
Supplementary Materials of Test-Time Scaling with World Models for Spatial Reasoning

Anonymous Author(s)

Affiliation

Address

email

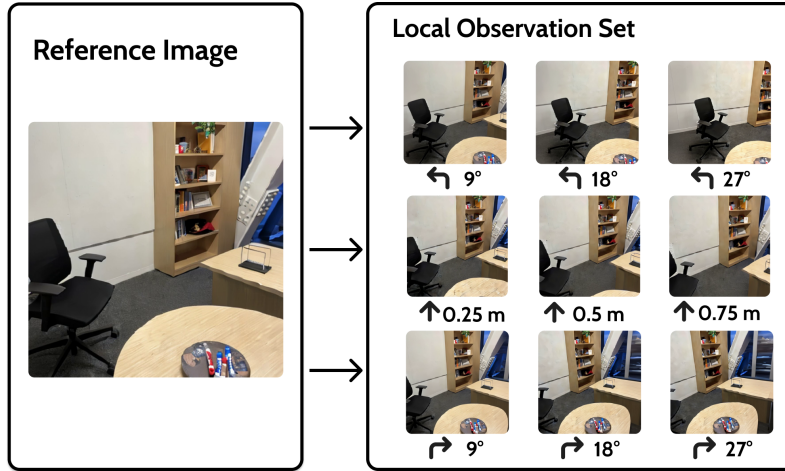


Figure 1: Trajectory Expansion example on SAT-Real.

1 The supplementary material is organized as follows:

- 2 • Sec. **A** details all experimental settings, including inference hyper-parameters and the
3 training recipe for our world model.
- 4 • Sec. **B** examines typical failure modes, distinguishing errors that arise in the world model
5 from those attributable to the VLM.
- 6 • Sec. **C** provides additional ablation studies, covering alternative world models and the role
7 of natural-language trajectory descriptions.
- 8 • Sec. **D** lists every prompt used in our pipeline to facilitate exact reproducibility.
- 9 • Sec. **E** discusses the broader societal impacts of this work.

10 To aid visual understanding, we supply .mp4 clips that juxtapose our simulated camera motions with
11 ground-truth trajectories in AI2-THOR, as well as .pdf files containing the final question-and-answer
12 prompts augmented with the evidence buffer obtained after search.

13 A Experiment Details

14 A.1 Inference Details

15 A.1.1 Search Configurations

16 We use the same search configuration for every experiment: a search depth of $n = 3$ steps, a beam
 17 size of 2, exploration and helpfulness thresholds $\gamma_{\text{exp}} = 8$ and $\gamma_{\text{help}} = 8$, and a maximum trajectory
 18 length of 8 starting from the given reference image. During each expansion, we allow up to $k = 3$
 19 consecutive repetitions per primitive action; each forward step moves the agent 0.25m and each
 20 rotation step turns it by 9° . We prepared more visual examples about the Trajectory Expansion
 21 process under our experiment settings at Fig. 1 and Fig. 2

22 A.1.2 Computational Resources

23 All inference experiments were run on high-performance NVIDIA GPUs: when using Search
 24 World Model as the world model, we employed A40 GPUs with 40GB of VRAM; when using
 25 Stable-Virtual-Camera as the world model, we ran on H100 GPUs with 80GB of VRAM; and for all
 26 experiments combining the InternVL3-14B VLM with the Search World Model, we also used H100
 27 GPUs to accommodate the larger memory footprint of the vision-language model.

