Human-Centric Architectural Priors Unlock Sample-Efficient Latent Spatial Reasoning

Brandon Ong [1] **Kingston Kuan** [1]

Abstract

As language models increasingly support robotic control, spatial understanding has become an area of growing interest across multimodal AI and robotics. Simulated data is a scalable source of spatial supervision for multimodal language models across images and video: SIMS-V showed that spatial question-answering (QA) synthesized from procedural ProcTHOR houses improves VSI-Bench. However, Cambrian-S's data ablation across spatial sources revealed that ProcTHOR is the least effective 3D-annotated source, suggesting that human-centric architectural priors may unlock latent spatial reasoning capabilities that procedural layouts miss. In this study, we ask whether generating embodied question-answering (EQA) data from human-crafted architectural layouts yields a stronger learning signal than procedural generation. We convert 29 GRScenes-100 apartments (Isaac Sim USD digital twins of real homes) into SIMS-V format, render Stretch-robot tours in Isaac Sim 5.1, and introduce **RealHomes5K**, a dataset of 2,378 unique VSI-style questions (4,756 multiple-choice and open-ended conversations). Fine-tuning `LLaVA-Video-7B-Qwen2` on **RealHomes5K** raises VSI-Bench from 35.0 to 43.8, outperforming a matched SIMS-VSI control (38.5) by +5.3 points, with a +5.5 gain on VSI-Bench-Debiased (37.2 vs. 31.7). A blind control trained on **RealHomes5K** questions with black video (39.8 VSI-Bench; 33.8 Debiased) shows that while roughly half of the +8.8 gain over the base model stems from question and answer text priors (+4.8), +4.0 points (+3.3 Debiased) require the rendered video. Notably, our +8.8 gain from 4,756 conversations matches the +8.4 gain SIMS-V reports from 25K examples, demonstrating substantial sample efficiency.

[1]Hebbian Robotics. Correspondence to: Hebbian Robotics Research Team <research@hebbianrobotics.com>.

Preprint. November 2, 2026.

Table 1. **Sample efficiency of simulated spatial video instruction tuning on VSI-Bench.** Prior procedural ProcTHOR datasets in Cambrian-S (Yang et al., 2025b) (VSI-590K single-source ablation) and SIMS-V (Brown et al., 2025) require 20K–25K samples for $+7.9$ to $+8.4$ points of gain over their respective base models. With only **4,756** samples ($4.2\times$–$5.3\times$ fewer), **RealHomes5K (Ours)** achieves a **+8.8** gain ($+5.3$ over matched SIMS-VSI).

Dataset / Setup	Scene Source	# Samples	VSI-Bench	Δ Base
Prior procedural simulation benchmarks (ProcTHOR / AI2-THOR):				
Cambrian-S (ProcTHOR) (Yang et al., 2025b)	Procedural houses	20,092	28.5 → 36.4	+7.9
SIMS-V (3Q mix) (Brown et al., 2025)	Procedural houses	25,000	36.0 → 44.4	+8.4
`LLaVA-Video-7B` *(matched 4,756-sample budget):*				
+ SIMS-VSI (5K)	Procedural houses	4,756	35.0 → 38.5	+3.5
+ RealHomes5K (w/o Video)	Black video (text prior)	4,756	35.0 → 39.8	+4.8
+ RealHomes5K (Ours)	**Real-world layouts**	**4,756**	**35.0 → 43.8**	**+8.8**

1. Introduction

Visual-spatial intelligence—the ability to perceive, remember, and reason about 3D space from egocentric video of indoor environments—remains a major bottleneck for multimodal large language models (MLLMs). On VSI-Bench (Yang et al., 2025a), which evaluates 5,130 spatial reasoning questions across 288 real indoor scan videos from ScanNet (Dai et al., 2017), ScanNet++ (Yeshwanth et al., 2023), and ARKitScenes (Baruch et al., 2021), open-weight video MLLMs perform poorly out of the box: `LLaVA-Video-7B` (Zhang et al., 2025c) scores 35.6 in the original benchmark evaluation (Yang et al., 2025a) and 36.0 in subsequent evaluations (Brown et al., 2025). Performance drops further on VSI-Bench-Debiased (Brown et al., 2026), which filters out questions solvable via non-visual language shortcuts to leave 2,362 strictly visual questions (30.7 for `LLaVA-Video-7B`).

Bridging this spatial reasoning gap requires 3D-grounded instruction tuning data, which currently comes from three main sources: (1) annotated real 3D scans (Dai et al., 2017; Yeshwanth et al., 2023; Baruch et al., 2021), which provide strong supervision in datasets like Cambrian-S's VSI-590K (Yang et al., 2025b) (reaching 67.5 on VSI-Bench with a 7B model) and ViCA-322K (Feng, 2025), but are limited in scene count and share the exact scan domains of VSI-Bench's test set; (2) pseudo-annotated web or robot videos (Chen et al., 2024; Yang et al., 2025b), which suffer from noisy 3D reconstruction; and (3) simulation. SIMS-V (Brown et al., 2025) demonstrated that generating spatial video QA from procedurally generated ProcTHOR

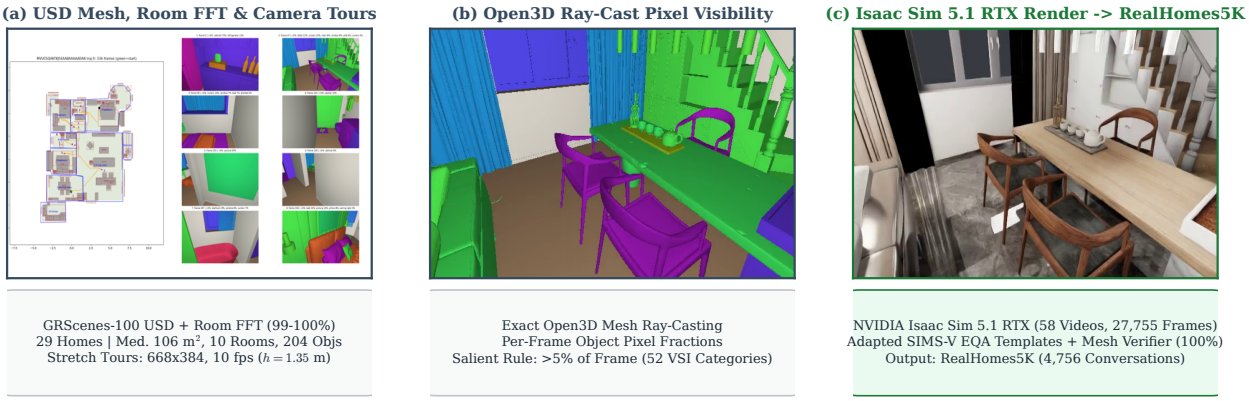


Figure 1. **Overview of the RealHomes5K data generation pipeline. (a)** Starting from 29 real-world apartment digital twins in GRScenes-100 (Wang et al., 2024) (navigation USD variant), we extract ground-truth 3D object bounding boxes directly from USD meshes, fit room polygons via 2D FFT cross-correlation against the house footprint (99–100% object containment), map categories and left-handed coordinates to SIMS-V (Brown et al., 2025), and plan Stretch-robot camera tours on occupancy maps. **(b)** Exact per-frame pixel visibility is computed via Open3D ray casting (Zhou et al., 2018) (retaining objects with $> 5\%$ frame area). **(c)** We render 58 tours (27,755 frames) in Isaac Sim 5.1 RTX (NVIDIA, 2025) and adapt SIMS-V's data generation pipeline and templates for embodied QA to produce **RealHomes5K** (2,378 unique questions / 4,756 conversations, 100% geometrically verified against the USD meshes).

houses (Deitke et al., 2022) rendered in AI2-THOR (Kolve et al., 2017) with a simulated Stretch robot camera (Kemp et al., 2022) raises `LLaVA-Video-7B` from 36.0 to 44.4 on VSI-Bench (30.7 to 38.4 on VSI-Bench-Debiased) using 25K examples.

