This CVPR Findings paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the accepted version; the final published version of the proceedings is available on IEEE Xplore.

UnrealSpace: Analyzing Spatial Understanding and Reasoning in Controllable Simulation

Wufei Ma*, **Sky Cen***, **Jianzhi Shen***, **Rex Lee***, **León Begiristain**, **Yan Zhuang**, **Jiawei Peng**
Zhifei Yu, **Tianao Song**, **Xinyuan Qi**, **Tianmin Shu**, **Adam Kortylewski**†, **Alan Yuille**
Johns Hopkins University †CISPA Helmholtz Center for Information Security

Abstract

*Accurate and robust spatial understanding is crucial for the control and safety of embodied AI and robotic systems. Yet current evaluation practices remain fragmented, where different tasks adopt distinct benchmarks and datasets. This impedes a comprehensive evaluation of spatial understanding models from a holistic scene-understanding perspective and limits fine-grained analysis of their performance under challenging out-of-distribution conditions. In this work we present **Unreal3DSpace**, a unified and controllable framework for comprehensive evaluation of spatial understanding. Our Unreal3DSpace features controllable and procedural generation of diverse and high-fidelity 3D scenes and the availability of comprehensive 2D and 3D ground-truth annotations. This enables the evaluation of a wide range of spatial understanding models within a single, consistent framework and allows systematic studies of model robustness under varying degrees of scene- and object-level distribution shifts. Moreover, the comprehensive 3D ground truths also support analysis of the inference trajectories of spatial reasoning models, allowing us to pinpoint the underlying factors that lead to failure cases. Experimental results on our Unreal3DSpace highlight key limitations of current spatial understanding models, particularly in handling distant objects, partial occlusion, and questions with multiple seemingly plausible answers. We further introduce an agentic model that integrates multiple spatial understanding modules to infer spatial relationships, which pinpoints the primary bottlenecks contributing to errors in final predictions. Overall, evaluation results on our Unreal3DSpace provide valuable insights into the weaknesses of existing spatial understanding models and inform directions for future research in this area. Our code and data is available from our project page.*

1. Introduction

Recent advancements of spatial understanding have driven remarkable progress in embodied AI and robotics. The capability to comprehend scene layouts, object relationships, and intrinsic physics of 3D world enables more effective 3D-aware reasoning in large multi-modal models [20, 26, 34], as well as enhanced instruction following and task planning prior to action execution [1, 52].

To systematically evaluate spatial understanding capabilities and reveal limitations of current multi-modal systems, a variety of spatial benchmarks have been proposed, targeting spatial reasoning from different modalities (*e.g.*, single image [19], multiple frames [43, 47, 49], and videos [25, 44]) or with different task formulations (*e.g.*, question answering [19, 22], grounding/localization [2, 6], and task success rate [1, 51]). However, although these benchmarks offer a high-level measurement of spatial understanding capabilities, they provide limited insight into the models' failure modes, in particular, what causes these errors and which components constitute the key bottlenecks.

Central to this challenge is the absence of comprehensive 2D and 3D ground-truth annotations that go beyond simple evaluation labels, such as question–answer pairs or segmentation masks, which hinders deeper diagnosis of model failures and prevent a unified understanding of how different spatial understanding components interact within a scene. Consequently, several recent spatial benchmarks were developed using synthetic data generated by simulation environments [31, 38, 39]. Those benchmarks include detailed 2D and 3D information about the scene that are typically hard to annotate, such as segmentation maps, 3D bounding boxes, and object orientations, which can be easily obtained from the simulation. However, these benchmarks remain largely object-centric, lacking realistic backgrounds, lighting, and contextual interactions, and thus fail to support a holistic, fine-grained evaluation of spatial understanding.

In this work, we introduce ***Unreal3DSpace***, a unified and controllable simulation suite for comprehensive evaluation of spatial understanding. Notably, our Unreal3DSpace features three key design choices: (1) *A data engine that enables fine-grained scene-level controllability*, enabling diverse and procedural generation of object placements across diverse scene contexts and sampling of key scene parameters such as density, weather, lighting conditions, and visual clutter; (2) *Comprehensive 2D and 3D ground-truth annotations*: We obtain a range of annotations from the simulation, including depth and surface normals, instance segmen-

*Equal first author.



Figure 1. **Overview of Unreal3DSpace.** We present Unreal3DSpace, a unified evaluation suite for 3D spatial understanding. Unreal3DSpace features controllable and procedural generation of 3D scenes, along with comprehensive 2D and 3D ground truth annotations. With Unreal3DSpace, we evaluate both task-oriented spatial understanding models and multi-modal reasoning models within a unified framework, enabling fine-grained analyses of their performance under varying scene- and object-level distribution shifts.

tation, visibility and occlusion from the z-buffer, 3D object poses and spatial relationships, and additional scene-level geometric and lighting properties; (3) *high-fidelity simulation from unreal engine*: our Unreal3DSpace is built on realistic simulation in Unreal Engine, as well as 3D scenes and assets made by professional artists.

The fine-grained controllability and comprehensive 2D and 3D ground truths in Unreal3DSpace address two key limitations in previous benchmarks and datasets:

First, previous research evaluated spatial understanding models on diverse datasets, each intended for unique domains and data processing. For instance, depth estimation models [45, 50] are commonly assessed on NYUv2 [23] and SUN RGB-D [35], while segmentation models [14, 32] rely on semantic datasets such as LVIS [11] and ADE20K [53]. Similarly, pose estimation models [17, 40] are analyzed using object-centric PASCAL3D+ [42] and ImageNet3D [18]. This fragmented evaluation hinders fair and coherent comparisons across models from a holistic spatial understanding perspective. For example, it remains unclear which component constitutes the key bottleneck when forming a comprehensive spatial interpretation of an image. In contrast, our Unreal3DSpace addresses these limitations by adopting a unified and comprehensive evaluation framework. With various 2D and 3D ground truths, our framework further enables fine-grained evaluation on different scene- and object-level properties, such as assessing the robustness of depth models under object partial occlusion or weather and scene distribution shifts.

