NAV3R: BENCHMARKING NAVIGATION-CONSTRAINED RGB CAPTURE FOR 3D RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Embodied 3D reconstruction requires an agent to decide not only how to reconstruct a scene, but also where to move and which observations to retain. Yet exploration and reconstruction are typically evaluated separately, leaving unclear whether better exploration actually produces better inputs for reconstruction. We introduce **NAV3R**, a benchmark for navigation-constrained RGB capture and 3D reconstruction. Agents explore unfamiliar indoor scenes under movement and frame budgets, and their retained observations are evaluated jointly by exploration coverage, pose-free geometric reconstruction, and novel-view synthesis. Across 51 scenes from three indoor environments, we benchmark diverse navigation and view-selection policies with multiple RGB reconstruction models. Our results reveal a consistent trade-off: broader exploration improves whole-scene completion and generally benefits novel-view synthesis, but is associated with higher reconstruction error over the observed region. We further find that reconstruction quality depends strongly on acquisition structure, including view overlap, spatial spread, and frame selection. These results show that coverage and reconstruction quality are not interchangeable objectives, motivating joint evaluation of exploration, geometry, and rendering for embodied 3D reconstruction.

1 INTRODUCTION

Embodied 3D reconstruction is fundamentally an *acquisition problem*: an agent must decide where to move and which RGB observations are worth retaining. Recent feed-forward reconstruction models can infer scene geometry and camera parameters directly from unposed RGB images (Wang et al., 2025b; Chen et al., 2026b; Wang et al., 2026b; 2025a), but their inputs are typically treated as given. In an embodied setting, they are not. Movement determines which surfaces can be observed, while frame selection determines the overlap, spatial extent, and visual evidence available to the reconstructor. Under finite motion and capture budgets, an agent must therefore trade off *exploring more of the environment* against *acquiring views that can be reconstructed reliably*. This raises a basic question that existing evaluation does not directly answer: *does better exploration actually produce better inputs for 3D reconstruction?*

Existing embodied benchmarks study complementary parts of this problem. Instruction-following, multi-goal and embodied question-answering benchmarks (Anderson et al., 2018; Krantz et al., 2020; Khanna et al., 2024; Majumdar et al., 2024; Yang et al., 2025a) evaluate task success, GLEAM-Bench (Chen et al., 2025) connects exploration to active mapping, and active reconstruction methods such as AREA3D (Xu et al., 2026), Creat3r (Tsai et al., 2026), FisherRF (Jiang et al., 2024) and COVER (Chen et al., 2026a) select informative viewpoints. We target a different interface: the output of an embodied policy is the *RGB capture set it acquires while navigating*, evaluated by downstream pose-free reconstruction and held-out rendering. A policy therefore cannot appear successful merely by reaching distant locations, nor by accurately reconstructing a small, highly overlapping region.

We introduce **NAV3R**, a benchmark for **navigation-constrained RGB capture for 3D reconstruction**. NAV3R contains 123,808 RGB observations across 51 indoor scenes from Matterport3D (Chang et al., 2017), iGibson (Shen et al., 2021), and ProcTHOR-10K (Deitke et al., 2022). Each scene is represented by a standardized offline pose graph. An agent starts from a fixed entrance, explores under a movement budget, and retains a bounded set of RGB observations for reconstruc-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

