tian, Q.: WorldGrow: Generating infinite 3D world. arXiv preprint arXiv:2510.21682 (2025) 5
43. Li, Y., Torralba, A.: Multimodal action conditioned video simulation. In: ICCV (2025) 5
44. Liang, J., Chen, R., Jiao, X., Liang, S., Liu, S., Zhang, Q., Hu, Z., Cao, X.: Explaining multimodal LLMs via intra-modal token interactions. arXiv preprint arXiv:2509.22415 (2025) 11
45. Linghu, X., Huang, J., Niu, X., Ma, X.S., Jia, B., Huang, S.: Multi-modal situated reasoning in 3D scenes. In: NeurIPS (2024) 2, 4, 6, 7, 10
46. Liu, J., Wang, Y., Ma, S., Li, B.: Analysis of stairs-climbing ability for a tracked reconfigurable modular robot. In: SSRR (2005) 2
47. Liu, J., Lin, S., Li, Y., Yang, M.H.: DynamicScaler: Seamless and scalable video generation for panoramic scenes. In: CVPR (2025) 5
48. Liu, Q., Huang, T., Zhang, Z., Tang, H.: Nav-R1: Reasoning and navigation in embodied scenes. arXiv preprint arXiv:2509.10884 (2025) 6
49. Liu, R., Zhang, J., Peng, K., Zheng, J., Cao, K., Chen, Y., Yang, K., Stiefelhagen, R.: Open scene understanding: Grounded situation recognition meets segment anything for helping people with visual impairments. In: CVPRW (2023) 3
50. Liu, R., Zhang, J., Schön, A., Müller, K., Zheng, J., Yang, K., Guo, A., Gerling, K., Stiefelhagen, R.: ObjectFinder: An open-vocabulary assistive system for interactive object search by blind people. arXiv preprint arXiv:2412.03118 (2024) 2
51. Liu, R., Zhang, J., Zheng, J., Chen, Y., Lee, P.S., Wen, D., Peng, K., Zhang, J., Yang, K., Mombaur, K., et al.: Not an obstacle for dog, but a hazard for human: A co-ego navigation system for guide dog robots. arXiv preprint arXiv:2603.20121 (2026) 2

52. Liu, R., Zheng, J., Chen, Y., Wang, Z., Peng, K., Yang, K., Zhang, J., Pollefeys, M., Stiefelhagen, R.: Situat3DChange: Situated 3D change understanding dataset for multimodal large language model. In: NeurIPS (2025) 2, 4, 6, 10
53. Liu, R., Zheng, J., Chen, Y., Wen, D., Quan, S., Wu, C., Zhang, J., Yang, K., Peng, K., Stiefelhagen, R.: Egoexomem: Cross-view memory reasoning over synchronized egocentric and exocentric videos. arXiv preprint arXiv:2605.18734 (2026) 4
54. Lu, T., Shu, T., Yuille, A., Khashabi, D., Chen, J.: Generative world explorer. In: ICLR (2025) 2, 5, 6
55. Ma, X., Yong, S., Zheng, Z., Li, Q., Liang, Y., Zhu, S.C., Huang, S.: SQA3D: Situated question answering in 3D scenes. In: ICLR (2023) 2, 4
56. Majumdar, A., Ajay, A., Zhang, X., Putta, P., Yenamandra, S., Henaff, M., Silwal, S., Mcvay, P., Maksymets, O., Arnaud, S., Yadav, K., Li, Q., Newman, B., Sharma, M., Berges, V., Zhang, S., Agrawal, P., Bisk, Y., Batra, D., Kalakrishnan, M., Meier, F., Paxton, C., Sax, A., Rajeswaran, A.: OpenEQA: Embodied question answering in the era of foundation models. In: CVPR (2024) 4, 6
57. Man, Y., Gui, L.Y., Wang, Y.X.: Situational awareness matters in 3D vision language reasoning. In: CVPR (2024) 4
58. Moulton, S.T., Kosslyn, S.M.: Imagining predictions: Mental imagery as mental emulation. In: Predictions in the Brain: Using Our Past to Generate a Future. Oxford University Press (2011) 2, 5