Moreover, comprehensive 2D and 3D ground truths enable a detailed analysis of spatial reasoning models, allowing us to pinpoint the underlying causes of these failures. Rather than directly predicting the answers to spatial-related questions, recent research explores chain-of-thought reasoning in spatial-aware vision-language models [15, 20] or agentic approaches that build on intermediate predictions from task-specific models [3, 41]. Beyond evaluating benchmark performance on the final predictions of reasoning models, the comprehensive 2D and 3D ground truths in our Unreal3DSpace offer valuable insights into their reasoning trajectories. This enables detailed analysis of where and why failures occur, such as by pinpointing the specific reasoning steps that lead to incorrect answers or identifying which modules (*e.g.*, depth estimation or pose estimation) that contribute most to these errors. Besides spatial reasoning models, our Unreal3DSpace also sheds light on the design of multi-step training data generation methods [5, 21] or reinforcement learning of multi-modal models with 3D-aware rewards [20].

We evaluate a variety of spatial understanding models, including scene-level visual foundation models for depth estimation and instance segmentation, object-level visual foundation models for pose estimation, and vision-language models for spatial reasoning VQA. By varying data sampling parameters during simulation, we study how models behave under different camera viewpoints, scene density, partial occlusion, and various weather and lighting conditions. Our results highlight that both instance segmentation and pose estimation models are highly susceptible to partial occlusion, due to their reliance on 2D object semantics and lack of 3D occlusion awareness. Meanwhile, pose estimation models exhibit relatively stronger robustness to occlusion by leveraging broader contextual cues. However, their performance degrades sharply once the object distance exceeds a certain threshold, leading to a substantial increase in absolute relative error.

With comprehensive 2D and 3D ground truths, we further analyze the reasoning trajectories of an agentic model built on a neural-symbolic reasoner [38, 48]. By iteratively correcting the errors that arise in intermediate steps, we find that pose estimation emerges as the primary bottleneck limiting the final answer accuracy. This type of evaluation on reasoning trajectories can be further extended to other reasoning models [4, 20], providing valuable insights into how their spatial reasoning capabilities can be further improved.

In summary, we present Unreal3DSpace, a unified eval-

uation suite for spatial understanding. Our Unreal3DSpace features high-fidelity simulation built on Unreal Engine with professionally designed assets, and achieves enhanced diversity through procedural generation and challenging out-of-distribution sampling. With comprehensive 2D and 3D ground truths, our benchmark unifies the evaluation of scene-level, object-level, and multi-modal spatial understanding models on a shared dataset. This unified framework enables consistent comparison across tasks and offers holistic perspectives that guide research toward general-purpose spatial understanding.

2. Related Works

Spatial understanding and reasoning. A comprehensive understanding of 3D scenes is a challenging yet fundamental research direction. It serves as the basis for both generative world modeling, such as synthesizing novel views while maintaining physical consistencies [7, 54] or placing objects into physically plausible locations [13], and a 3D-aware reasoning that is crucial for embodied agents and robots. Prior efforts in this direction span a variety of visual foundation models for segmentation [14, 30], depth estimation [45, 46, 50], and object pose estimation [18, 40]. To advance toward reasoning about scene-level spatial information, such as static spatial relationships or temporal spatial events [19, 38, 39], recent studies have post-trained vision-language models for spatial reasoning [5, 8, 20, 21].

Evaluating spatial understanding models. Thoroughly quantifying spatial understanding capabilities and pinpointing their limitations is difficult. Previous research often explored task-specific datasets, *e.g.*, NYUv2 [23] and SUN RGB-D [35] for depth estimation and PASCAL3D+ [42] and ImageNet3D [18] for pose estimation. Moreover, the lack of 3D ground truths in spatial VQA datasets hinders the analysis of reasoning trajectories besides accuracy of the final answer. Lastly, it is largely understudied how these spatial understanding models perform in challenging, out-of-distribution conditions, due to the limited availability of 3D ground truths. Our Unreal3DSpace addressed these issues by presenting a unified evaluation suite for spatial understanding, supporting the evaluation of various tasks on the same data, and simulating diverse, controllable, and high-fidelity data for evaluation on different weather conditions, lighting conditions, scene density, etc.

Synthetic datasets from simulation. Synthetic datasets offer accurate and valuable 2D and 3D ground truths that are often unavailable in real-world image datasets [27, 28, 33]. However, generating 3D scenes with sufficient diversity, controllability, and photorealism requires substantial efforts. Previous synthetic datasets are often limited to specific domains [10, 29, 38], constrained by fixed scene layouts [33, 37], or lacking realistic lighting and material appearance [9, 10]. Our Unreal3DSpace instead builds on Unreal Engine 5 and high-quality 3D scene assets created by professional artists, enabling procedural generation of diverse, high-fidelity scenes that support a comprehensive and fine-grained evaluation of spatial understanding models.

3. Unreal3DSpace

In this section, we introduce Unreal3DSpace, a unified evaluation suite for spatial understanding. We begin by outlining the 3D scene setup and procedural generation process in Section 3.1. Building on this, we describe the collection of 2D and 3D ground-truth annotations derived from the data generation pipeline in Section 3.2. Then in Section 3.3, we discuss different generation parameters and sampling distributions used to diversify the data, ranging from uncluttered scenes under clear midday lighting to cluttered, foggy environments at dawn. Finally, we summarize key dataset statistics in Section 3.4 and refer the readers to Section B in the supplementary materials for additional technical details about our simulation pipeline.