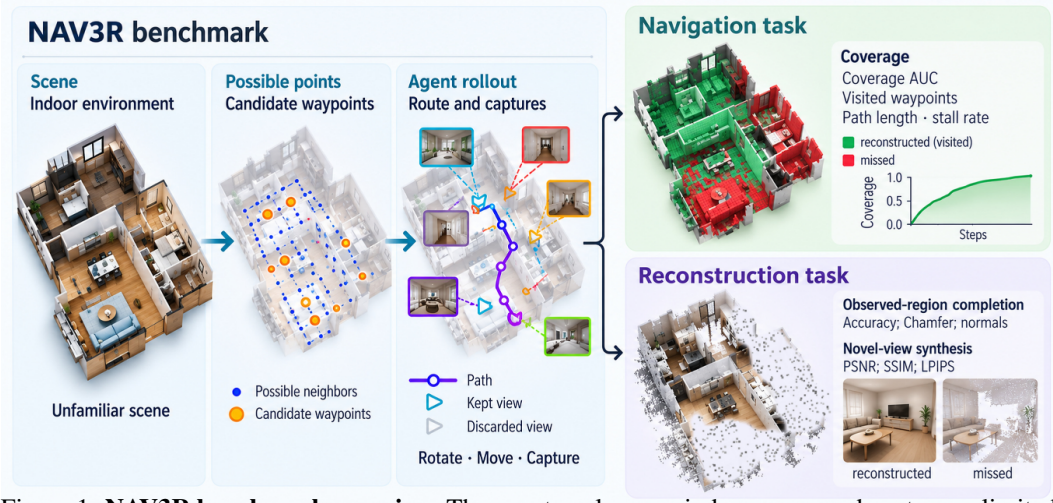

Figure 1: **NAV3R benchmark overview.** The agent explores an indoor scene and captures a limited set of RGB views from candidate viewpoints. Its route and captures determine navigation coverage, while the retained images determine reconstruction quality. Green marks overlap between the reconstructed point cloud and ground truth; red marks missed surfaces. The agent visits the bottom-left room but captures no image there, leaving it largely unreconstructed.

tion. Policies observe the current RGB image, their own pose, and an online graph containing only previously visited and locally revealed waypoints; ground-truth depth and scene geometry are reserved for evaluation. NAV3R evaluates every capture set along three complementary axes: **exploration coverage**, **geometric reconstruction** of observed and whole-scene surfaces, and **novel-view synthesis** on held-out views. Because trajectory generation is separated from reconstruction, the same capture set can be replayed through multiple reconstruction backends without changing the acquisition policy.

Our experiments expose a tension between *coverage* and *local geometric fidelity*. Broader exploration is associated with higher completion error over the surfaces the agent actually observed, yet improves whole-scene completion and generally improves held-out novel-view synthesis. The effect is not explained by frame count alone: greater co-visibility between views is associated with lower geometric error, spreading a fixed budget over a larger area with higher error, and changing frame selection along a fixed trajectory improves reconstruction without changing where the agent travels.

Our contributions are: 1) **Navigation-constrained RGB capture for reconstruction.** We formulate navigation-constrained RGB capture, coupling embodied exploration, frame selection, and downstream 3D reconstruction; 2) **NAV3R, a unified benchmark for exploration and reconstruction.** NAV3R standardizes 51 scenes with protocols for coverage, pose-free geometry, and held-out novel-view synthesis; and 3) **A systematic analysis of the exploration–reconstruction trade-off.** We expose a coverage-reconstruction trade-off and identify overlap, spatial spread, and frame selection as key acquisition factors.

2 Related Work

Embodied navigation and understanding. Instruction-following (Anderson et al., 2018; Krantz et al., 2020), object-goal navigation (Yokoyama et al., 2024), multi-goal navigation (Khanna et al., 2024), question answering (Majumdar et al., 2024), and embodied-agent benchmarks (Yang et al., 2025a) evaluate whether an agent completes its assigned task (Table 1). NAV3R instead treats the retained image set as the agent's output and evaluates the reconstruction it supports. We test both prompted VLMs and navigation specialists, including OmniNav (Xue et al., 2026), JanusVLN (Zeng et al., 2026), NaVILA (Cheng et al., 2025), and VLN-Analogical (Zhang et al., 2025), to examine how their navigation behaviour transfers to image acquisition.

Table 1: Comparison with existing embodied benchmarks. Syn: synthetic; Cov: coverage; Geo: geometry; PF: pose-free; RGB: reconstruction from RGB; NVS: novel-view synthesis; IL/RL: imitation/reinforcement learning.

Benchmark	Scenes	Real	Syn	Task	Policies	Cov	Geo	PF	RGB	NVS
R2R	29	✓	×	Instruction	IL	×	×	×	–	×
VLN-CE	29	✓	×	Instruction	IL	×	×	×	–	×
HM3D-OVON	36	✓	×	ObjectNav	IL, RL, modular	×	×	×	–	×
GOAT-Bench	36	✓	×	Multi-goal	RL, modular	×	×	×	–	×
OpenEQA	180+	✓	×	QA	LLM, VLM	×	×	×	–	×
EmbodiedBench	–	×	✓	Instruction	MLLM	×	×	×	–	×
GLEAM-Bench	128	✓	✓	Mapping	Planner, RL	✓	✓	×	×	×
NAV3R (ours)	51	✓	✓	3D capture	VLM, VLN	✓	✓	✓	✓	✓