Yet procedural houses differ substantially from real human dwellings: their room connectivity, floor-plan geometry, furniture density, and object co-occurrence statistics are governed by a procedural sampler rather than real architectural layouts. Indeed, in Cambrian-S's single-source ablation of VSI-590K (Yang et al., 2025b), ProcTHOR improves the 28.5 baseline to only 36.4 (+7.9 from 20,092 samples; Table 1)—the weakest of all sources with exact 3D annotations, trailing out-of-benchmark real scans and artist scenes (ADT 41.0, S3DIS 41.6, Hypersim 45.6 (Roberts et al., 2021)) as well as in-domain 3D scans (ARKitScenes 51.0, ScanNet / ScanNet++ V2 56.3, full VSI-590K mix 63.2), and beating only pseudo-annotated video (OXE+AgiBot 30.6, YouTube 32.5).

A promising middle route between procedural generation and in-domain real scans is **simulation-ready digital twins of real homes**: artist-built Universal Scene Description (USD) scenes that faithfully follow real-world apartment floor plans and furniture arrangements. GRScenes-100, released as part of the GRUtopia project (Wang et al., 2024), provides such scenes for NVIDIA Isaac Sim (NVIDIA, 2025). These scenes provide simulator-exact 3D ground truth on realistic residential layouts while remaining completely disjoint from VSI-Bench's scan sources.

In this paper, we investigate two concrete empirical questions: *With the embodied QA generation pipeline, training recipe, and dataset size held strictly fixed, does spatial QA generated from real-world house layouts improve VSI-Bench more than QA from procedural houses? And how much of that improvement requires the rendered video, versus coming from question and answer text priors?*

To answer these questions, we introduce **RealHomes5K**: we convert 29 real-world apartments from GRScenes-100 into SIMS-V's scene format, render 58 robot tours in Isaac Sim 5.1 RTX, and adapt SIMS-V's data generation pipeline and templates for embodied QA to synthesize 2,378 unique questions (4,756 conversations across multiple-choice and open-ended formats; Figure 1). We then run a controlled study on `LLaVA-Video-7B-Qwen2` comparing four model variants: the un-finetuned **Base** model, **+ RealHomes5K (Ours)** trained on our 4,756 real-layout conversations, **+ SIMS-VSI (5K)** trained on a strictly matched SIMS-VSI subset with identical task and format quotas (4,756 conversations across 31 ProcTHOR houses), and **+ RealHomes5K (w/o Video)**, a blind control trained on our exact question-answer pairs paired with black video frames.

Our experiments establish three main findings, each accompanied by important methodological caveats:

- **Real-layout simulation outperforms matched procedural simulation:** Fine-tuning on 4,756 conversations from **RealHomes5K (Ours)** reaches **43.8** on VSI-Bench and **37.2** on VSI-Bench-Debiased, beating the matched **+ SIMS-VSI (5K)** baseline (38.5 / 31.7) by **+5.3** points on VSI-Bench and **+5.5** points on VSI-Bench-Debiased. Because **RealHomes5K** and **SIMS-VSI (5K)** differ simultaneously in house layout, 3D assets, renderer (Isaac Sim RTX vs. Unity), and apartment size, this difference measures the combined real-layout simulation package rather than layout geometry in isolation.

- **Decomposing text priors versus visual learning via a blind control:** Training on our exact QA pairs with black video frames (**+ RealHomes5K (w/o Video)**) reaches 39.8 on VSI-Bench (33.8 Debiased)—a gain of **+4.8** over the **Base** model (35.0) from answer format and numeric range priors alone. At the same time, comparing **+ RealHomes5K (Ours)** to **+ RealHomes5K (w/o Video)** proves that **+4.0** points on VSI-Bench (**+3.3** on Debiased) require the rendered real-layout video. Decomposing our headline $+5.3$ advantage over **+ SIMS-VSI (5K)** shows that only $+1.2$ points come from text priors alone (**+ RealHomes5K (w/o Video)** vs. **+ SIMS-VSI (5K)**), while $+4.0$ points require the rendered video.
- **Sample efficiency and task-level trade-offs:** As highlighted in Table 1, our $+8.8$ gain from just 4,756 conversations in **RealHomes5K** (29 apartments) exceeds Cambrian-S's $+7.9$ ProcTHOR gain from 20,092 samples (Yang et al., 2025b) ($4.2\times$ fewer samples) and matches SIMS-V's $+8.4$ gain from 25,000 samples (Brown et al., 2025) ($5.3\times$ fewer samples, though SIMS-V's 25K run uses a 3-question task mix and a different step budget, making the comparison indicative rather than controlled). Task-level analysis reveals that whole-apartment room-size questions explain a large fraction of **+ SIMS-VSI (5K)**'s deficit (where it collapses to 4.4 due to over-predicting on large ProcTHOR houses), but even excluding room size, **+ RealHomes5K (Ours)** outperforms **+ SIMS-VSI (5K)** by **+2.5** on VSI-Bench (**+2.6** Debiased), and that remaining VSI-Bench advantage is 100% video-dependent (-0.1 without video).

Contributions.

1. **RealHomes5K** and a conversion pipeline that adapts SIMS-V's data generation pipeline and templates for embodied QA to real-world USD house layouts (GRScenes-100), with independent geometric answer verification directly against the 3D USD meshes (100% agreement across all 2,378 questions).
2. A matched ablation design for synthetic spatial data comparing **+ RealHomes5K (Ours)**, a matched procedural-house baseline (**+ SIMS-VSI (5K)**), and a blind text-prior control (**+ RealHomes5K (w/o Video)**), evaluated across overall benchmarks and pre-specified task subsets.
3. Empirical evidence that real-layout QA beats matched procedural-house QA on VSI-Bench and VSI-Bench-Debiased, and a decomposition of the gain into a video-dependent part and a text-prior part.

2. Related Work

Evaluating Visual-Spatial Intelligence and Non-Visual Shortcuts. VSI-Bench (Yang et al., 2025a) standardized the evaluation of egocentric 3D spatial reasoning in video MLLMs across eight tasks spanning configurational reasoning (object count, relative distance, relative direction across easy/medium/hard splits, route planning), metric measurement (object size, absolute distance, room size, evaluated via mean relative accuracy [MRA]), and spatiotemporal reasoning (appearance order), using real indoor scans from ScanNet (Dai et al., 2017), ScanNet++ (Yeshwanth et al., 2023), and ARKitScenes (Baruch et al., 2021) executed through `lmms-eval` (Zhang et al., 2025b). Recently, Brown et al. (2026) demonstrated that benchmark designers should "train on the test set" without visual inputs to expose exploitable language priors, releasing VSI-Bench-Debiased (2,362 questions) after filtering shortcut-vulnerable items. Our study complements Brown et al. (2026) from the training side: we evaluate all models on both VSI-Bench and VSI-Bench-Debiased, and we introduce a blind training control (**+ RealHomes5K (w/o Video)**) to measure how much synthetic spatial instruction tuning improves test scores via non-visual answer priors alone.