59. Mullins, C.D.: Cognitive Behavioral Group Therapy for Blind and Visually Impaired Adults: Acceptance, Problem-Solving, and Cognitive Distortions. Ph.D. thesis, Philadelphia College of Osteopathic Medicine (2019) 2
60. van Munster, E.P.J., van der Aa, H.P.A., Verstraten, P., van Rens, G.H.M.B., van Nispen, R.M.A.: Barriers and facilitators to recognize and discuss depression and anxiety experienced by adults with vision impairment or blindness: a qualitative study. BMC Health Services Research (2021) 2
61. Nie, D., Guo, X., Duan, Y., Zhang, R., Chen, L.: WMNav: Integrating vision-language models into world models for object goal navigation. In: IROS (2025) 2
62. OpenAI: Sora: Creating video from text. Tech. rep., OpenAI (2024), https://openai.com/sora, accessed: 2025-11-12 9
63. OpenAI: Function calling (Aug 2025), https://developers.openai.com/api/docs/guides/function-calling/, openAI API documentation, accessed: 2026-03-02 14
64. OpenAI: Gpt-5 technical overview. https://openai.com/index/gpt-5 (2025), accessed: 2025-11-12 7
65. Parker-Holder, J., Fruchter, S.: Genie 3: A new frontier for world models. https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models (2025), accessed: 2025-11-12 5
66. Qin, Y., Shi, Z., Yu, J., Wang, X., Zhou, E., Li, L., Yin, Z., Liu, X., Sheng, L., Shao, J., Bai, L., Ouyang, W., Zhang, R.: WorldSimBench: Towards video generation models as world simulators. arXiv preprint arXiv:2410.18072 (2024) 4
67. Ramakrishnan, S.K., Gokaslan, A., Wijmans, E., Maksymets, O., Clegg, A., Turner, J., Undersander, E., Galuba, W., Westbury, A., Chang, A.X., Savva, M., Zhao, Y., Batra, D.: Habitat-matterport 3D dataset (HM3D): 1000 large-scale 3D environments for embodied AI. In: NeurIPS (2021) 6
68. Ray, A., Duan, J., Brown, E., Tan, R., Bashkirova, D., Hendrix, R., Ehsani, K., Kembhavi, A., Plummer, B.A., Krishna, R., Zeng, K.H., Saenko, K.: SAT:

Dynamic spatial aptitude training for multimodal language models. In: COLM (2025) 4, 6, 8, 14
69. Richens, J., Everitt, T.: Robust agents learn causal world models. In: ICLR (2024) 5
70. inVia Robotics, I.: Overcoming the challenges of robotics in the warehouse. https://inviarobotics.com/blog/overcoming-challenges-robotics-warehouse (2016), accessed: 2025-11-12 2
71. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., Parikh, D., Batra, D.: Habitat: A Platform for Embodied AI Research. In: ICCV (2019) 6
72. Seo, T., Ryu, S., Won, J.H., Kim, Y., Kim, H.S.: Stair-climbing robots: A review on mechanism, sensing, and performance evaluation. IEEE Access (2023) 2
73. Song, I., Lee, S., Joo, M., Lee, J.: Anomaly detection for people with visual impairments using an egocentric 360-degree camera. In: WACV (2025) 5
74. Sugandhika, C., Li, C., Rajan, D., Fernando, B.: Situational scene graph for structured human-centric situation understanding. In: WACV (2025) 4
75. Tan, Z., Yang, H., Qin, L., Gong, J., Yang, M., Li, H.: Omni-Video: Democratizing unified video understanding and generation. arXiv preprint arXiv:2507.06119 (2025) 5