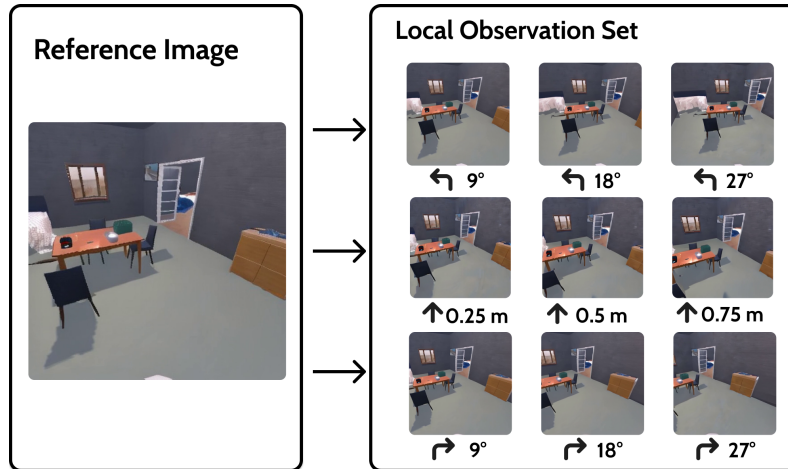


Figure 2: Trajectory Expansion example on SAT-Synthesized.

28 A.2 Search World Model Training

29 A.2.1 Dataset

30 The training set for our Search World Model (SWM) comprises three components: HM3D,
 31 DL3DV-10K, and RealEstate10K [Szot et al., 2022, Ling et al., 2024, Zhou et al., 2018].

32 **HM3D.** We generate 50K simulated navigation clips using Habitat on HM3D scenes. For each
 33 episode, we sample a random start and goal position and follow the shortest path for up to 500 steps.
 34 To improve robustness to camera tilt, we uniformly draw an initial pitch from $[-30^\circ, 20^\circ]$, and in
 35 10% of episodes we hold the agent at a fixed location and vary only its pitch by rotating up and down.

36 **RealEstate10K.** This collection of 10K real indoor videos enriches our data with diverse residential
 37 environments, balancing the simulated HM3D distribution.

38 **DL3DV-10K.** Similarly, we incorporate 10K real outdoor videos from DL3DV-10K to capture natural
 39 scenes and broaden our model’s generalization to exterior settings.

40 To ensure consistent camera dynamics across all sources, we normalize each clip’s frame rate and
 41 spatial resolution before training.

42 A.2.2 Implementation Details

43 **Model Architecture.** We adopt CogVideoX-5B [Yang et al., 2024]¹ as our backbone. To incorporate
 44 camera embeddings, we expand the input channels of its PatchEmbed layer—duplicating the pre-
 45 trained convolutional weights into the appropriate submatrix and zero-initializing all new parameters.
 46 Furthermore, we introduce LoRA adapters within each self-attention block, configured with rank
 47 $r = 128$ and scaling factor $\alpha = 128$, enabling parameter-efficient fine-tuning without disturbing the
 48 original transformer weights.

49 **Action Representation.** We represent each relative camera transform using the Plücker embed-
 50 ding [Sitzmann et al., 2021]. Given intrinsic matrix \mathbf{K} and extrinsic matrix $\mathbf{E} = [\mathbf{R} \mid \mathbf{t}]$, we build a
 51 6-channel “image”

$$P \in \mathbb{R}^{h \times w \times 6},$$

52 where at pixel (u, v) we store the pair

$$(o_{(u,v)} \times d_{(u,v)}, d_{(u,v)}).$$

53 Here

$$o_{(u,v)} = \mathbf{t}, \quad d_{(u,v)} = \mathbf{R} \mathbf{K}^{-1} \begin{pmatrix} u \\ v \\ 1 \end{pmatrix}.$$

54 This encoding lets the video model reason directly over 3D ray origins and directions for precise
 55 spatial control.

56 **Decoupling Pitch.** Our primitive actions {move forward, turn right, turn left} always pre-
 57 serve the camera’s pitch angle θ . To enforce this, we factor each extrinsic transform $\mathbf{E} \in \text{SE}(3)$ into
 58 a pure pitch rotation and a remaining horizontal motion:

$$\mathbf{E} = R_{\text{pitch}}(\theta) T_{\text{horiz}},$$

59 where $R_{\text{pitch}}(\theta)$ is a rotation about the camera’s local x -axis:

$$R_{\text{pitch}}(\theta) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos \theta & -\sin \theta & 0 \\ 0 & \sin \theta & \cos \theta & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

60 and the horizontal motion component is

$$T_{\text{horiz}} = R_{\text{pitch}}(-\theta) \mathbf{E}.$$

61 We then apply the Plücker embedding only to T_{horiz} , producing a tensor in $\mathbb{R}^{h \times w \times 6}$. Finally, we
 62 broadcast the scalar pitch θ into an $(h \times w \times 1)$ map and concatenate it along the channel dimension
 63 to obtain the full $\mathbb{R}^{h \times w \times 7}$ action representation.