Spatial Instruction Tuning from Real Scans and Architecture Modifications. Built upon general-purpose visual instruction tuning recipes such as LLaVA (Liu et al., 2023; 2024), LLaVA-OneVision (Li et al., 2024), and LLaVA-Video (Zhang et al., 2025c) (combining Qwen2 (Yang et al., 2024) and SigLIP (Zhai et al., 2023)), recent works have scaled spatial supervision from 2D images and 3D scans. SpatialVLM (Chen et al., 2024) and SpatialRGPT (Cheng et al., 2024) lift 2D web images into pseudo-3D scenes to synthesize spatial QA. In the video domain, Cambrian-S (Yang et al., 2025b) curates VSI-590K across real scans, simulation, and pseudo-labelled video; SPAR (Zhang et al., 2025a) builds a 2D spatial QA generation pipeline on 3D ground-truth scan data (SPAR-7M and SPAR-Bench); and ViCA (Feng, 2025) generates 322K QA pairs directly from real indoor scan videos (ARKitScenes, ScanNet, ScanNet++) to reach strong performance across all eight VSI-Bench tasks. Other 2025 approaches modify the training objective or model architecture: SpaceR (Ouyang et al., 2025) combines 91K verifiable video spatial QA pairs with 60K general samples (SpaceR-151k) under reinforcement learning with verifiable rewards (RLVR); Spatial-MLLM (Wu et al., 2025) augments the 2D semantic encoder with a feed-forward 3D visual-geometry encoder; VLM-3R (Fan et al., 2025) injects implicit 3D tokens from monocular frames via 3D-reconstructive instruction tuning; and Multi-SpatialMLLM (Xu et al., 2026) scales multi-frame training to over 27M samples from 3D/4D scenes. Our work is orthogonal to architectural modifications and real-scan scaling:

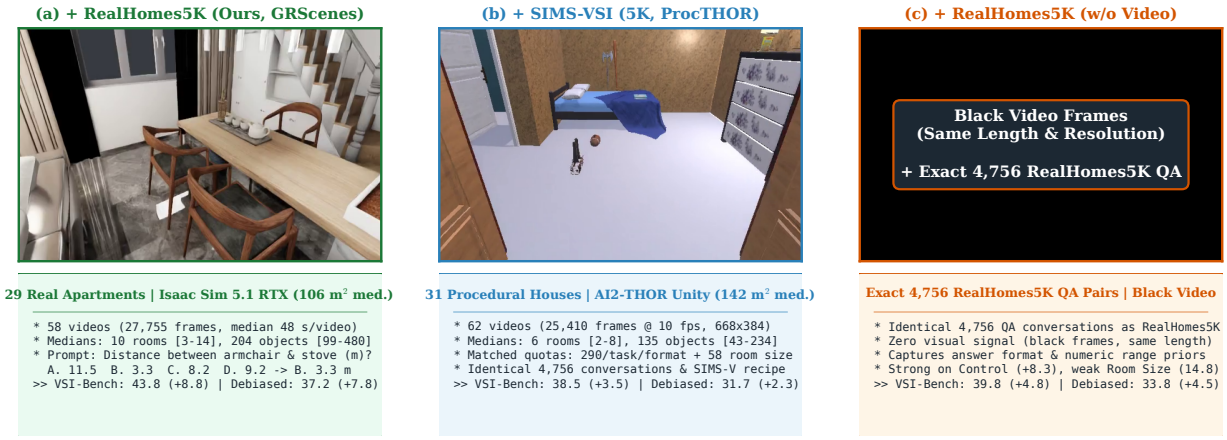


Figure 2. **Visual and experimental comparison across + RealHomes5K (Ours), + SIMS-VSI (5K) procedural houses, and + RealHomes5K (w/o Video).** Isaac Sim 5.1 RTX camera tour frames from our converted GRScenes-100 apartments are shown alongside procedural ProcTHOR/AI2-THOR environments (Deitke et al., 2022; Brown et al., 2025) and black-frame blind training inputs.

we hold the base model (`LLaVA-Video-7B-Qwen2`), training recipe, and embodied QA templates strictly fixed, and vary only the underlying synthetic scene source.

Simulated Environments and Sim-to-Real Spatial Transfer. Using randomized or structured simulation to transfer perception models to the real world is a foundational paradigm in robotics (Tobin et al., 2017). For embodied AI and indoor perception, major 3D scene datasets include real-world reconstructions like Habitat-Matterport 3D (HM3D) (Savva et al., 2019; Ramakrishnan et al., 2021), synthetic professional interiors like Structured3D (Zheng et al., 2020) and Hypersim (Roberts et al., 2021), and procedural environments like AI2-THOR (Kolve et al., 2017) and ProcTHOR (Deitke et al., 2022) populated with Objaverse 3D assets (Deitke et al., 2023). For MLLMs, SAT (Ray et al., 2025) provided the most direct early evidence that static and dynamic spatial QA generated in procedural ProcTHOR/AI2-THOR scenes transfers to real-image spatial benchmarks. SIMS-V (Brown et al., 2025) extended this paradigm to video MLLMs by simulating Stretch robot tours (Kemp et al., 2022) in ProcTHOR houses to generate SIMS-VSI. Rather than relying on procedurally sampled floor plans, we leverage GRScenes-100 from GRUtopia (Wang et al., 2024), which provides artist-authored Isaac Sim (NVIDIA, 2025) USD digital twins of real apartments, allowing us to test whether realistic residential layouts and rendering yield superior spatial video instruction tuning data at a fixed question budget.

3. Method: GRScenes to SIMS-V Spatial QA

To test the effect of scene source without confounding changes to question generation logic, we build a deterministic conversion pipeline (Figure 1) that adapts SIMS-V's data generation pipeline and templates for embodied QA to real-world USD apartments from GRScenes-100.

3.1. Source Scenes

We source indoor environments from the GRScenes-100 home subset (Wang et al., 2024), which consists of Isaac Sim USD scenes modeled in a Z-up, centimeter-scale, right-handed coordinate system under a CC BY-NC-SA 4.0 license. Thirty home scenes are distributed with full benchmark metadata (room polygons, navigation occupancy maps, and semantic object dictionaries). We use the navigation variant of each scene (`start_result_navigation.usd`), in which interior doors are removed to permit continuous multi-room robot traversal. We apply an automated geometric alignment check across all 30 benchmark home scenes and retain **29 scenes**; one scene (`MV7J6NIKTKJZ2AABAAAAAEA8`) fails our alignment threshold (only 82% of objects fall in their labelled room and 76% occupancy fit) and is excluded.

3.2. Conversion and Verification Pipeline

Our conversion pipeline executes nine sequential stages:

1. **Object boxes from USD meshes:** We compute 3D bounding boxes directly from the live USD meshes (`UsdGeom.BBoxCache` at the top-most semantically labelled prim) rather than GRScenes' `object_dict.json`. Because the navigation USD drops doors and some tableware and moves some objects by up to 0.52 m, the meshes are the ground truth that matches the render.
2. **Room polygons:** GRScenes' `annotation.json` stores room polygons in an undocumented per-scene image frame. We fit the pixel-to-world transform per scene by 2D FFT cross-correlation of the room union

Table 2. **Dataset statistics and Step-0 ground-truth numeric answer distributions.** *Left (a):* Scene, video, and QA statistics for **RealHomes5K (Ours)** vs. the matched **SIMS-VSI (5K)** subset. *Right (b):* Step-0 ground-truth answer medians and interquartile ranges [IQR] before training across **RealHomes5K**, **SIMS-VSI (5K)**, and the VSI-Bench test set.