Active mapping and view selection. GLEAM-Bench (Chen et al., 2025) evaluates exploration using coverage and geometric mapping quality, making it a close point of comparison. NAV3R focuses on the interaction between navigation-constrained RGB capture and pose-free reconstruction. NextBestPath (Li et al., 2025) predicts long-horizon goals using accumulated coverage gain and an obstacle map, extending local next-best-view selection to path planning in unknown indoor environments. GenNBV (Chen et al., 2024a) learns a generalizable next-best-view policy that moves through 5D free space to reconstruct unseen objects and scenes. Other active-reconstruction methods, including AREA3D (Xu et al., 2026), Creat3r (Tsai et al., 2026), FisherRF (Jiang et al., 2024), and COVER (Chen et al., 2026a), provide alternative strategies for selecting informative views. We evaluate adaptations of these four selectors separately because they choose from a viewpoint pool without traversing the graph. Their results contextualize capture quality but do not isolate the effect of navigation constraints.

Reconstruction from RGB observations. DUSt3R (Wang et al., 2024), CUT3R (Wang et al., 2025b), TTT3R (Chen et al., 2026b), π^3 (Wang et al., 2026b), and VGGT (Wang et al., 2025a; 2026a) estimate geometry from RGB observations. NAV3R uses reconstruction models as downstream evaluators of an agent's capture set, allowing the same policy to be assessed with different backends. We also evaluate rendering from Gaussian splats (Kerbl et al., 2023; Fan et al., 2024) to determine whether geometric and appearance-based evaluations yield similar policy rankings.

3 NAV3R Benchmark

NAV3R evaluates an agent's ability to collect RGB images for 3D reconstruction while moving through an unfamiliar indoor scene, where the agent explores and retains a bounded image set, which is then evaluated for coverage, geometry, and rendering quality (Figure 1).

3.1 Overview of Dataset

NAV3R contains 51 scenes from Matterport3D (MP3D) (Chang et al., 2017), iGibson (Shen et al., 2021), and ProcTHOR-10K (Deitke et al., 2022). MP3D and iGibson provide environments derived from real scans; ProcTHOR-10K provides procedurally generated houses. The evaluation includes the 18 MP3D test scenes, 15 interactive iGibson scenes, and 18 selected ProcTHOR-10K test houses with 7–10 rooms. Scene statistics, identifiers and selection details are in Appendices A.1–A.3. We use a shared acquisition pipeline to standardize waypoint separation, camera height, and ground-truth construction across the three simulators.

Image collection. In every scene we place waypoints over the walkable floor, at least 1 meter(m) apart. The simulators' own samplers do not guarantee this spacing, so we order the waypoints in a back-and-forth (boustrophedon) sweep and greedily drop any waypoint that is less than 1.0 m from one already kept. At each waypoint we take 8 RGB images, one every 45° of heading, with the

camera fixed at 1.25 m above the floor in all three datasets. This gives each scene a dense, evenly spaced set of panoramic views, from which every agent in the benchmark draws its images.

Ground-truth geometry. The three datasets store scene geometry in different forms (Table 6), so we build the ground truth the same way for all of them. Depth images rendered at every one of these viewpoints are back-projected to world coordinates and fused into one point cloud per scene (0.02 m voxels), with normals estimated by local PCA ($k = 30$). This cloud supplies the reference surface for evaluation; completion is evaluated on the observed region or the whole scene as specified below.

Offline pose graph. Running a simulator for the length of a long rollout is slow, so we turn the images of each scene into an offline pose graph. The agent's state is a waypoint and one of its 8 headings. Its primitive actions are rotation, movement and capture. ROTATE turns the agent in place to another heading. MOVE takes the agent to the nearest waypoint within 45° of the direction it faces, or of the opposite direction. That waypoint must be no more than 1 m above or below the current one, no more than four times the median waypoint spacing away, and reachable by a direct walkable path in the simulator, so the agent can climb stairs one hop at a time. CAPTURE adds the current view to the frames used for reconstruction. The planner interface also accepts GOTO requests specifying a destination waypoint; movement remains subject to the graph constraints and shared hop budget. We take each image's viewing direction from its stored ground-truth camera rotation.