64 A.2.3 Training Details

65 We subsample training clips with a frame-skip stride uniformly drawn between 1 and 3 to expose the
 66 model to varied camera motions. Optimization is performed with Adam and a linear warmup schedule
 67 to a peak learning rate of $1e - 4$, using bfloat16 precision for efficiency and clipping gradients to a
 68 maximum norm of 1.0 for stability. All video-diffusion models are trained on eight NVIDIA H100
 69 GPUs over approximately three days. At training time we adopt the v-prediction objective [Salimans
 70 and Ho, 2022] and, during inference, employ the DDIM sampler [Rombach et al., 2022] with 50
 71 sampling steps—producing nine-frame videos in roughly nine seconds.

72 B Failure Case Analysis

73 Despite the strong overall performance of our model, we conducted a detailed analysis of the failure
 74 cases to uncover its limitations. The following examples highlight typical scenarios where the model
 75 does not perform as expected.

76 Unlike the previous figures, the action labels in Fig. 3 represent delta action at each step.

¹<https://github.com/THUDM/CogVideo>; Apache-2.0

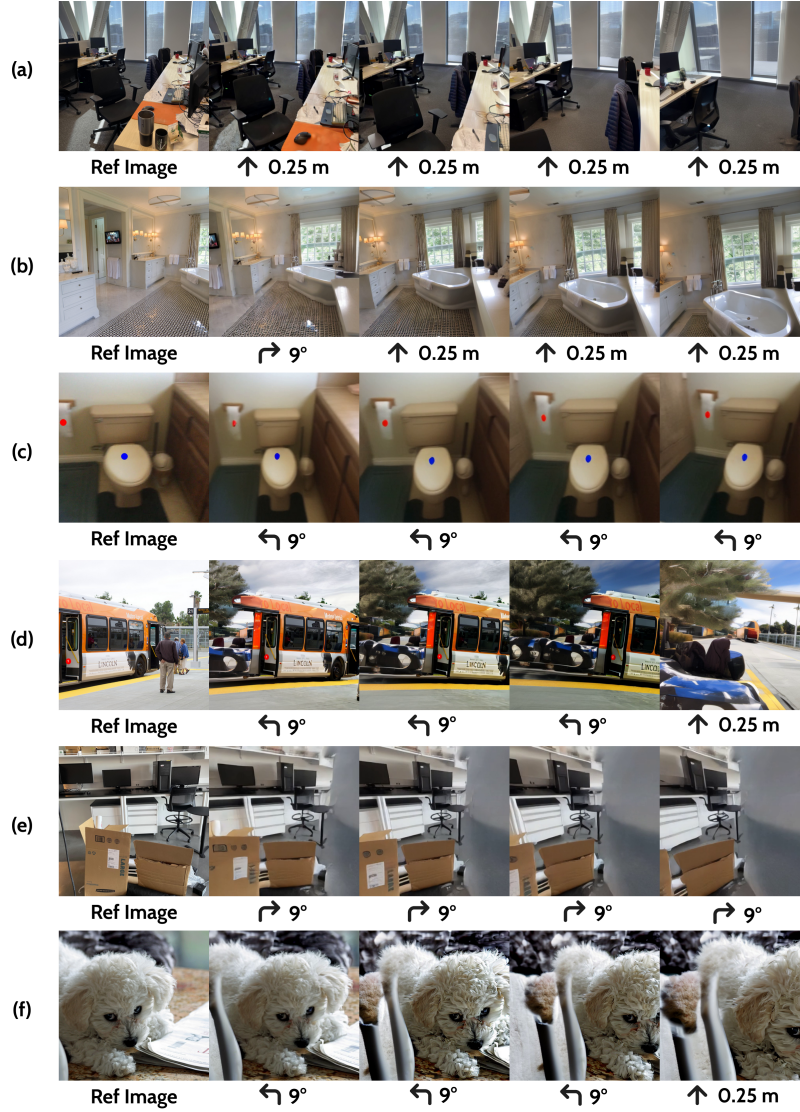


Figure 3: **Failure Cases of World Models.** Group (a) shows inaccurate forward movement; group (b) shows unintended roll movement leading to a tilted scene; group (c) shows the unstable egocentric rotation that introduces viewpoint movement; group (d) shows the model generate artifacts for unseen regions; group (e) shows misinterpretation when inference on real-world scene; group (f) shows a failure case on out-of-domain animal data.

77 **B.1 World Model Capabilities**

78 **B.1.1 Case 1: Inaccurate Forward Movement**

79 In group (a), the imagined trajectories systematically under- or over-estimate forward translations: the
80 actual step lengths no longer match the intended distances, and the displacement between successive
81 frames varies unpredictably. As a result, the agent repeatedly overshoots its targets and exhibits
82 jittery, erratic motion within the simulated environment.

83 We hypothesize that these errors arise from scale inconsistencies across our training sources. In the
84 Search World Model, only HM3D provides metrically accurate movement distances, whereas the
85 Stable Virtual Camera relies on datasets with more heterogeneous scale calibrations. When these
86 conflicting scale conventions are combined during training, the model learns incompatible motion
87 priors—manifesting exactly as the misaligned, erratic forward movements seen in group (a).

88 **B.1.2 Case 2: Unintended Roll Movement**

89 In group (b), the predicted images exhibit an unnatural tilt of the scene, where the horizon line
90 is significantly misaligned. This indicates that the model sometimes introduces unintended roll