(a) Scene, Video, and QA Statistics

Property	RealHomes5K (Ours)	SIMS-VSI (5K)
Houses	29 real apartments	31 ProcTHOR houses
Floor area (m^2, med. [min–max])	106 [37–186]	142 [36–517]
Rooms / house (med. [min–max])	10 [3–14]	6 [2–8]
Objects / house (med. [min–max])	204 [99–480]	135 [43–234]
Videos / Frames (10 fps, 668×384)	58 / 27,755 (med. 48 s)	62 / 25,410
Unique Qs / Conversations (MC+OE)	2,378 / 4,756	2,378 / 4,756
Renderer	Isaac Sim 5.1 RTX	AI2-THOR (Unity)
3D assets	GRScenes artist models	ProcTHOR / Objaverse

(b) Step-0 Ground-Truth Answer Distributions (Median [IQR])

Task	RealHomes5K	SIMS-VSI (5K)	VSI-Bench Test
Abs. distance (m)	3.1 [1.5–5.2]	4.2 [2.6–7.4]	1.8 [1.1–2.9]
Object size (cm)	100 [60–185]	140 [105–200]	96 [59–169]
Object count	1 [1–3]	1 [1–2]	2 [2–4]
Room size (m^2)	106 [102–119]	142 [103–176]	17.5 [11–27]

Note: Distance and object-size priors in RealHomes5K lie closer to VSI-Bench than SIMS-VSI (5K)'s, motivating our blind control (+ RealHomes5K w/o Video). Both synthetic sets ask room size for the whole visited apartment, whereas VSI-Bench asks it for single scanned rooms.

against the house footprint (8 axis swaps/flips plus a shift), accepting a fit when 99–100% of objects lie inside their labelled room.

3. **Categories:** Mapped onto SIMS-V's 52 VSI object categories (e.g., `couch` → `sofa`, `washingmachine` → `washer`).

4. **Coordinates:** GRScenes (x, y, z) maps to SIMS-V (x, z, y). SIMS-V is Y-up and left-handed, and this single axis swap also flips handedness ($\det = -1$), so left/right answers stay correct (unit-tested; see Appendix A.2).

5. **Camera tours:** We replicate SIMS-V's Stretch robot camera (Kemp et al., 2022): 0.2 m steps, 30° turns, a 12-frame scan in each room, pitch $27° \pm 2°$, vertical FOV $59° \pm 1°$, 668×384 pixels, 10 fps, with an assumed camera height of 1.35 m. Paths are planned on GRScenes' occupancy maps (two tours per scene). Rooms that no tour reaches (35 of 274) are excluded from room-count and room-size answers.

6. **Visibility:** Exact ray casting with Open3D (Zhou et al., 2018) gives the pixel fraction of every object in every frame. An object is salient if it covers $> 5\%$ of at least one frame (SIMS-V's rule).

7. **Rendering:** Isaac Sim 5.1 RTX renderer (NVIDIA, 2025), 8 subframes, 668×384, 10 fps, producing 58 videos (27,755 frames, median 48 s/video) in $\sim$ 1.5 GPU-hours on an RTX A6000.

8. **QA generation:** We adapt SIMS-V's data generation pipeline and templates for embodied QA (5 questions per task type per video), formatting each question in both multiple-choice (MC) and open-ended (OE) modes (Figure 2).

9. **Verification:** An independent geometric verifier rederives every answer directly from the USD bounding boxes and room polygons: **2,378 of 2,378 agree (100%)**.

3.3. Dataset Statistics and Step-0 Answer Distributions

Table 2a compares **RealHomes5K (Ours)** against the matched **SIMS-VSI (5K)** subset. Our 29 apartments contain 5,965 objects across 70 GRScenes categories (most common: `book`, `cabinet`, `pillow`, `curtain`, `window`), with more rooms per house (median 10 vs. 6) and denser objects (204 vs. 135) than ProcTHOR houses. Both sets contain 2,378 unique questions (4,756 MC+OE conversations): 290 per task per format for absolute distance, relative distance, appearance order, object count, object size, and relative direction easy/medium/hard, plus 58 per format for room size (one per video). Route planning is not generated by SIMS-V and is not trained.

Auditing the ground-truth numeric answers before training (Table 2b and Figure 3) reveals three key properties: (1) our distance (3.1 m vs. 4.2 m; VSI-Bench 1.8 m) and object size (100 cm vs. 140 cm; VSI-Bench 96 cm) distributions are closer to VSI-Bench than SIMS-VSI's are, motivating our blind control (**+ RealHomes5K w/o Video**); (2) both synthetic sets ask room size for the **whole apartment** (106 m^2 and 142 m^2) whereas VSI-Bench asks it for single scanned rooms (17.5 m^2), making room size a weak test of layout realism for both; and (3) MC answer letters (25–38% per letter) and relative-direction labels are balanced, preventing option-letter shortcuts, while object counts are concentrated at 1 in both synthetic sets (vs. median 2 in VSI-Bench).

4. Experimental Setup

Model and Training Recipe. Following Appendix A of SIMS-V (Brown et al., 2025), we fine-tune `LLaVA-Video-7B-Qwen2` (Zhang et al., 2025c) (`Qwen2-7B` LLM (Yang et al., 2024), `SigLIP-so400m` vision encoder (Zhai et al., 2023)) in the `LLaVA-NeXT` codebase (Liu et al., 2023; 2024; Li et al., 2024). We unfreeze the LLM, projector, and vision tower with learning rate 1×10^{-6}, AdamW, weight decay 0, cosine schedule, 3% warmup, global batch size 32 (8 GPUs $\times 1 \times 4$ grad. accum.), `bf16`, max sequence length 32K, DeepSpeed ZeRO-3 (Rajbhandari et al., 2020), FlashAttention-2 (Dao, 2024), and gradient checkpointing. Videos use up to 64 uniformly

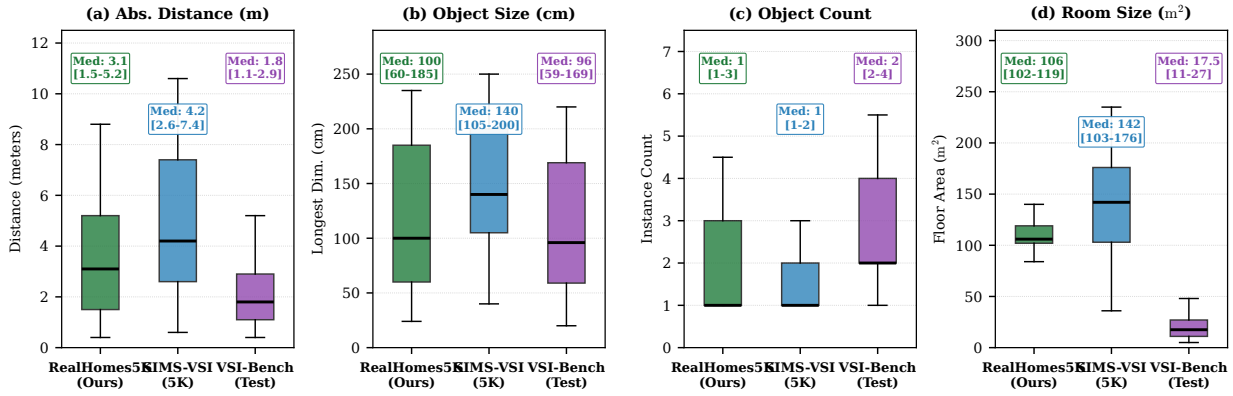


Figure 3. **Ground-truth numeric answer distributions (median and interquartile range) across RealHomes5K (Ours), SIMS-VSI (5K), and the VSI-Bench test set.** Our real-layout apartments have object sizes and inter-object distances much closer to real VSI-Bench scans than ProcTHOR houses do, motivating our blind control (**+ RealHomes5K w/o Video**) to isolate answer-prior effects.

sampled frames with `LLaVA-Video`'s time instruction and spatial pool stride 2. Every fine-tuned variant runs for 3 epochs (444 steps) on spatial training data only (1 h 57 min on 8 × NVIDIA H100 80GB).