3.1. Scene Simulation

Generating 3D scenes with diversity, controllability, and high fidelity requires substantial efforts. Existing synthetic scene datasets are often constrained to specific domains [10, 29, 38], rely on fixed scene layouts [33, 37], or suffer from limited realism in lighting and appearances [9, 10]. To overcome these limitations, we build upon Unreal Engine and leverage high-quality 3D assets created by professional artists. To enable controllable and procedural generation, we further define a set of procedural rules that govern the configuration and object placement within the 3D scenes.

Procedural generation. For each 3D scene, we define four types of procedural rules: vehicle trajectories, pedestrian trajectories, road areas, and sidewalk areas. Vehicle and pedestrian trajectories are used to simulate typical movement patterns, such as cars driving along streets or pedestrians walking on sidewalks. On the other hand, road and sidewalk areas specify where corresponding objects can be placed. For example, road areas may include traffic barriers and bicycles, whereas sidewalk areas may contain vending machines or hand carts. Our procedural generation pipeline builds on these rules to achieve a balance between scene diversity and visual fidelity.

3.2. 2D and 3D Ground Truths

We obtain a range of 2D and 3D annotations by simulating realistic 3D scenes in Unreal Engine with the LychSim plugin [16]. These include 2D ground truths, *e.g.*, segmentation masks, and 3D ground truths such as depth maps and object 3D bounding boxes, which are also considered in prior works [27, 28, 38]. We further consider the following ground truths not available in previous studies yet critical for the evaluation of spatial understanding models.

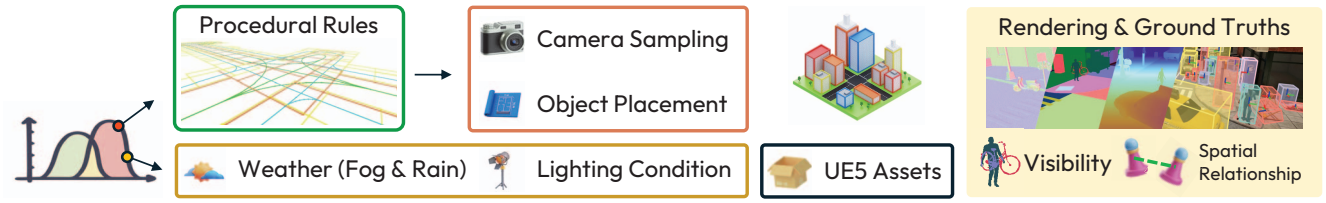


Figure 2. **Overview of the procedural generation pipeline for Unreal3DSpace.** To enable controllable generation of 3D scenes and evaluate 3D spatial understanding on data sampled from different distributions, we annotate procedural rules for object and camera sampling, parameterize weather and lighting simulations, and draw random data samples from Gaussian mixtures. This enables a unified and fine-grained evaluation of spatial understanding under different data distributions.

Visibility and occlusion from z-buffer. We extend the LychSim plugin to extract z-buffer data from the rendering pipeline. This enables analysis not only of the visible regions from the camera viewpoint but also of areas occluded by other objects. From the z-buffer data, we derive crucial ground-truth information about the 3D scene, including object occlusion ordering, amodal segmentation masks, as well as object-level visibility and occlusion rates. These data enable a fine-grained evaluation of spatial understanding models under challenging conditions, such as model robustness to different levels of partial occlusion.

Spatial relationships. Previous spatial understanding models have leveraged various visual foundation models to infer object-level 3D information and derive spatial relationships from 3D pseudo-labels [5, 21]. However, the resulting spatial relationships are often noisy, as even minor errors at any stage of the data processing pipeline can propagate and lead to highly inaccurate predictions. Meanwhile, the spatial relationships in our Unreal3DSpace are constructed from accurate 3D ground truths directly obtained from simulation. These accurate 3D ground truths enable two key improvements crucial for evaluating spatial understanding models: (1) *object physical contact*: In addition to the standard 3D and 6D spatial relationships studied in prior works [19, 38], we further extract physical contact information from the Unreal Engine simulation, which cannot be reliably derived from noisy pseudo-labels or even human annotations. (2) *question difficulty*: Besides binary labels indicating the spatial relationships, *e.g.*, whether "A is on the left or right of B", we further estimate a difficulty metric for each spatial relationship, reflecting how challenging the question is to answer. For example, when comparing whether "A" is closer to "B" or "C", a low difficulty indicates a large difference between the distances of "AB" and "AC", whereas a high difficulty suggests that the two distances are similar and thus harder to distinguish. These enhancements are made possible by the availability of accurate 3D ground truths, which enables a more comprehensive and fine-grained evaluation.

3.3. Sampling Distributions

A key feature of our Unreal3DSpace is its fine-grained scene-level controllability. We parameterize the procedural sampling with weather, camera, and lighting configurations. By drawing samples from predefined Gaussian mixtures, we largely enhance the diversity of the synthetic data and enhance the diversity of the synthetic data and enable comprehensive, fine-grained evaluation of spatial understanding models under diverse and challenging conditions. Please refer to Section B.2, where we present qualitative results sampled from different parameter distributions.



Figure 3. **Weather and lighting sampling.** From left to right, we show example images of fog, a clear day, dawn, and rain. As discussed in Section 3.3, we parameterize the weather and lighting simulations and draw random samples from Gaussian mixtures.