91 movements, resulting in a distorted camera orientation.

92 **B.1.3 Case 3: Unstable Egocentric Rotation**

93 In group (c), the predicted images exhibit unstable viewpoints during egocentric rotation. The
94 transitions between consecutive frames are inconsistent, and the visual perspective appears to undergo
95 a rightward translation while simultaneously rotating.

96 This issue happens more often for our world model SWM, which happens because we blend some
97 RealEstate10K data when fine-tuning SWM. RealEstate10K contains numerous segments in which
98 the camera trajectory exhibits simultaneous translation motion and rotation, leading to a distributional
99 bias in training and causing the model to develop systematic prediction errors.

100 **B.1.4 Case 4: Visual Artifacts**

101 In group (d), the predicted images contain noticeable visual artifacts, particularly in regions that are
102 occluded or unseen in the input view. These artifacts manifest as texture distortions, unnatural edges,
103 or inconsistent object boundaries, which significantly degrade the visual realism of the generated
104 images.

105 This issue may stem from the model’s limited ability to hallucinate plausible content in areas with
106 insufficient visual context or out of domain data. In particular, when the target view includes regions
107 not visible in the source image, the model may rely on weak priors or overfit to spurious patterns
108 seen during training.

109 **B.1.5 Case 5: Out of domain data - scene misinterpretation**

110 In group (e), the model exhibits clear failures when processing scenes that fall outside the distribution
111 of the training data. The predicted images demonstrate significant misinterpretation of scene structure,
112 such as incorrect boundary extension as shown in the example. These failures are especially prominent
113 in complex real-world environments with lighting, textures, or layouts not observed during training.

114 We attribute this behavior to the model’s limited generalization ability when confronted with out-
115 of-distribution inputs. Without adequate exposure to diverse scene types during training, the model
116 tends to rely on learned priors that do not transfer well, resulting in hallucinated or semantically
117 inconsistent content.

118 **B.1.6 Case 6: Out of domain data - human or animal**

119 In group (f), as shown in the images, the body of the dog is missing. The model fails to generate
120 plausible predictions when encountering humans or animals, which are underrepresented or absent
121 in the training data. The generated images often exhibit severe distortions in body shape or texture
122 consistency, making the predictions semantically incorrect.