Model Variants and Matching. We compare four configurations (Table 3): **Base** (untrained `LLaVA-Video-7B`), **+ RealHomes5K (Ours)** (4,756 conversations from 29 GRScenes apartments rendered in Isaac Sim RTX), **+ SIMS-VSI (5K)** (a SIMS-VSI subset matched to **RealHomes5K** with identical counts per task and format: 290/task/format + 58 room size, drawn by shuffling SIMS-VSI houses until quotas were met across 31 houses / 62 videos), and **+ RealHomes5K (w/o Video)** (**RealHomes5K**'s exact 4,756 questions and answers with every video replaced by black frames of the same length and resolution). All fine-tuned variants use the SIMS-VSI release wrapper (Appendix A.1), and we removed an extra multiple-instance sentence from absolute-distance prompts in **RealHomes5K** that is absent from the SIMS-VSI release and VSI-Bench.

Evaluation and Pre-Specified Subsets. We evaluate with `lmms-eval` v0.7.3 (Zhang et al., 2025b) (model `llava_vid`) on `vsibench` (5,130 questions) and `vsibench_debiased` (2,362 questions) using 64 frames, greedy decoding, and batch size 1 (∼ 70 min/model on 8 × H100). Alongside VSI-Bench and Debiased overall scores, we evaluate two pre-specified task subsets: *Layout-sensitive* (absolute distance, room size, relative distance, relative direction) and *Control* (object count, object size, appearance order), with untrained route planning outside both. The **Base** model scores 35.0, within 1 point of published 35.6 (Yang et al., 2025a) and 36.0 (Brown et al., 2025) baselines.

Table 3. **Main results on VSI-Bench and VSI-Bench-Debiased** with sample counts (top), alongside reference points from the literature (bottom). All fine-tuned variants run for 444 steps (1h 57m on 8 × H100; mean training loss: 0.079 for **+ RealHomes5K (Ours)** and **+ SIMS-VSI (5K)**, 0.084 for **+ RealHomes5K (w/o Video)**). SIMS-V's 25K run uses its "3Q" task mix and constant steps, so the size comparison is indicative rather than controlled.

Model / Setup	Training Data	# Samples	VSI-Bench	Debiased
Controlled study on `LLaVA-Video-7B` (444 steps on 8 × H100):				
Base	None (zero-shot)	0	35.0	29.4
+ SIMS-VSI (5K)	SIMS-VSI subset (matched)	4,756	38.5	31.7
+ RealHomes5K (w/o Video)	RealHomes5K QA, black video (blind)	4,756	39.8	33.8
+ RealHomes5K (Ours)	**RealHomes5K QA, real-layout video**	**4,756**	**43.8**	**37.2**
Literature reference points (same base model, different data sizes / recipes):				
`LLaVA-Video-7B` (Brown et al., 2025)	None (base)	0	36.0	30.7
+ SIMS-VSI (25K) (Brown et al., 2025)	Procedural sim. (3Q mix)	25,000	44.4	38.4
`LLaVA-Video-72B` (Brown et al., 2025)	None (10× params)	0	41.2	—
`Cambrian-S-7B` (Yang et al., 2025b)	VSI-590K + full mix	590,667	67.5	59.9

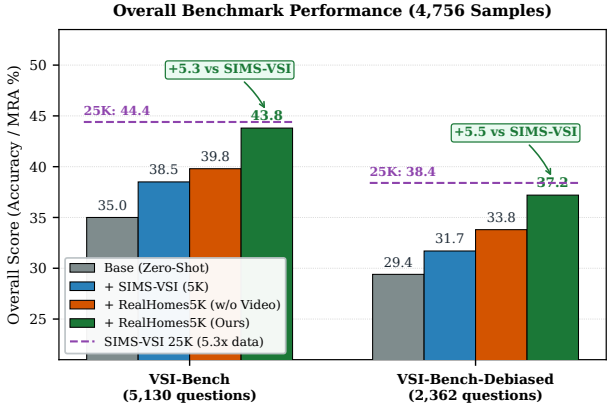


Figure 4. **Overall VSI-Bench and VSI-Bench-Debiased scores** across **Base**, **+ SIMS-VSI (5K)**, **+ RealHomes5K (w/o Video)**, and **+ RealHomes5K (Ours)**. Dashed lines show SIMS-V's 25K-example scores (Brown et al., 2025).

5. Results

Table 3 and Figure 4 summarize overall performance and training statistics, Table 4 gives per-task scores on VSI-Bench and VSI-Bench-Debiased, and Table 5 reports pairwise differences across overall benchmarks and pre-specified subsets.

Table 4. **Per-task accuracy / MRA scores on VSI-Bench (top, 5,130 questions) and VSI-Bench-Debiased (Brown et al., 2026) (bottom, 2,362 questions).** Relative direction averages easy ($n = 217$), medium ($n = 378$), and hard ($n = 373$) splits first; on VSI-Bench easy/medium/hard, **Base** scores 48.8 / 43.1 / 35.4, **+ SIMS-VSI (5K)** 48.8 / 46.0 / 38.1, **+ RealHomes5K (w/o Video)** 48.4 / 44.2 / 39.7, and **+ RealHomes5K (Ours)** 50.2 / 44.7 / 39.7. Route planning is not trained.