76. Team, H., Wang, Z., Liu, Y., Wu, J., Gu, Z., Wang, H., Zuo, X., Huang, T., Li, W., Zhang, S., Lian, Y., Tsai, Y., Wang, L., Liu, S., Jiang, P., Yang, X., Guo, D., Tang, Y., Mao, X., Yu, J., Yu, J., Zhang, J., Chen, M., Dong, L., Jia, Y., Zhang, C., Tan, Y., Zhang, H., Ye, Z., He, P., Wu, R., Chen, M., Li, Z., Qin, W., Wang, L., Sun, Y., Niu, L., Yuan, X., Yang, X., He, Y., Xiao, J., Tao, Y., Zhu, J., Xue, J., Liu, K., Zhao, C., Wu, X., Liu, T., Chen, P., Wang, D., Liu, Y., Linus, Jiang, J., Wang, T., Guo, C.: HunyuanWorld 1.0: Generating immersive, explorable, and interactive 3D worlds from words or pixels. arXiv preprint arXiv:2507.21809 (2025) 5
77. Veerabadran, V., Xiao, F., Kamra, N., Matias, P., Chen, J., Drooff, C., Roads, B.D., Williams, R., Henderson, E., Zhao, X., Carlberg, K., Tighe, J., Ridgeway, K.: Benchmarking egocentric multimodal goal inference for assistive wearable agents. arXiv preprint arXiv:2510.22443 (2025) 5
78. Wang, A., Wu, B., Chen, S., Chen, Z., Guan, H., Lee, W.N., Li, L.E., Gan, C.: SOK-Bench: A situated video reasoning benchmark with aligned open-world knowledge. In: CVPR (2024) 4
79. Wang, A., Ai, B., Wen, B., Mao, C., Xie, C., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Meng, X., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z., Liu, Z.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025) 8, 9
80. Wang, H., Liang, W., Van Gool, L., Wang, W.: Dreamwalker: Mental planning for continuous vision-language navigation. In: ICCV (2023) 2
81. Wang, J., Yuan, Y., Zheng, R., Lin, Y., Gao, J., Chen, L., Bao, Y., Zhang, Y., Zeng, C., Zhou, Y., Long, X., Zhu, H., Zhang, Z., Cao, X., Yao, Y.: SpatialVID: A large-scale video dataset with spatial annotations. arXiv preprint arXiv:2509.09676 (2025) 5

82. Wang, Q., Li, W., Mou, C., Cheng, X., Zhang, J.: 360DVD: Controllable panorama video generation with 360-degree video diffusion model. In: CVPR (2024) 5
83. Wang, S., Zhou, D., Xie, L., Xu, C., Yan, Y., Yin, E.: PanoGen++: Domain-adapted text-guided panoramic environment generation for vision-and-language navigation. Neural Networks (2025) 5
84. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024) 14
85. Wang, X., Yang, X., Xu, Y., Wu, Y., Li, Z., Zhao, N.: AffordBot: 3D fine-grained embodied reasoning via multimodal large language models. In: NeurIPS (2025) 2, 4
86. Wang, Y., Jia, B., Zhu, Z., Huang, S.: Masked point-entity contrast for open-vocabulary 3D scene understanding. In: CVPR (2025) 3
87. Wei, J., Zheng, J., Liu, R., Hu, J., Zhang, J., Stiefelhagen, R.: OneBEV: Using one panoramic image for bird's-eye-view semantic mapping. In: ACCV (2024) 5
88. Wei, J., Yuan, S., Li, P., Hu, Q., Gan, Z., Ding, W.: OccLLaMA: An occupancy-language-action generative world model for autonomous driving. arXiv preprint arXiv:2409.03272 (2024) 5
89. Weiss, M., Chamorro, S., Girgis, R., Luck, M., Kahou, S.E., Cohen, J.P., Nowrouzezahrai, D., Precup, D., Golemo, F., Pal, C.: Navigation agents for the visually impaired: A sidewalk simulator and experiments. In: CoRL (2020) 5
90. Wen, Y., Zhao, Y., Liu, Y., Jia, F., Wang, Y., Luo, C., Zhang, C., Wang, T., Sun, X., Zhang, X.: Panacea: Panoramic and controllable video generation for autonomous driving. In: CVPR (2024) 5
91. Wu, B., Yu, S., Chen, Z., Tenenbaum, J.B., Gan, C.: STAR: A benchmark for situated reasoning in real-world videos. In: NeurIPS (2021) 4
92. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: ICCV (2023) 5