123 This failure can be attributed to the model’s lack of exposure to articulated and deformable entities
124 during training. Humans and animals involve complex structures and dynamic poses that require
125 specialized representation and learning. Without sufficient domain-specific data, the model struggles
126 to generalize, leading to implausible reconstructions or complete semantic failures.

Question:
If I turn left by 16 degrees, will I be facing away from GarbageCan (near the mark 3 in the image)?

Answer Choices:
['yes', 'no']

Correct Answer: yes
LLM Response: no (wrong)

These are the images that pair with the question.
Image 1:


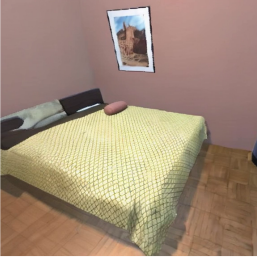


Image 1 is your current egocentric view

Below are the imagined views you would obtain if you took the corresponding actions. These are provided to help you answer the question.

Action: turn left
turn left 18 degrees



turn left 27 degrees




Figure 4: Failure case - VLM’s Q&A ability is not sufficient.

127 **B.2 VLM Capabilities**

128 **B.2.1 Case 1: VLM Q&A**

129 As illustrated in Fig. 4, although the world model generates great visualizations that would intuitively
130 help the VLM with spatial reasoning, the VLM can still be confused and cannot answer the question
131 correctly. Therefore, for spatial reasoning question, the question answering ability of the base VLM
132 is still very important.

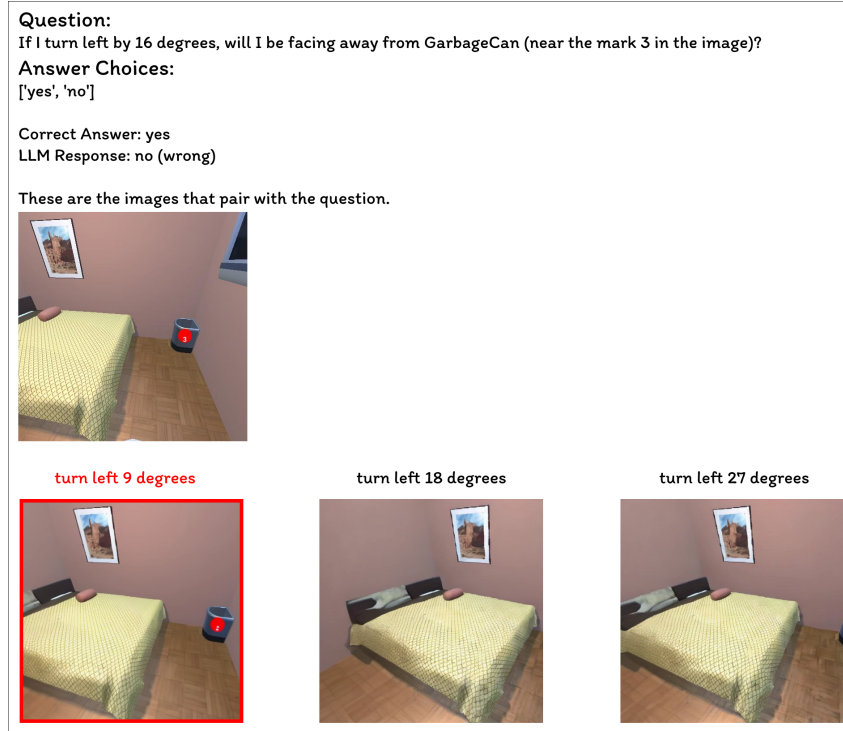


Figure 5: Failure case - VLM’s scoring ability is not sufficient.

B.2.2 Case 2: VLM Scoring

Given the same question mentioned in Case 1, the VLM is not able to keep one of the most informative image after the scoring process. As illustrated in Fig. 5, the "turn left 9 degrees" is the most informative image as it contains the garbage can mentioned in the question. However, the VLM scoring process does not keep the image in the final evidence buffer, which leads to a wrong answer. The improvement of VLM capability will also benefit the VLM scoring process and improve the overall performance implicitly.