Benchmark / Model	Overall	Count	Abs. Dist	Obj. Size	Room Size	Rel. Dist	Rel. Dir	Route	Appear. Order
(a) VSI-Bench *(full benchmark, $n = 5,130$ questions; task counts: 565 | 834 | 953 | 288 | 710 | 968 | 194 | 618)*									
Base	35.0	**53.7**	9.8	47.7	18.4	44.9	42.5	**37.1**	25.9
+ SIMS-VSI (5K)	38.5	29.3	**29.3**	54.1	4.4	50.3	44.3	36.1	60.2
+ RealHomes5K (w/o Video)	39.8	44.0	25.5	52.2	14.8	47.5	44.1	34.0	56.0
+ RealHomes5K (Ours)	**43.8**	38.6	28.1	**55.6**	**29.3**	**51.0**	**44.9**	**37.1**	**65.7**
(b) VSI-Bench-Debiased *(shortcut-filtered benchmark, $n = 2,362$ questions; task counts: 251 | 434 | 353 | 200 | 310 | 382 | 114 | 318)*									
Base	29.4	**38.0**	13.8	34.5	14.2	44.8	37.3	**21.9**	30.5
+ SIMS-VSI (5K)	31.7	19.2	**26.8**	35.8	3.0	**49.7**	**41.4**	18.4	59.1
+ RealHomes5K (w/o Video)	33.8	37.4	22.0	38.1	11.3	49.0	37.5	17.5	57.9
+ RealHomes5K (Ours)	**37.2**	28.2	22.5	**45.4**	**29.0**	49.4	37.9	21.1	**64.2**

Table 5. **Pairwise score differences ($Y - X$ in percentage points) across overall benchmarks and task subsets on VSI-Bench (top) and VSI-Bench-Debiased (bottom).** Pre-specified subset scores on **VSI-Bench**: Layout-sensitive Base 28.9, + SIMS-VSI (5K) 32.1, + RealHomes5K (w/o Video) 32.9, + RealHomes5K (Ours) 38.3; Control Base 42.4, + SIMS-VSI (5K) 47.9, + RealHomes5K (w/o Video) 50.7, + RealHomes5K (Ours) 53.3; Overall excl. room size Base 37.4, + SIMS-VSI (5K) 43.4, + RealHomes5K (w/o Video) 43.3, + RealHomes5K (Ours) 45.9. On **Debiased**: Layout-sensitive Base 27.5, + SIMS-VSI (5K) 30.2, + RealHomes5K (w/o Video) 30.0, + RealHomes5K (Ours) 34.7; Control Base 34.3, + SIMS-VSI (5K) 38.1, + RealHomes5K (w/o Video) 44.4, + RealHomes5K (Ours) 45.9; Overall excl. room size Base 31.5, + SIMS-VSI (5K) 35.8, + RealHomes5K (w/o Video) 37.1, + RealHomes5K (Ours) 38.4.

Comparison (Y vs. X)	Interpretation	Overall	Layout-Sensitive	Control	Overall Excl. Room Size
(a) VSI-Bench Pairwise Differences *($n = 5,130$)*					
+ RealHomes5K (Ours) vs. Base	Total effect of our data	**+8.8**	**+9.4**	**+10.9**	**+8.5**
+ SIMS-VSI (5K) vs. Base	Effect of matched SIMS-VSI	**+3.5**	**+3.2**	**+5.4**	**+6.0**
+ RealHomes5K (w/o Video) vs. Base	Text/answer priors only	**+4.8**	**+4.1**	**+8.3**	**+6.0**
+ RealHomes5K (Ours) vs. + SIMS-VSI (5K)	**Real vs. procedural package**	**+5.3**	**+6.2**	**+5.4**	**+2.5**
+ RealHomes5K (Ours) vs. + RealHomes5K (w/o Video)	**Video-dependent part**	**+4.0**	**+5.4**	**+2.6**	**+2.5**
+ RealHomes5K (w/o Video) vs. + SIMS-VSI (5K)	Prior diff. (blind vs. SIMS-VSI)	+1.2	+0.9	+2.9	−0.1
(b) VSI-Bench-Debiased Pairwise Differences *($n = 2,362$)*					
+ RealHomes5K (Ours) vs. Base	Total effect of our data	**+7.8**	**+7.1**	**+11.6**	**+6.8**
+ SIMS-VSI (5K) vs. Base	Effect of matched SIMS-VSI	**+2.3**	**+2.7**	**+3.7**	**+4.2**
+ RealHomes5K (w/o Video) vs. Base	Text/answer priors only	**+4.5**	**+2.4**	**+10.1**	**+5.5**
+ RealHomes5K (Ours) vs. + SIMS-VSI (5K)	**Real vs. procedural package**	**+5.5**	**+4.4**	**+7.8**	**+2.6**
+ RealHomes5K (Ours) vs. + RealHomes5K (w/o Video)	**Video-dependent part**	**+3.3**	**+4.7**	+1.5	+1.3
+ RealHomes5K (w/o Video) vs. + SIMS-VSI (5K)	Prior diff. (blind vs. SIMS-VSI)	+2.2	−0.2	+6.4	+1.3

Fine-tuning on 4,756 conversations from **RealHomes5K (Ours)** raises VSI-Bench from **35.0** (**Base**) to **43.8** (**+8.8**) and VSI-Bench-Debiased from **29.4** to **37.2** (**+7.8**). Fine-tuning on the matched **+ SIMS-VSI (5K)** subset (4,756 conversations) reaches **38.5** (**+3.5**) on VSI-Bench and **31.7** (**+2.3**) on Debiased. Directly comparing **+ RealHomes5K (Ours)** to **+ SIMS-VSI (5K)** yields **+5.3** points on VSI-Bench and **+5.5** points on VSI-Bench-Debiased.

Meanwhile, the blind control trained on our questions with black video (**+ RealHomes5K w/o Video**) scores **39.8** on VSI-Bench (**+4.8** over **Base**) and **33.8** on Debiased (**+4.5** over **Base**). Comparing **+ RealHomes5K (Ours)** against **+ RealHomes5K (w/o Video)** shows that **+4.0** points on VSI-Bench and **+3.3** points on VSI-Bench-Debiased strictly require the rendered video. All fine-tuned variants ran for 444 steps (1 h 57 min on $8 \times$ H100) with mean training loss 0.079 for **+ RealHomes5K (Ours)** and **+ SIMS-VSI (5K)**, and 0.084 for **+ RealHomes5K (w/o Video)** (Table 3).

6. Analysis and Discussion

We structure our analysis around eight empirical findings from Tables 3 to 5.

1. Real-layout QA beats matched procedural-house QA. **+ RealHomes5K (Ours)** outperforms **+ SIMS-VSI (5K)** by **+5.3** on VSI-Bench and **+5.5** on the debiased set (Table 5). On VSI-Bench, the advantage is largest on the pre-specified layout-sensitive subset (**+6.2**, 38.3 vs. 32.1), while also improving the control subset by **+5.4** (53.3 vs. 47.9).

2. Half of the gain over Base is text priors. **+ RealHomes5K (w/o Video)** gains **+4.8** over **Base** on VSI-Bench (**+4.5** Debiased) with no visual input during training (Table 5). Most of this gain is in the control subset (**+8.3** VSI-Bench; **+10.1** Debiased), especially appearance order (25.9 to 56.0 on VSI-Bench; 30.5 to 57.9 on Debiased): training teaches expected answer formats and plausible numeric answer ranges. This is a general methodological caution for synthetic spatial data, not specific to our dataset.

3. About 4 points require the real-layout video. Comparing **+ RealHomes5K (Ours)** to **+ RealHomes5K (w/o Video)** yields **+4.0** on VSI-Bench and **+3.3** on VSI-Bench-Debiased, concentrated in the layout-sensitive subset (**+5.4** VSI-Bench; **+4.7** Debiased). On VSI-Bench (Table 4a), the largest video-dependent gains are on room size (14.8 to 29.3, $+14.5$), appearance order (56.0 to 65.7, $+9.7$), relative distance (47.5 to 51.0, $+3.5$), and object size (52.2 to 55.6, $+3.4$).

4. Decomposing the advantage over + SIMS-VSI (5K). By adding and subtracting the blind control (Table 5), **+ RealHomes5K (Ours)**'s $+5.3$ VSI-Bench advantage over **+ SIMS-VSI (5K)** decomposes into $+1.2$ from our question/answer distribution without video (**+ RealHomes5K w/o Video** vs. **+ SIMS-VSI 5K**) plus $+4.0$ from our rendered video (**+ RealHomes5K Ours** vs. **+ RealHomes5K w/o Video**). On the layout-sensitive subset, the blind-vs.-SIMS-VSI difference is $+0.9$ (Debiased -0.2). The text-prior part of our advantage is small; most of the advantage requires the video.