93. Wu, S., Teng, F., Shi, H., Jiang, Q., Luo, K., Wang, K., Yang, K.: QuaDreamer: Controllable panoramic video generation for quadruped robots. In: CoRL (2025) 5
94. Xia, Y., Weng, S., Yang, S., Liu, J., Zhu, C., Teng, M., Jia, Z., Jiang, H., Shi, B.: PanoWan: Lifting diffusion video generation models to 360° with latitude/longitude-aware mechanisms. In: NeurIPS (2025) 5
95. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. In: ICLR (2025) 5, 14
96. Xie, J., Yang, Z., Shou, M.Z.: Show-o2: Improved native unified multimodal models. arXiv preprint arXiv:2506.15564 (2025) 14
97. Xing, J., Xia, M., Zhang, Y., Chen, H., Yu, W., Liu, H., Liu, G., Wang, X., Shan, Y., Wong, T.T.: DynamiCrafter: Animating open-domain images with video diffusion priors. In: ECCV (2024) 5
98. Xiong, H., Zhuge, Y., Zhu, J., Zhang, L., Lu, H.: 3UR-LLM: An end-to-end multimodal large language model for 3D scene understanding. IEEE Transactions on Multimedia (2025) 3
99. Xiu, J., Hong, F., Li, Y., Li, M., Wang, W., Han, S., Pan, L., Liu, Z.: EgoTwin: Dreaming body and view in first person. arXiv preprint arXiv:2508.13013 (2025) 5

100. Xu, J., Xu, S., Ma, M., Ma, J., Li, C.: Research and implementation of travel aids for blind and visually impaired people. Sensors (2025) 2
101. Yan, Z., Li, S., Wang, Z., Wu, L., Wang, H., Zhu, J., Chen, L., Liu, J.: Dynamic open-vocabulary 3D scene graphs for long-term language-guided mobile manipulation. IEEE Robotics and Automation Letters (2025) 2
102. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025) 9
103. Yang, J., Zhang, H., Li, F., Zou, X., Li, C., Gao, J.: Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V. arXiv preprint arXiv:2310.11441 (2023) 7
104. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces. In: CVPR (2025) 4
105. Yang, L., Duan, H., Tao, R., Cheng, J., Wu, S., Li, Y., Liu, J., Min, X., Zhai, G.: ODI-Bench: Can MLLMs understand immersive omnidirectional environments? arXiv preprint arXiv:2510.11549 (2025) 4
106. Yang, S., Yang, J., Huang, P., Brown, E., Yang, Z., Yu, Y., Tong, S., Zheng, Z., Xu, Y., Wang, M., Lu, D., Fergus, R., LeCun, Y., Fei-Fei, L., Xie, S.: Cambrian-S: Towards spatial supersensing in video. arXiv preprint arXiv:2511.04670 (2025) 4
107. Yang, S., Xu, R., Xie, Y., Yang, S., Li, M., Lin, J., Zhu, C., Chen, X., Duan, H., Yue, X., Lin, D., Wang, T., Pang, J.: MMSI-Bench: A benchmark for multi-image spatial intelligence. arXiv preprint arXiv:2505.23764 (2025) 4
108. Yang, Y., Liu, J., Zhang, Z., Zhou, S., Tan, R., Yang, J., Du, Y., Gan, C.: MindJourney: Test-time scaling with world models for spatial reasoning. In: NeurIPS (2025) 2, 9, 13, 14
109. Yang, Y., Yang, H., Zhou, J., Chen, P., Zhang, H., Du, Y., Gan, C.: 3D-Mem: 3D scene memory for embodied exploration and reasoning. In: CVPR (2025) 2, 4
110. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., Yin, D., Zhang, Y., Wang, W., Cheng, Y., Xu, B., Gu, X., Dong, Y., Tang, J.: CogVideoX: Text-to-video diffusion models with an expert transformer. In: ICLR (2025) 9
111. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: ScanNet++: A high-fidelity dataset of 3D indoor scenes. In: ICCV (2023) 6