Weather conditions. The safety of autonomous vehicles, robots, and embodied AI systems under extreme weather conditions remains a critical challenge for real-world deployment. To examine how spatial understanding models perform under different weather conditions, we simulate rain and fog within our environment. Unlike image-level corruptions and perturbations explored in prior works [12], our simulations operate directly in 3D, providing enhanced 3D-related fidelity. For instance, nearby objects remain relatively unaffected by the fog, while distant objects appear progressively less visible. We parameterize fog by its density and rain by the droplet spawn rate and quantity. By sampling these parameters from Gaussian mixtures, we generate synthetic data exhibiting varying levels of fog and rain intensity.

Camera parameters. Spatial understanding models are fundamental to a wide range of embodied systems, including quadruped robots, humanoid robots, and unmanned aerial vehicles (UAV), where accurate perception of 3D

environments is essential for safe and effective operation. We sample diverse camera parameters to facilitate a thorough evaluation aligned with potential downstream deployments. In particular, our sampling covers different camera heights, corresponding to quadruped, humanoid, and UAV platforms, and varying fields-of-view (FoV) that simulates different sensing configurations and perceptual ranges.

Lighting sampling. Our Unreal3DSpace also samples from different lighting conditions that simulate the lighting during different times of the day. We parameterize lighting by the elevation and azimuth angles of the directional sunlight, as well as the color temperature that controls its warmth. We sample these parameters from Gaussian mixtures that simulates dawn/dusk, noon, and night.

Scene density and visual clutter. Lastly, we sample 3D scenes with varying levels of scene density and visual clutter, characterized by a different numbers of vehicles and objects in each scene, as well as whether objects are standing upright or have fallen over.

3.4. Dataset Statistics

Our Unreal3DSpace comprises 24 high-quality 3D scenes created by professional artists. We further populate the 3D scenes with 217 3D objects spanning 72 categories, create diverse lighting conditions by varying sun height and color temperature, and sample random yet plausible camera extrinsics. In total, we generate 19,200 samples with paired renderings and corresponding 2D and 3D ground truths. The generated data cover diverse distributions of lighting, scene density, object placement, and partial occlusion. Please refer to Section B for more technical details about the synthetic data in our Unreal3DSpace.

Scalability. We note that our Unreal3DSpace is built on a semi-automatic and highly scalable data engine. For each 3D scene, minimal efforts are necessary to specify procedural rules. Afterwards, abundant and diverse synthetic data with comprehensive 2D and 3D ground truths can be generated without any human intervention. Although beyond the scope of this work, our Unreal3DSpace can be adopted as a scalable approach to create synthetic training data to improve the robustness and generalization ability of visual foundation models or vision-language systems.

4. Experiments

Our Unreal3DSpace presents a unified and controllable evaluation suite that enables fine-grained evaluation of different spatial understanding models on the same data sampled from different distributions. We begin in Section 4.1 with a fine-grained evaluation of task-oriented spatial understanding, studying their behavior under levels of partial occlusion and distribution shifts. Then, in Section 4.2, we focus on spatial reasoning models, particularly their performance on different levels of problem difficulties due to object occlusion and ambiguity between answer choices. Finally in Section D, informed by the insights from both task-level and reasoning-level evaluations, we analyze the failure cases of an agentic method to pinpoint which module serves as the primary bottleneck hindering holistic scene understanding.

4.1. Task-Oriented Spatial Understanding

Experimental setup. In this section, we focus on three spatial understanding tasks: instance segmentation, depth estimation, and object pose estimation. We choose a state-of-the-art visual foundation model for each task, *i.e.*, SAM 2 [30], Depth Anything v2 [46], and Orient Anything [40], and study their performance under different data distributions. In terms of evaluation metrics, we use Intersection over Union (IoU) for segmentation, absolute relative error (AbsRel) for depth estimation, and pose accuracy under $\pi/6$ threshold (Acc@$\pi/6$) for pose estimation. For quantitative comparisons with parallel methods, we refer the readers to Section C of the supplementary materials.

Object distance and partial occlusion. We evaluate visual foundation models on instance segmentation, depth estimation, and pose estimation tasks using data that varies in object distance, degree of occlusion, and weather and lighting conditions. From the results illustrated in Figure 4, we draw the following conclusions. **(1) All visual foundation models maintain a relatively stable performance across a moderate range of object distances**, but their performance drops sharply once the distance exceeds this range. **(2) Model performance begins to degrade even under low levels of partial occlusion.** Notably, low levels of partial occlusion are common in the training data of these visual foundation models. This highlights that, despite the success of data-driven approaches, current visual foundation models still lack intrinsic knowledge of the 3D world. As a result, they remain largely constrained to 2D appearance cues rather than reasoning about occlusion ordering in 3D space. **(3) Depth estimation is generally more robust to occlusion than instance segmentation.** We hypothesize that this is because segmentation models rely heavily on semantic cues and are hence more sensitive to the degree of partial occlusion. In contrast, depth estimation models can often remain reliable, as the depth of partially occluded objects can be inferred from surrounding contextual information. **(4) Pose estimation models also show better robustness to object distance**, whereas performance drops significantly even when the partial occlusion rate is below 0.25. We attribute this to the feed-forward ViT backbone of Orient Anything [40], which struggles to infer full-body orientation when only partial observations are available.