C More Ablation Studies

C.1 World Models

We evaluated the performance of two world models, Search World Model (SWM) and Stable Virtual Camera (SVC), on a dataset generated through the AI2-THOR simulator, as AI2-THOR is out-of-domain for both world models. The evaluation includes both quantitative metrics, measuring the accuracy of predictions and the quality of the generated images, and qualitative comparisons through visualizations of representative samples.

During inference, both models are executed with 50 diffusion steps. Specifically, to get the metrics of generated quality, we generated 10 episodes for each of the 208 scenes in AI2-THOR. Each episode consists of an action sequence of 8 steps, where at each step, an action is randomly selected from the primitive action set:

{move forward 0.25 meter, turn left 9 degrees, turn right 9 degrees}

C.1.1 Quantitative Comparison

More specifically, following the previous work of stable virtual camera, we tested the prediction result of our world models using standard metrics-peak signal-to-noise ratio (PSNR), learned perceptual image patch similarity (LPIPS), and structural similarity index measure (SSIM). Results are shown in

Table 1: Video Generation Results. Comparison of SWM and SVC in both visual quality and consistency.

Method	PSNR \uparrow	SSIM \uparrow	LPIPS(1e-4) \downarrow
SVC	64.51 \pm 0.27	0.994 \pm 0.01	0.49 \pm 0.01
SWM	66.59 \pm 0.21	0.997 \pm 0.01	0.31 \pm 0.01

Table 1, a quantitative comparison between two video generation models, SWM and SVC. These metrics jointly assess both visual fidelity and perceptual consistency.

SWM outperforms SVC in terms of PSNR (66.59 vs. 64.51) and SSIM (0.997 vs. 0.994), indicating more accurate and structurally consistent predictions. It also achieves a lower LPIPS score (0.31 vs. 0.49), suggesting that SWM generates images that are more perceptually similar to the ground truth. Overall, SWM demonstrates superior performance in terms of visual accuracy and perceptual similarity. This suggests that SWM is more effective at generating coherent and visually faithful video sequences for the primitive actions we defined.

C.1.2 Qualitative Comparison

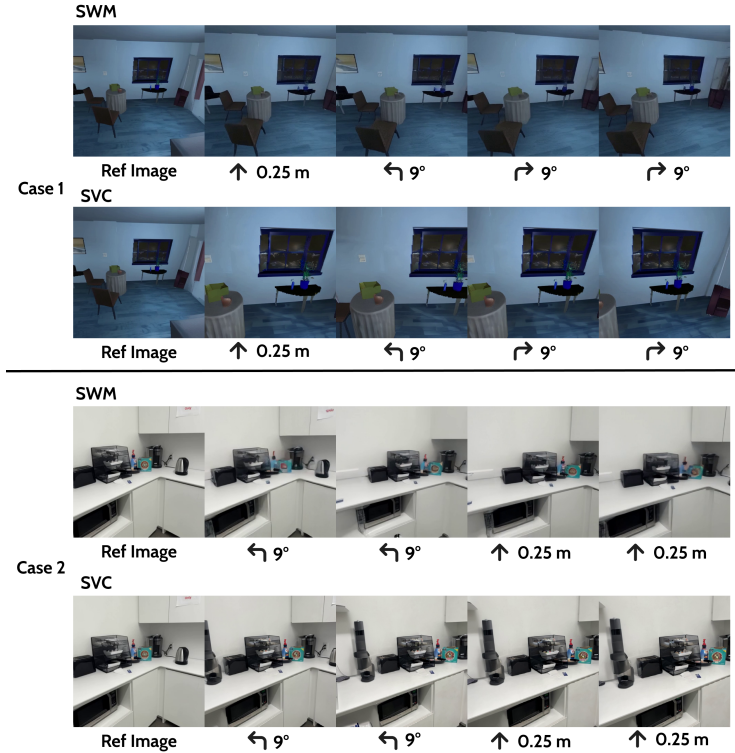


Figure 6: **Comparison of World Models.** Case 1 comes from validation split of SAT dataset; Case 2 comes from real-world test split of SAT dataset.