3.2 EVALUATION PROTOCOL AND METRICS

All navigation policies use a shared episode loop with the same entrance waypoint, candidate poses and dataset-specific budgets: 30 retained frames on iGibson, 100 on ProcTHOR-10K and 200 on MP3D, with respective hop limits of 75, 250 and 500. Agents receive RGB observations, their own pose and an online map of visited and adjacency-revealed waypoints; ground-truth depth and surfaces are reserved for evaluation. VLMs choose captures explicitly, while VLN policies capture one frame per newly reached waypoint. A shared stall rule terminates episodes when consecutive model calls change neither the waypoint nor the retained frame set. See detailed description in Appendix A.4

We summarize the evaluation metrics used to evaluate navigation, geometry, and novel-view synthesis. Coverage measures the fraction of scene waypoints within 2 m of a captured waypoint. Coverage AUC averages coverage over the frame budget, holding the final value for unused slots when an agent stops early. Geometry is evaluated primarily by observed-region completion: the mean distance from reference surfaces seen by the captured frames to the reconstructed point cloud. We also report accuracy, Chamfer distance, normal consistency and depth errors; whole-scene completion is evaluated separately in Section 4.3. Novel-view synthesis is scored with PSNR, SSIM and LPIPS on held-out views. Detailed alignment and rendering settings are provided in Appendix A.5.

4 EXPERIMENTS

We evaluate whether agents explore effectively, whether broader exploration improves reconstruction, which properties of the captured frames are associated with geometric error, and whether novel-view synthesis yields similar policy rankings.

4.1 EXPERIMENTAL SETUP

We compare capture policies under the protocol in Section 3.2 and evaluate their retained images with four reconstruction models. Navigation policies share entrance waypoints and dataset-specific frame and movement budgets; view selectors are evaluated separately because they do not traverse the pose graph.

Capture policies. We evaluate a random walk and two scripted greedy oracles, which select frontiers by predicted mass gain or proximity. The eight VLM planners are Qwen2.5-VL-7B and Qwen2.5-VL-3B (Bai et al., 2025), VST-7B (Yang et al., 2025b), InternVL3-8B (Zhu et al., 2025), InternVL2.5-8B (Chen et al., 2024b), Kimi-VL-A3B (Kimi Team, 2025), Ovis2-8B (Lu et al., 2024), and GPT-5.6-luna (OpenAI, 2026). Each receives the current RGB frame, an online map, and a short

Table 2: Final coverage, Cov AUC and observed-region completion error (mean over the four reconstruction models) for every navigation policy. Per-model results: Tables 9 and 10; stall rates: Table 11.

	iGibson			ProcTHOR-10K			MP3D		
Policy	Cov↑	AUC↑	Comp↓	Cov↑	AUC↑	Comp↓	Cov↑	AUC↑	Comp↓
Random walk	0.678	0.518	1.01	0.484	0.353	1.24	0.331	0.245	1.45
Greedy (mass gain)	0.843	0.587	1.24	0.842	0.508	1.89	**0.661**	**0.393**	2.30
Greedy (nearest)	0.842	0.569	1.18	0.824	0.493	2.01	0.656	0.370	2.16
Qwen2.5-VL-7B	0.559	0.404	0.92	0.428	0.292	1.16	0.223	0.169	1.00
Qwen2.5-VL-3B	0.312	0.275	0.73	0.110	0.104	1.02	0.076	0.063	**0.74**
VST-7B	0.737	0.561	1.22	0.473	0.364	1.48	0.361	0.275	1.69
InternVL3-8B	0.867	0.612	1.26	0.833	0.517	2.06	0.616	0.389	2.19
InternVL2.5-8B	0.655	0.497	1.11	0.455	0.338	1.52	0.265	0.205	1.32
Kimi-VL-A3B	0.180	0.176	0.63	0.047	0.047	**0.99**	0.020	0.020	0.75
Ovis2-8B	0.752	0.553	1.00	0.613	0.422	1.63	0.312	0.237	1.23
GPT-5.6-luna	**0.903**	**0.656**	1.27	**0.883**	**0.581**	1.79	0.669§	0.427	2.19
OmniNav	0.545	0.467	0.76	0.309	0.270	1.12	0.287	0.240	1.29
VLN-Analogical	0.680	0.569	1.07	0.291	0.265	1.28	0.212	0.195	1.17
JanusVLN	0.404	0.363	**0.62**	0.095	0.093	1.08	0.168	0.148	0.83
NaVILA	0.204	0.199	0.98	0.070	0.069	1.25	0.024	0.024	0.99

§ the endpoint's content filter refused the remaining trajectories (Appendix A.9); not eligible for bold.

text state, and returns actions using the same prompt and sampling settings. The four VLN policies are OmniNav (Xue et al., 2026), JanusVLN (Zeng et al., 2026), NaVILA (Cheng et al., 2025), and VLN-Analogical (Zhang et al., 2025). OmniNav, JanusVLN, and NaVILA use released checkpoints; VLN-Analogical runs on the Qwen2.5-VL-7B planner's weights. VLN policies emit motion, with one frame captured per newly reached waypoint, while VLM planners choose captures explicitly. We use Coverage AUC for comparisons across these interfaces. Stochastic policies run five trajectories per scene and deterministic policies run one. AREA3D (Xu et al., 2026), Creat3r (Tsai et al., 2026), FisherRF (Jiang et al., 2024), and COVER (Chen et al., 2026a) select views from a pool without navigation (Appendix A.6).