5. Room size explains much of the headline gap, but a video-dependent gap remains without it. **+ SIMS-VSI (5K)** collapses on room size (4.4 vs. **Base** 18.4 on VSI-Bench; 3.0 vs. 14.2 on Debiased) because SIMS-VSI room answers are large (median 142 m^2 vs. VSI-Bench 17.5 m^2) and the model learned to over-predict. Excluding room size (Table 5, rightmost column), **+ RealHomes5K (Ours)** still beats **+ SIMS-VSI (5K)** by **+2.5** on VSI-Bench (45.9 vs. 43.4), and that remaining gap is entirely video-dependent: **+ RealHomes5K (w/o Video)** vs. **+ SIMS-VSI (5K)** is -0.1, while **+ RealHomes5K (Ours)** vs. **+ RealHomes5K (w/o Video)** is **+2.5**. On the debiased set, the excluding-room-size gaps are **+2.6** for **+ RealHomes5K (Ours)** vs. **+ SIMS-VSI (5K)** (38.4 vs. 35.8) and $+1.3$ over the blind control.

6. Counting regresses across all fine-tuned variants. Every fine-tuned variant loses accuracy on object count relative to the **Base** model (VSI-Bench: **Base** 53.7 vs. **+ RealHomes5K (Ours)** 38.6, **+ SIMS-VSI (5K)** 29.3, **+ RealHomes5K (w/o Video)** 44.0; Debiased: **Base** 38.0 vs. 28.2, 19.2, 37.4). Both synthetic sets have count answers concentrated at 1 (median 1) while VSI-Bench's median is 2 (Table 2b), so fine-tuning pulls predicted counts downward. The blind variant counts best among the fine-tuned models, consistent with a prior effect.

7. Relative direction and route planning barely move. Relative direction moves little in any variant (42.5 in **Base** to at most 44.9 in **+ RealHomes5K (Ours)** on VSI-Bench; 37.3 in **Base** to 41.4 in **+ SIMS-VSI (5K)** and 37.9 in **+ RealHomes5K (Ours)** on Debiased), and untrained route planning stays near baseline (37.1 in **Base** and **+ RealHomes5K (Ours)** on VSI-Bench). Direction questions inherit SIMS-V's multiple-instance ambiguity: about 44–45% of SIMS-VSI direction questions and 40–60% of ours name a category with another visible instance in the video.

8. Sample efficiency. 4,756 real-layout conversations give a VSI-Bench gain ($+8.8$, 35.0 to 43.8) similar to SIMS-V's 25K-example run ($+8.4$, 36.0 to 44.4 (Brown et al., 2025)), whereas 4,756 SIMS-VSI conversations give $+3.5$. A controlled scaling curve (SIMS-VSI at 5K/10K/25K) is needed to turn this comparison into an exact data exchange rate.

7. Limitations and Future Work

While our experiments demonstrate strong sample efficiency from real-world apartment layouts, several limitations point toward natural directions for future work: (1) **Dataset and model scale:** Our current dataset spans 29 scenes (2,378 unique questions) and evaluates a single base architecture (`LLaVA-Video-7B-Qwen2`); scaling this pipeline to larger collections of human-centric 3D residential scenes and broader model families is an important next step. (2) **Licensing:** Because GRScenes-100 is released under CC BY-NC-SA 4.0, our derived dataset and checkpoints inherit non-commercial terms. (3) **General video retention:** Because fine-tuning was performed exclusively on spatial QA conversations, retention on general video understanding benchmarks was not measured and could be addressed by co-training with general instruction mixtures. (4) **Transfer to downstream robotics tasks:** Finally, VSI-Bench evaluates spatial understanding solely through Embodied Question Answering (EQA). An exciting direction for future work is evaluating whether the spatial representations learned from **RealHomes5K** transfer directly to closed-loop robotics tasks such as indoor navigation, mobile manipulation, and embodied action planning.

8. Conclusion

With the base model (`LLaVA-Video-7B-Qwen2`), training recipe, and embodied QA generation pipeline held fixed at 4,756 conversations, converting 29 real-world GRScenes-100 apartments into **RealHomes5K** improves VSI-Bench by **+5.3** points (43.8 vs. 38.5) and VSI-Bench-Debiased by **+5.5** points (37.2 vs. 31.7) over a matched **+ SIMS-VSI (5K)** ProcTHOR baseline. A blind black-video control (**+ RealHomes5K w/o Video**: 39.8 VSI-Bench, 33.8 Debiased) demonstrates that while roughly half of the $+8.8$ gain over the base model comes from question/answer priors alone, **+4.0** points require the rendered video. Together, these findings show that grounding synthetic visual supervision in human-centric 3D scenes is an effective and sample-efficient path for unlocking latent spatial reasoning in multimodal

language models.

References

Baruch, G., Chen, Z., Dehghan, A., Dimry, T., Feigin, Y., Fu, P., Gebauer, T., Joffe, B., Kurz, D., Schwartz, A., and Shulman, E. ARKitScenes: A diverse real-world dataset for 3D indoor scene understanding using mobile RGB-D data. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. arXiv:2111.08897.

Brown, E., Ray, A., Krishna, R., Girshick, R., Fergus, R., and Xie, S. SIMS-V: Simulated instruction-tuning for spatial video understanding. *arXiv preprint arXiv:2511.04668*, 2025.

Brown, E., Yang, J., Yang, S., Fergus, R., and Xie, S. Benchmark designers should "train on the test set" to expose exploitable non-visual shortcuts. In *Conference on Language Modeling (COLM)*, 2026. arXiv:2511.04655.

Chen, B., Xu, Z., Kirmani, S., Ichter, B., Driess, D., Florence, P., Sadigh, D., Guibas, L., and Xia, F. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14455–14465, 2024.

Cheng, A.-C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., and Liu, S. SpatialRGPT: Grounded spatial reasoning in vision language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.01584.

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., and Nießner, M. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2432–2443, 2017.

Dao, T. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2307.08691.

Deitke, M., VanderBilt, E., Herrasti, A., Weihs, L., Salvador, J., Ehsani, K., Han, W., Kolve, E., Farhadi, A., Kembhavi, A., and Mottaghi, R. ProcTHOR: Large-scale embodied AI using procedural generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2206.06994.

Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., and Farhadi, A. Objaverse: A universe of annotated 3D objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13142–13153, 2023.

Fan, Z., Zhang, J., Li, R., Zhang, J., Chen, R., Hu, H., Wang, K., Qu, H., Zhou, S., Wang, D., Yan, Z., Xu, H., Theiss, J., Chen, T., Li, J., Tu, Z., Wang, Z., and Ranjan, R. VLM-3R: Vision-language models augmented with instruction-aligned 3D reconstruction. *arXiv preprint arXiv:2505.20279*, 2025.

Feng, Q. Visuospatial cognitive assistant. *arXiv preprint arXiv:2505.12312*, 2025.

Kemp, C. C., Edsinger, A., Clever, H. M., and Matulevich, B. The design of Stretch: A compact, lightweight mobile manipulator for indoor human environments. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3156–3163, 2022. doi: 10.1109/ICRA46639.2022.9811922.

Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., Deitke, M., Ehsani, K., Gordon, D., Zhu, Y., Kembhavi, A., Gupta, A., and Farhadi, A. AI2-THOR: An interactive 3D environment for visual AI. *arXiv preprint arXiv:1712.05474*, 2017.

Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., and Li, C. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2304.08485.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296, 2024. arXiv:2310.03744.

NVIDIA. NVIDIA Isaac Sim. https://developer.nvidia.com/isaac/sim, 2025. Version 5.1; accessed 2026-11-01.

Ouyang, K., Liu, Y., Wu, H., Liu, Y., Zhou, H., Zhou, J., Meng, F., and Sun, X. SpaceR: Reinforcing MLLMs in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025.

Rajbhandari, S., Rasley, J., Ruwase, O., and He, Y. ZeRO: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–16, 2020. arXiv:1910.02054.

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., and Batra, D. Habitat-matterport 3D dataset (HM3D): 1000 large-scale 3D environments for embodied AI. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. arXiv:2109.08238.

Ray, A., Duan, J., Brown, E., Tan, R., Bashkirova, D., Hendrix, R., Ehsani, K., Kembhavi, A., Plummer, B. A., Krishna, R., Zeng, K.-H., and Saenko, K. SAT: Dynamic spatial aptitude training for multimodal language models. In *Conference on Language Modeling (COLM)*, 2025. arXiv:2412.07755.

Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M. A., Paczan, N., Webb, R., and Susskind, J. M. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10892–10902, 2021.

Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., Parikh, D., and Batra, D. Habitat: A platform for embodied AI research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9338–9346, 2019.

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P. Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 23–30, 2017. arXiv:1703.06907.

Wang, H., Chen, J., Huang, W., Ben, Q., Wang, T., Mi, B., Huang, T., Zhao, S., Chen, Y., Yang, S., Cao, P., Yu, W., Ye, Z., Li, J., Long, J., Wang, Z., Wang, H., Zhao, Y., Tu, Z., Qiao, Y., Lin, D., and Pang, J. GRUtopia: Dream general robots in a city at scale. *arXiv preprint arXiv:2407.10943*, 2024.

Wu, D., Liu, F., Hung, Y.-H., and Duan, Y. Spatial-MLLM: Boosting MLLM capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*, 2025.

Xu, R., Wang, W., Tang, H., Chen, X., Wang, X., Chu, F.-J., Feiszli, M., and Liang, K. J. Multi-SpatialMLLM: Multi-frame spatial understanding with multi-modal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. arXiv:2505.17015.

Yang, A. et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Yang, J., Yang, S., Gupta, A. W., Han, R., Fei-Fei, L., and Xie, S. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643, 2025a. arXiv:2412.14171.

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., and Xie, S. Cambrian-S: Towards spatial supersensing in video. *arXiv preprint arXiv:2511.04670*, 2025b.

Yeshwanth, C., Liu, Y.-C., Nießner, M., and Dai, A. ScanNet++: A high-fidelity dataset of 3D indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12–22, 2023.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952, 2023.

Zhang, J., Chen, Y., Zhou, Y., Xu, Y., Huang, Z., Mei, J., Chen, J., Yuan, Y.-J., Cai, X., Huang, G., Quan, X., Xu, H., and Zhang, L. From flatland to space: Teaching vision-language models to perceive and reason in 3D. *arXiv preprint arXiv:2503.22976*, 2025a.

Zhang, K., Li, B., Zhang, P., Pu, F., Cahyono, J. A., Hu, K., Liu, S., Zhang, Y., Yang, J., Li, C., and Liu, Z. LMMs-Eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 881–916, 2025b. doi: 10.18653/v1/2025.findings-naacl.51.

Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., and Li, C. LLaVA-Video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025c. arXiv:2410.02713.

Zheng, J., Zhang, J., Li, J., Tang, R., Gao, S., and Zhou, Z. Structured3D: A large photo-realistic dataset for structured 3D modeling. In *European Conference on Computer Vision (ECCV)*, 2020. arXiv:1908.00222.

Zhou, Q.-Y., Park, J., and Koltun, V. Open3D: A modern library for 3D data processing. *arXiv preprint arXiv:1801.09847*, 2018.

A. Additional Pipeline, Prompt, and Verification Details

A.1. Exact Prompt Formatting and Disambiguation Removal

To ensure zero formatting confounds between **+ RealHomes5K (Ours)**, **+ SIMS-VSI (5K)**, and **+ RealHomes5K (w/o Video)**, all conversations use the exact prompt wrapper from the official SIMS-VSI dataset release:

- **Multiple-Choice (MC) wrapper:** `<image>These are frames of a video.\n{question}\n` followed by `"Please select the correct option from the previous choices."`
- **Open-Ended (OE) wrapper:** `<image>These are frames of a video.\n{question}\n` followed by `"Please answer the question using a single word or phrase."`
- **Concrete Example (RealHomes5K, MC):** `<image>These are frames of a video. Measuring from the closest point of each object, what is the direct distance between the armchair and the stove (in meters)? A. 11.5 B. 3.3 C. 8.2 D. 9.2 Please select the correct option from the previous choices.` → Answer: `B. 3.3.`

Removal of the Extra Absolute-Distance Sentence. SIMS-V's open-source template generator appends a multiple-instance sentence to absolute-distance questions (*"If there are multiple instances of an object category, use the pair of instances with the shortest distance."*). Because neither the SIMS-VSI release nor VSI-Bench contains this sentence, we removed it from **RealHomes5K** before training.

A.2. Coordinate System Transformation and Handedness Preservation

GRScenes-100 USD stages are authored in a right-handed, Z-up coordinate system in centimeters: $(x_{\text{GRS}}, y_{\text{GRS}}, z_{\text{GRS}})$. SIMS-V operates in a left-handed, Y-up coordinate system in meters: $(x_{\text{SIMS}}, y_{\text{SIMS}}, z_{\text{SIMS}})$. Mapping $(x, y, z) \mapsto \frac{1}{100}(x, z, y)$ swaps the second and third axes (determinant -1), simultaneously converting Z-up to Y-up and flipping handedness from right-handed to left-handed so that egocentric left/right relative-direction answers remain exact (verified across all 2,378 unique questions).

A.3. Hugging Face Checkpoint Cross-Reference

Because our public Hugging Face repositories retain the internal run identifiers from our training logs, Table 6 provides a direct cross-reference between the model variants reported in the paper and their corresponding Hugging Face checkpoint IDs.

Table 6. **Cross-reference between paper model variants and public Hugging Face checkpoints.**

Paper Model Variant	Training Data & Visual Input	Hugging Face Repository ID	VSI-Bench / Debiased
Base	Zero-shot (un-finetuned base model)	`lmms-lab/LLaVA-Video-7B-Qwen2`	35.0 / 29.4
+ SIMS-VSI (5K)	4,756 SIMS-VSI conv. + AI2-THOR video	`ob11/grs-llava-video-7b-arm_C_seed0`	38.5 / 31.7
+ RealHomes5K (w/o Video)	4,756 RealHomes5K conv. + black video (blind)	`ob11/grs-llava-video-7b-arm_E_seed0`	39.8 / 33.8
+ RealHomes5K (Ours)	**4,756 RealHomes5K conv. + Isaac Sim RTX video**	`ob11/grs-llava-video-7b-arm_B_seed0`	**43.8 / 37.2**