112. Ying, K., Liu, R., Chen, C., Tao, M., Shi, H., Yang, K., Zhang, J., Stiefelhagen, R.: mmWalk: Towards multi-modal multi-view walking assistance. In: NeurIPS (2025) 7, 10
113. Yuan, Z., Jiang, S., Feng, C.M., Zhang, Y., Cui, S., Li, Z., Zhao, N.: SceneR1: Video-grounded large language models for 3D scene reasoning without 3D annotations. arXiv preprint arXiv:2506.17545 (2025) 4
114. Yun, H., Yu, Y., Yang, W., Lee, K., Kim, G.: Pano-AVQA: Grounded audio-visual question answering on 360° videos. In: ICCV (2021) 4
115. Zhang, W., Wang, M., Liu, G., Xu, H., Jiang, Y., Shen, Y., Hou, G., Zheng, Z., Zhang, H., Li, X., Lu, W., Li, P., Zhuang, Y.: Embodied-reasoner: Synergizing visual search, reasoning, and action for embodied interactive tasks. arXiv preprint arXiv:2503.21696 (2025) 2

116. Zhang, X., Fu, T., Zheng, X.: Omnidirectional spatial modeling from correlated panoramas. In: MM (2025) 4
117. Zhang, X., Ye, Z., Zheng, X.: Towards omnidirectional reasoning with 360-R1: A dataset, benchmark, and GRPO-based method. arXiv preprint arXiv:2505.14197 (2025) 5
118. Zhang, Y., Xu, Z., Shen, Y., Kordjamshidi, P., Huang, L.: SPARTUN3D: Situated spatial understanding of 3D world in large language models. In: ICLR (2025) 2, 4
119. Zhang, Z., Wang, Z., Zhang, G., Dai, W., Xia, Y., Yan, Z., Hong, M., Zhao, Z.: DSI-Bench: A benchmark for dynamic spatial intelligence. arXiv preprint arXiv:2510.18873 (2025) 4
120. Zhao, J., Hou, R., Tian, Z., Chang, H., Shan, S.: HIS-GPT: Towards 3D human-in-scene multimodal understanding. In: ICCV (2025) 4
121. Zhen, H., Qiu, X., Chen, P., Yang, J., Yan, X., Du, Y., Hong, Y., Gan, C.: 3D-VLA: A 3D vision-language-action generative world model. arXiv preprint arXiv:2403.09631 (2024) 2, 5
122. Zheng, D., Huang, Z., Liu, H., Zou, K., He, Y., Zhang, F., Zhang, Y., He, J., Zheng, W., Qiao, Y., Liu, Z.: VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. arXiv preprint arXiv:2503.21755 (2025) 9
123. Zheng, D., Huang, S., Wang, L.: Video-3D LLM: Learning position-aware video representation for 3D scene understanding. In: CVPR (2025) 4
124. Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., Feng, J.: MagicVideo: Efficient video generation with latent diffusion models. arXiv preprint arXiv:2211.11018 (2022) 5
125. Zhou, H., Cheng, X., Yu, W., Tian, Y., Yuan, L.: HoloDreamer: Holistic 3D panoramic world generation from text descriptions. arXiv preprint arXiv:2407.15187 (2024) 5
126. Zhou, J., Gao, H., Voleti, V., Vasishta, A., Yao, C.H., Boss, M., Torr, P., Rupprecht, C., Jampani, V.: Stable virtual camera: Generative view synthesis with diffusion models. arXiv preprint arXiv:2503.14489 (2025) 9
127. Zhou, S., Du, Y., Chen, J., Li, Y., Yeung, D.Y., Gan, C.: RoboDreamer: Learning compositional world models for robot imagination. In: ICML (2024) 2
128. Zhu, C., Wang, T., Zhang, W., Chen, K., Liu, X.: ScanReason: Empowering 3D visual grounding with reasoning capabilities. In: ECCV (2024) 3
129. Zhu, Z., Wang, X., Li, Y., Zhang, Z., Ma, X., Chen, Y., Jia, B., Liang, W., Yu, Q., Deng, Z., Huang, S., Li, Q.: Move to understand a 3D scene: Bridging visual grounding and exploration for efficient and versatile embodied navigation. In: ICCV (2025) 4