Reconstruction and rendering. We evaluate captured images with CUT3R (Wang et al., 2025b), TTT3R (Chen et al., 2026b), π^3 (Wang et al., 2026b), and VGGT-Ω (Wang et al., 2026a). These models estimate geometry and camera parameters from RGB inputs without supplied calibration or bundle adjustment. For eight capture policies, we also fit Gaussian splats from the reconstructed points and evaluate held-out views using the protocol in Appendix A.5.

4.2 Do agents explore effectively?

We compare coverage and progress across VLM planners, VLN policies, and scripted baselines. Table 2 reports coverage and downstream completion error for every navigation policy, with per-model results in Tables 9 and 10; we examine the geometric results in Section 4.3.

Exploration coverage. InternVL3-8B and GPT-5.6-luna exceed the random walk's final coverage on all three datasets. GPT-5.6-luna reaches 0.903 coverage on iGibson and 0.883 on ProcTHOR-10K, above both greedy oracles, while InternVL3-8B has the highest coverage among the open VLM planners on each dataset. Among VLN policies, VLN-Analogical has the highest Coverage AUC on iGibson (0.569), and OmniNav leads on ProcTHOR-10K and MP3D (0.270 and 0.240). These VLN results exceed the random walk's Coverage AUC on iGibson but fall below it on the other two datasets. GPT-5.6-luna's MP3D results include 81 of 90 trajectories because the endpoint refused images in the remaining runs (Appendix A.9). *Exploration performance varies substantially across policies; only two of the eight VLM planners exceed the random walk's final coverage on all three datasets under the reported evaluation.*

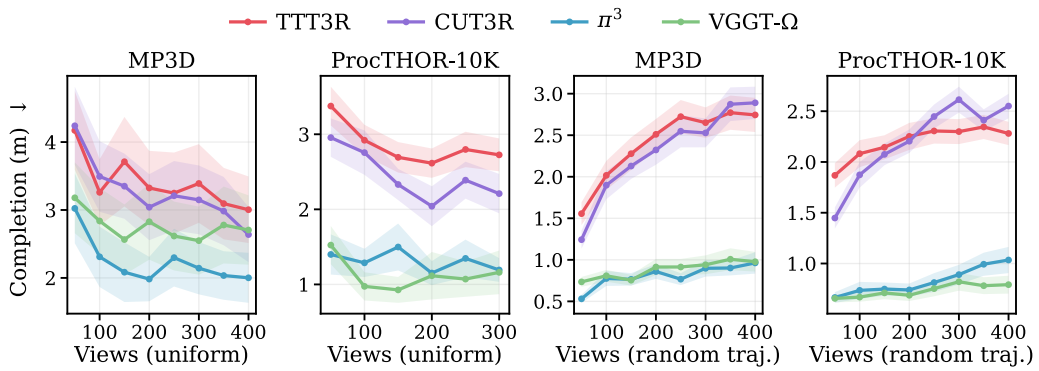

Figure 2: Observed-region completion against capture size. Columns show Matterport3D and ProcTHOR-10K. Top: uniform sampling with one heading per waypoint over 18 scenes; requested view budgets are capped by the number of available waypoints. Bottom: 90 random trajectories over 18 scenes with increasing capture budgets. Bands are ± 1 SEM over scenes or trajectories, respectively. Additional metrics are in Appendices 24 and 26.

Stalling and action parsing. On ProcTHOR-10K, every open VLM planner stalls on at least 44% of trajectories, and six of seven stall on more than 85%. GPT-5.6-luna's stall rates are 19%, 8%, and 3% on iGibson, ProcTHOR-10K, and MP3D, respectively (Appendix 11). The open VLMs have zero reported parse failures, so successful action parsing does not imply progress. Among VLN policies, VLN-Analogical never stalls, whereas NaVILA stalls on every trajectory and JanusVLN on 42 of 51. *Under the shared stall rule, many policies produce valid actions yet fail to sustain exploration.*