We present a qualitative comparison between the two models, SWM and SVC, using two representative examples from the synthesized validation set and real-world test set of SAT dataset. As shown in Case 1 from Fig. 6, SVC sometimes performs inaccurate forward motion. After moving forward by 0.25m, while the visual consequence from SWM seems reasonable, the outcome from SVC shows its inconsistency. In Case 2 Fig. 6, the SVC shows better capability of keeping object-level details. In the generated results from SWM, the objects become blurry as the video extends, but SVC successfully keeps details of each existing object. Generally, we observed that while SWM is more consistent in the scale of movement, SVC preserves more details during the camera movements.

Table 2: **Ablation on Trajectory Description.** Accuracy for large proprietary and MLMs on SAT-Real.

	SAT Real					
	Avg	EgoM	ObjM	EgoAct	GoalAim	Pers
GPT-4o	60.3	56.5	85.0	50.0	64.0	45.0
+ SN (SWM)	68.0	73.9	69.6	75.7	73.5	48.5
+ SN (SWM) , w/o Traj. Desc.	66.8	60.0	76.7	71.0	70.0	48.3
+ SN (SVC)	69.3	78.3	60.9	78.4	70.6	57.6
+ SN (SVC) , w/o Traj. Desc.	66.5	73.5	65.0	74.5	66.3	53.1
GPT-4.1	67.3	81.0	76.4	69.5	73.9	36.0
+ SN (SWM)	82.6	95.0	78.2	89.0	85.0	66.6
+ SN (SWM), w/o Traj. Desc.	73.0	100.0	78.2	67.7	75.8	53.1

Table 3: **Ablation on Trajectory Description.** Accuracy for large proprietary MLMs on SAT-Synthesized.

	SAT Synthesized					
	Avg	EgoM	ObjM	EgoAct	GoalAim	Pers
GPT-4o	61.0	64.7	86.8	51.9	68.7	43.4
+ SN (SWM)	70.8	77.6	82.6	70.1	84.5	45.8
+ SN (SWM) , w/o Traj. Desc.	64.1	67.9	83.1	63.0	71.8	42.3
+ SN (SVC)	72.3	80.0	84.8	65.0	89.3	51.4
+ SN (SVC) , w/o Traj. Desc.	65.8	64.7	83.3	68.4	65.8	47.8
GPT-4.1	66.4	75.3	89.0	57.8	78.3	41.5
+ SN (SWM)	75.4	88.2	92.4	70.8	89.3	45.8
+ SN (SWM), w/o Traj. Desc.	72.6	89.3	92.3	65.0	88.1	33.4

174 C.2 Ablation on Trajectory Description

175 In our current method, for each observation in the evidence buffer, we also provide a natural language
 176 trajectory description that explains its relationship with the initial reference image. We demonstrate
 177 that the trajectory description is necessary for our method in Table 2 and Table 3. According to the
 178 tables, we observe that the performance of our method drops on both SAT-Real and SAT-Synthesized
 179 for all VLMs and world models.

180 D Prompts

181 Here we provide 4 different prompts used in our method. The baseline prompt is shown in Fig. 7.
 182 The exploration Scoring prompt is shown in Fig. 8. The helpful Scoring prompt is shown in Fig. 9.
 183 The question-answering prompt using SpatialNavigator is shown in Fig. 10.

<p>System Prompt:</p> <p>Task: You are an AI assistant designed to help with spatial reasoning in a 3D indoor scene. You must analyze any provided images or observations and answer the question.</p> <p>Content Prompt:</p> <p>These are the images that pair with the question. Image 1: {image} Image 2: {image} ...</p> <p>Question: {question} Answer Choices: {answer choices}</p> <p>Output the exact answer from the choices. Answer:</p>

Figure 7: **Prompt for Baseline Q&A.**

System Prompt:

Task: You are an AI assistant designed to help with spatial reasoning in a 3D indoor scene. You must analyze any provided images and score imagined images based on how suitable they are for exploring these action consequences in order to answer the question from the choices.