Distribution shifts of weather and lighting conditions. In Figure 4, we report task-oriented spatial understanding

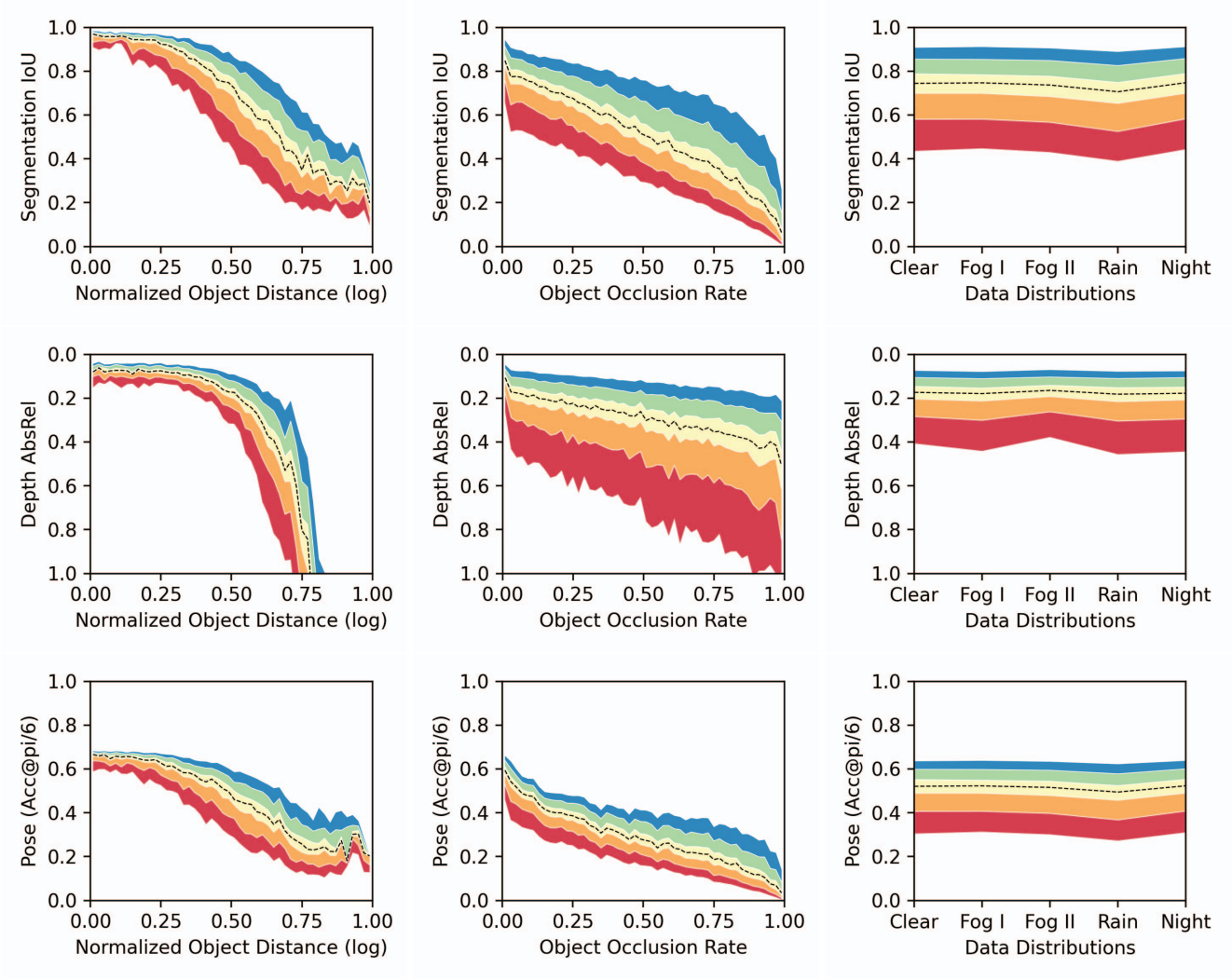


Figure 4. **The comprehensive 3D annotations in Unreal3DSpace enables fine-grained task-oriented spatial understanding on data varying in object distances, levels of partial occlusion, and weather and lighting conditions.** Results show that depth estimation models struggle to estimate accurate depth as object distances increase, while instance segmentation and pose estimation models are more adversely affected by partial occlusions. Please refer to the detailed discussions in Section 4.1.

results under different semantic distribution shifts, including clear day, fog level 1, fog level 2, rainy, and nighttime. Results show that visual foundation models exhibit relatively robust performance under different weather and lighting conditions. However, we note that when generating scenes under these less common conditions, we primarily vary the semantic attributes of the environment, *e.g.*, from "day" to "night" or from "clear day" to "raining", and our simulation engine still guarantees sufficient illumination and object visibility. Thanks to the large-scale image pretraining, *e.g.*, with DINOv2 [24], these visual foundation demonstrate robustness to the scene-level semantic changes.

4.2. Spatial Reasoning

Experimental setup. We follow the 3D spatial question types in [19] and generate about 22,500 spatial reasoning QA pairs. Using this data, we benchmark the spatial reasoning performance of state-of-the-art vision-language models and conduct fine-grained analyses of spatial reasoning capabilities under various scene- and object-level distribution shifts, as well as across different levels of question difficulty (see Section 3.2).

Baselines VLMs. We evaluate a collection of VLMs, including open-weight Cambrian-1 [36] and Qwen2.5VL [4], SpatialReasoner [20] with spatial-aware post-training, and proprietary GPT-5.

Spatial reasoning under distribution shifts. Similar to the results in Figure 4, we find that spatial reasoning models also struggle with increasing object distances and partial occlusion, while being robust to semantic changes of weather and lighting conditions.

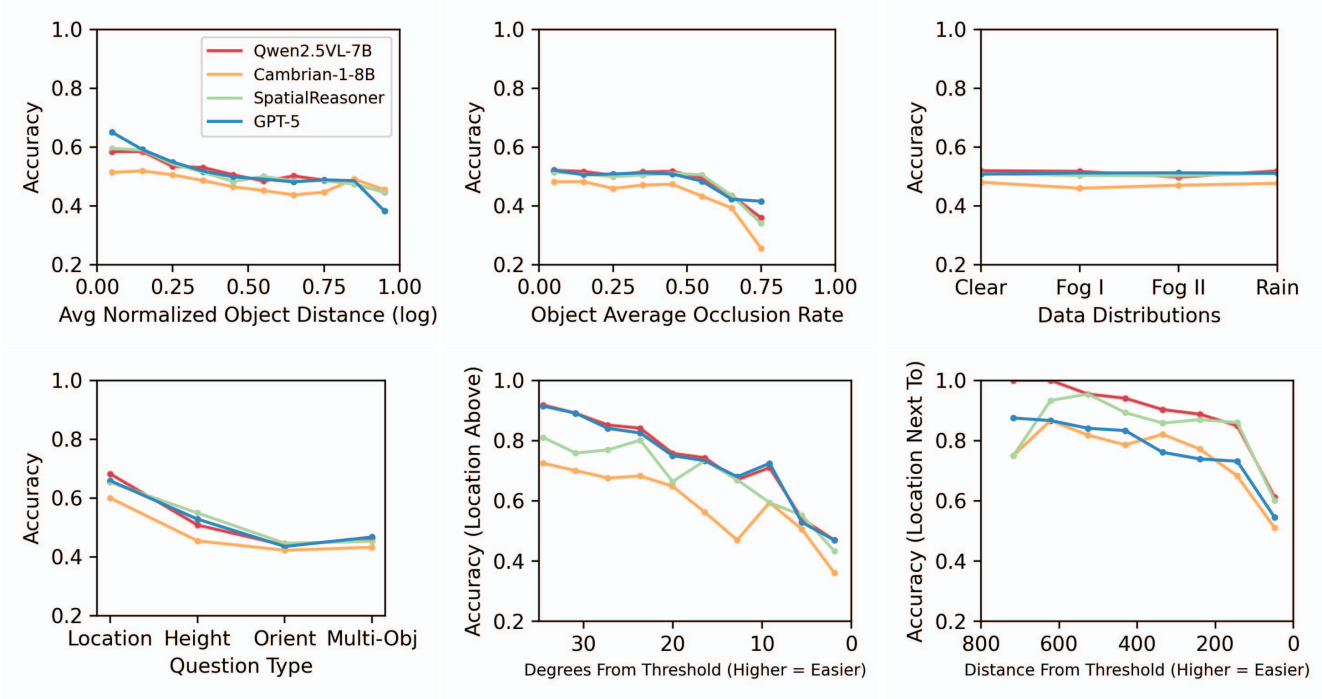


Figure 5. **Spatial reasoning performance on data varying in scene- and object-level distribution shifts (top), and with different levels of question difficulty (bottom).** Results highlight that state-of-the-art VLMs are not robust to object partial occlusion and struggle with spatial VQAs that demand higher levels of precision. Please refer to the detailed discussions in Section 4.2.

Question difficulty. From the bottom plots in Figure 5, we find that: (1) Consistent with prior findings [19], all VLMs struggle with questions involving object orientations or multi-object spatial reasoning; (2) Spatial reasoning performance drop substantially, from over 80% to 40%, as question difficulty increases. This degradation arises because VLMs can often rely on shortcuts or other 2D cues to answer simpler questions, whereas more challenging questions with multiple seemingly plausible answers require genuinely 3D-aware reasoning. Overall the fine-grained spatial reasoning evaluation enabled by Unreal3DSpace highlights the limited 3D awareness of current multi-modal models and their frequent failures on spatial VQA tasks that demand higher levels of precision.

4.3. Evaluating Reasoning Trajectory

In this section, we analyze the reasoning trajectories when answering questions about spatial relationships. In particular, we focus on identifying which components of the reasoning process most often lead to incorrect final answers.

Baseline agentic model for spatial reasoning. We consider a baseline agentic model composed of two parts: (1) *spatial understanding tools*, including various state-of-the-art visual foundation models such as for metric depth estimation and pose estimation, and (2) *neural-symbolic reasoner* [38, 48] which parses questions into executable programs and aggregates intermediate predictions to produce the final answer. The overall pipeline of this agentic model is illustrated in Figure 6. We note that this design effectively abstracts away linguistic reasoning, allowing us to isolate and evaluate holistic scene understanding and the effectiveness of visual foundation models.

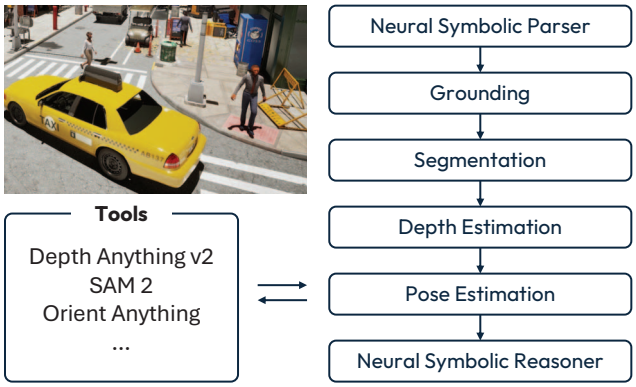


Figure 6. **Baseline agentic model with neural-symbolic reasoner.**

Experimental setup. Given the baseline agentic model, we aim to identify the key bottlenecks that lead to errors in the final prediction and limit the effectiveness of the overall system. Our assumption is that "improving" the predictions of the key bottleneck would lead to a significant increase in the final performance. Therefore, with the comprehensive 3D ground truth in Unreal3DSpace, we iteratively correct