Rules:

1. You'll be provided with images (including imagined images), a question, and a set of answer choices. You should score all imagined images.
2. You should output a list of scores from 0 to 9, separated by ','. For example: Output: 3,5,2,9,0,1

Content Prompt:

These are the images that pair with the question.
Image 1: {image}
Image 2: {image}
...
Image 1 is your current egocentric view.

Question: {question}
Answer Choices: {answer choices}

Below are the imagined views after taking actions.
Imagined image of index {index} if you {action sequence}:
{image}

Below are the imagined views after taking actions.
...
Output a list of scores.
Output:

Figure 8: **Prompt for Exploration Scoring.**

System Prompt:

Task: You are an AI assistant designed to help with spatial reasoning in a 3D indoor scene. You must analyze any provided images and score imagined images based on how helpful they are for answering the questions.

Hint: They may not be correct for answering the questions, but they will be helpful for excluding the wrong answers. The scores should also consider the image quality. If the image quality is very bad, it should receive a low score. Otherwise, the score should be augmented.

Rules:

1. You'll be provided with images (including imagined images), a question, and a set of answer choices. You should score all imagined images.
2. You should output a list of scores from 0 to 9, separated by ','. For example: Output: 3,5,2,9,0,1"

Content Prompt:

These are the images that pair with the question.
Image 1: {image}
Image 2: {image}
...
Image 1 is your current egocentric view.

Question: {question}
Answer Choices: {answer choices}

Below are the imagined views after taking actions.
Imagined image of index {index} if you {action sequence}:
{image}

Below are the imagined views after taking actions.
...
Output a list of scores.
Output:

Figure 9: **Prompt for Helpful Scoring.**

System Prompt:

Task: You are an AI assistant designed to help with spatial reasoning in a 3D indoor scene. You must analyze any provided images or observations and answer the question.

Rules:

1. You should output the exact answer from the choices.
2. You will be provided with multiple imagined views if you take corresponding actions to help you answer the questions.
3. You can include minimal reasoning, but your final line must only include the exact answer choice.

Content Prompt:

These are the images that pair with the question.

Image 1: {image}

Image 2: {image}

...

Image 1 is your current egocentric view.

Question: {question}

Answer Choices: {answer choices}

Below are the imagined views you would obtain if you took the corresponding actions. These are provided to help you answer the question.

You can include them in your reasoning, but you should still only output the exact answer at the last line.

Action: {action catalog}

{action sequence}

{image}

Action: {action catalog}

...

Output the exact answer from the choices.

Answer:

Figure 10: Prompt for Q&A using SpatialNavigator.

184 E Broader Impacts

185 By allowing vision–language models to build and interrogate a physically consistent “mental
 186 workspace,” our method could accelerate progress in assistive robotics, remote inspection, and
 187 immersive training: robots that better understand 3D space can navigate cluttered homes for elder
 188 care, inspect hazardous sites without human entry, and deliver richer AR/VR experiences for ed-
 189 ucation or therapy. At the same time, safer decision-making from imagined roll-outs may reduce
 190 real-world trial-and-error, lowering both cost and risk. Yet the technology also raises concerns. More
 191 capable spatial reasoning can enhance autonomous surveillance systems or military platforms; and
 192 greater autonomy could displace certain manual-labor jobs. Finally, training large video world models
 193 consumes considerable energy and inherits any biases present in the data (e.g., under-representation
 194 of certain environments). Researchers and practitioners should therefore pair technical advances
 195 with robust provenance tracking for generated content, scenario-specific safety constraints, and
 196 data-diversity audits, while favouring energy-efficient architectures and openly reporting compute
 197 footprints.

References

- Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. DI3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 2
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL <https://arxiv.org/abs/2112.10752>. 3
- Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models, 2022. URL <https://arxiv.org/abs/2202.00512>. 3
- Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. *Advances in Neural Information Processing Systems*, 34:19313–19325, 2021. 3
- Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat, 2022. URL <https://arxiv.org/abs/2106.14405>. 2
- Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 3
- Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snaveley. Stereo magnification: Learning view synthesis using multiplane images, 2018. URL <https://arxiv.org/abs/1805.09817>. 2