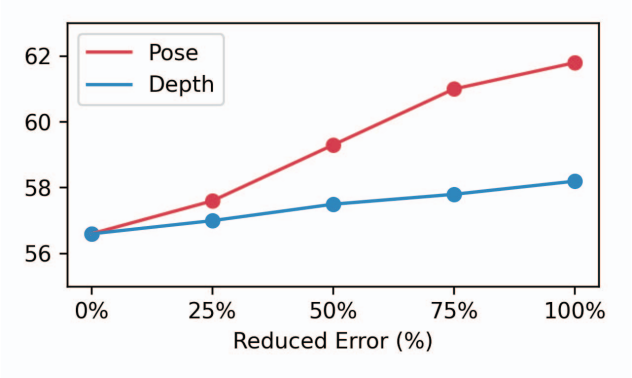


Figure 7. **Analyzing reasoning failures of an agentic model.**

a certain percentage of the prediction errors made by the visual foundation models, in particular, object depth from Metric3D [50] and pose error from Orient Anything [40]. Then we re-evaluate the neural symbolic model and observe the change in final answer accuracy.

Results. With the comprehensive 3D ground truths in Unreal3DSpace, we iteratively refine the metric depth predictions and object poses by reducing the errors by 25%, 50%, 75%, and 100%. After each refinement stage, we re-run the neural symbolic reasoner with the updated intermediate results. From Figure 7, we conclude that pose estimation is the primary bottleneck limiting the spatial reasoning performance of the agentic model. Analyses of the model's intermediate spatial reasoning trajectories offer valuable insights into its underlying mechanisms and inform effective strategies for further improving its performance.

Extensibility. Beyond agentic methods built on neural–symbolic models, our Unreal3DSpace can be readily extended to evaluate inference trajectories of other reasoning models, such as SpatialReasoner [20] that predicts explicit 3D measurements before producing the final answer, or agentic models equipped with a reasoning LLM [41]. This is made possible by the comprehensive 2D and 3D ground truths in Unreal3DSpace, which support unified and fine-grained evaluation of spatial reasoning models.

5. Conclusions

Accurate and robust spatial understanding is essential for improving the control and safety of embodied AI and robotic systems. However, existing spatial understanding evaluation remain fragmented, where different tasks adopted distinct benchmarks and datasets. This hinders comprehensive evaluation of spatial understanding models from a holistic scene-understanding perspective and makes it difficult to pinpoint the key bottlenecks in reasoning models that must perform multiple spatial tasks. In this work, we introduce Unreal3DSpace, a unified evaluation suite for spatial understanding. Our benchmark provides a high-fidelity Unreal Engine simulation and a controllable, procedural data engine. This framework enables systematic evaluation of spatial understanding models across diverse data distributions, including variations in scene density, partial occlusion, and weather conditions. Our simulation engine further generates comprehensive 3D ground truth data, enabling fine-grained analysis of model behavior and evaluation of the intermediate trajectories produced by reasoning models. Evaluation results on our Unreal3DSpace shed new light on the current limitations of spatial understanding models and open new directions for advancing models that achieve unified, holistic spatial understanding.

Acknowledgements

WM and AY acknowledge support from ONR with N00014-23-1-2641 National Eye Institute (NEI) with Award ID: R01EY037193. This work was also supported in part by the Whiting School of Engineering at Johns Hopkins University.

References

[1] Abbas Abdolmaleki, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Ashwin Balakrishna, Nathan Batchelor, Alex Bewley, Jeff Bingham, Michael Bloesch, et al. Gemini robotics 1.5: Pushing the frontier of generalist robots with advanced embodied reasoning, thinking, and motion transfer. *arXiv preprint arXiv:2510.03342*, 2025. 1

[2] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas J. Guibas. ReferIt3D: Neural listeners for fine-grained 3d object identification in real-world scenes. In *16th European Conference on Computer Vision (ECCV)*, 2020. 1

[3] Anonymous. Pursuing minimal sufficiency in spatial reasoning. In *Submitted to The Fourteenth International Conference on Learning Representations*, 2025. under review. 2

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 6, 5

[5] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024. 2, 3, 4

[6] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. *16th European Conference on Computer Vision (ECCV)*, 2020. 1

[7] Zhaoxi Chen, Tianqi Liu, Long Zhuo, Jiawei Ren, Zeng Tao, He Zhu, Fangzhou Hong, Liang Pan, and Ziwei Liu. 4dnex: Feed-forward 4d generative modeling made easy. *arXiv preprint arXiv:2508.13154*, 2025. 3

[8] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. In *NeurIPS*, 2024. 3

[9] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. *Advances in Neural Information Processing Systems*, 35:5982–5994, 2022. 3

[10] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4340–4349, 2016. 3

[11] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[12] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 4

[13] Ian Huang, Yanan Bao, Karen Truong, Howard Zhou, Cordelia Schmid, Leonidas Guibas, and Alireza Fathi. Fireplace: Geometric refinements of llm common sense reasoning for 3d object placement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13466–13476, 2025. 3

[14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 2, 3

[15] Yuecheng Liu, Dafeng Chi, Shiguang Wu, Zhanguang Zhang, Yaochen Hu, Lingfeng Zhang, Yingxue Zhang, Shuang Wu, Tongtong Cao, Guowei Huang, et al. Spatialcot: Advancing spatial reasoning through coordinate alignment and chain-of-thought for embodied task planning. *arXiv preprint arXiv:2501.10074*, 2025. 2

[16] Wufei Ma, Siyi Chen, Rex Lee, Jiawei Peng, and Alan Yuille. LychSim: Analyzing vision systems in controllable and interactive simulation. Accessed on 2025-09-15. 3

[17] Wufei Ma, Angtian Wang, Alan Yuille, and Adam Kortylewski. Robust category-level 6d pose estimation with coarse-to-fine rendering of neural features. In *European Conference on Computer Vision*, pages 492–508. Springer, 2022. 2

[18] Wufei Ma, Guofeng Zhang, Qihao Liu, Guanning Zeng, Adam Kortylewski, Yaoyao Liu, and Alan Yuille. Imagenet3d: Towards general-purpose object-level 3d understanding. *Advances in Neural Information Processing Systems*, 37:96127–96149, 2024. 2, 3

[19] Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6924–6934, 2025. 1, 3, 4, 6, 7

[20] Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso M de Melo, Jianwen Xie, and Alan Yuille. Spatialreasoner: Towards explicit and generalizable 3d spatial reasoning. In *Advances in Neural Information Processing Systems*, 2025. 1, 2, 3, 6, 8, 5

[21] Wufei Ma, Luoxin Ye, Celso de Melo, Alan L Yuille, and Jieneng Chen. Spatialllm: A compound 3d-informed design towards spatially-intelligent large multimodal models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2025. 2, 3, 4

[22] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*, 2022. 1

[23] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2, 3

[24] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 6

[25] Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025. 1

[26] Zekun Qi, Wenyao Zhang, Yufei Ding, Runpei Dong, Xinqiang Yu, Jingwen Li, Lingyun Xu, Baoyu Li, Xialin He, Guofan Fan, Jiazhao Zhang, Jiawei He, Jiayuan Gu, Xin Jin, Kaisheng Ma, Zhizheng Zhang, He Wang, and Li Yi. Sofar: Language-grounded orientation bridges spatial reasoning and object manipulation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 1

[27] Weichao Qiu, Fangwei Zhong, Yi Zhang, Siyuan Qiao, Zihao Xiao, Tae Soo Kim, and Yizhou Wang. Unrealcv: Virtual worlds for computer vision. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1221–1224, 2017. 3

[28] Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zuo, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, Alejandro Newell, Hei Law, Ankit Goyal, Kaiyu Yang, and Jia Deng. Infinite photorealistic worlds using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12630–12641, 2023. 3

[29] Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, Zeyu Ma, and Jia Deng. Infinigen indoors: Photorealistic indoor scenes using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21783–21794, 2024. 3

[30] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3, 5

[31] Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, et al. Sat: Dynamic spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024. 1

[32] Tianhe Ren, Yihao Chen, Qing Jiang, Zhaoyang Zeng, Yuda Xiong, Wenlong Liu, Zhengyu Ma, Junyi Shen, Yuan Gao, Xiaoke Jiang, Xingyu Chen, Zhuheng Song, Yuhong Zhang, Hongjie Huang, Han Gao, Shilong Liu, Hao Zhang, Feng Li,

Kent Yu, and Lei Zhang. Dino-x: A unified vision model for open-world object detection and understanding, 2024. 2

[33] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *International Conference on Computer Vision (ICCV) 2021*, 2021. 3

[34] Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15768–15780, 2025. 1

[35] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 2, 3

[36] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024. 6, 5

[37] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. 2020. 3

[38] Xingrui Wang, Wufei Ma, Zhuowan Li, Adam Kortylewski, and Alan L Yuille. 3d-aware visual question answering about parts, poses and occlusions. *Advances in Neural Information Processing Systems*, 36:58717–58735, 2023. 1, 2, 3, 4, 7

[39] Xingrui Wang, Wufei Ma, Tiezheng Zhang, Celso M de Melo, Jieneng Chen, and Alan Yuille. Spatial457: A diagnostic benchmark for 6d spatial reasoning of large multimodal models. *CVPR*, 2025. 1, 3

[40] Zehan Wang, Ziang Zhang, Tianyu Pang, Chao Du, Hengshuang Zhao, and Zhou Zhao. Orient anything: Learning robust object orientation estimation from rendering 3d models. *arXiv:2412.18605*, 2024. 2, 3, 5, 8

[41] Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. *arXiv preprint arXiv:2506.09965*, 2025. 2, 8

[42] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE winter conference on applications of computer vision*, pages 75–82. IEEE, 2014. 2, 3

[43] Runsen Xu, Weiyao Wang, Hao Tang, Xingyu Chen, Xiaodong Wang, Fu-Jen Chu, Dahua Lin, Matt Feiszli, and Kevin J. Liang. Multi-spatialmllm: Multi-frame spatial understanding with multi-modal large language models. *arXiv preprint arXiv:2505.17015*, 2025. 1

[44] Jihan Yang, Shusheng Yang, Anjali Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in Space: How Multimodal Large Language Models See, Remember and Recall Spaces. *arXiv preprint arXiv:2412.14171*, 2024. 1

[45] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10371–10381, 2024. 2, 3

[46] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 3, 5

[47] Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. Mmsi-bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*, 2025. 1

[48] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B Tenenbaum. Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. In *Advances in Neural Information Processing Systems (NIPS)*, 2018. 2, 7, 1

[49] Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. In *Structural Priors for Vision Workshop at ICCV'25*, 2025. 1

[50] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9043–9053, 2023. 2, 3, 8

[51] Wenqi Zhang, Mengna Wang, Gangao Liu, Xu Huixin, Yiwei Jiang, Yongliang Shen, Guiyang Hou, Zhe Zheng, Hang Zhang, Xin Li, et al. Embodied-reasoner: Synergizing visual search, reasoning, and action for embodied interactive tasks. *arXiv preprint arXiv:2503.21696*, 2025. 1

[52] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1702–1713, 2025. 1

[53] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[54] Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv preprint arXiv:2503.14489*, 2025. 3