Action interface and policy adaptation. OmniNav predicts a waypoint, and VLN-Analogical selects among candidate views. JanusVLN emits 25 cm steps and turns, while NaVILA's motion commands are decomposed into 25 cm steps. NaVILA visits only 2.3–2.5 waypoints per trajectory; JanusVLN requires 6–19 calls to cross one graph edge. Since the stall rule counts model calls, these results depend on the adaptation to the graph interface. Using the same Qwen2.5-VL-7B weights, VLN-Analogical achieves higher Coverage AUC than the prompted planner on iGibson and MP3D (0.569 vs. 0.404 and 0.195 vs. 0.169), but lower AUC on ProcTHOR-10K (0.265 vs. 0.292). Exposing the full floor layout also changes planner rankings (Appendix A.10). *The observed rankings reflect both model capabilities and their action and observation interfaces.*

4.3 Does more exploration improve reconstruction?

We examine this question through capture-budget sweeps, comparisons across policies, and within-scene correlations. The distinction between observed-region and whole-scene completion matters: each trajectory observes a different surface subset, whereas whole-scene completion uses the same reference surface for every trajectory in a scene.

Capture budget and sampling strategy. Figure 2 compares uniformly sampled waypoints with five random walks in each of 18 scenes, evaluated at eight budgets from 50 to 400 frames. All 90 walks reach every budget. Uniform sampling is capped by each scene's waypoint count, so its nominal budget can exceed the achieved number of inputs; plateaus after this cap do not demonstrate model saturation. Observed-region completion error decreases with budget under uniform sampling but increases along the evaluated random walks. On MP3D, CUT3R's error rises from 1.243 m to 2.890 m and TTT3R's from 1.556 m to 2.744 m. The increases are smaller for π^3 (0.531 m to 0.963 m) and VGGT-Ω (0.735 m to 0.977 m). Longer walks change both the input sequence and the region evaluated, so these curves do not isolate accumulated drift. *In these sweeps, the relationship between frame count and completion error depends on how the frames are acquired.*

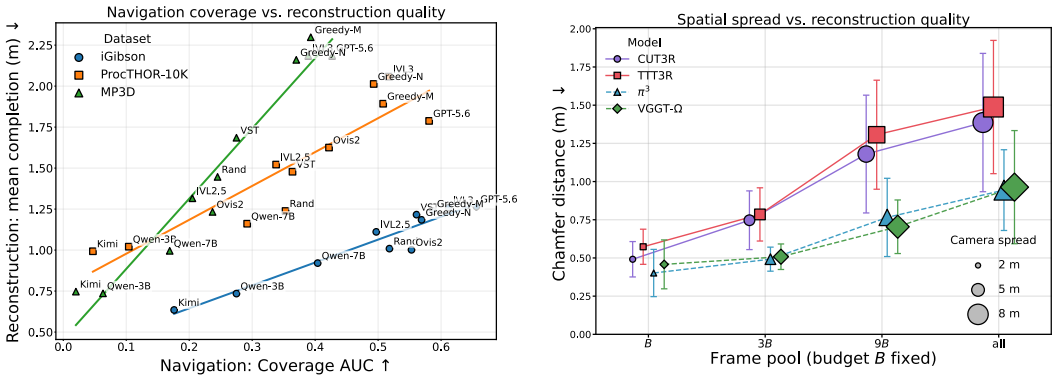

Figure 3: Navigation coverage and reconstruction. Left: Coverage AUC against observed-region completion error for the policies in Table 9; marker shape denotes the dataset, and completion is averaged over four reconstruction models. Right: effect of spatial spread at a fixed frame budget. Frames are sampled from pools of at least B, $3B$, or $9B$ views around a random centre, or from the whole reachable scene. Points average the three datasets, error bars show the standard deviation across datasets, and marker size encodes mean pairwise camera spread.

Table 3: Within-scene Spearman correlation between navigation coverage and reconstruction error over trajectories from the policies in Section 4.2 and Appendix A.6. CD denotes Chamfer distance, and AbsRel denotes absolute relative depth error. Positive values indicate that greater navigation coverage is associated with larger reconstruction error. Grey values have a 95% CI that includes zero.

Dataset	Model	CD (observed)	Accuracy (scene)	Completion (observed)	Completion (scene)	Depth AbsRel
iGibson	CUT3R	+0.24	+0.39	+0.42	−0.50	+0.42
	TTT3R	+0.37	+0.30	+0.48	−0.46	+0.40
	π^3	−0.04	+0.45	+0.30	−0.75	+0.23
	VGGT-Ω	+0.08	+0.57	+0.40	−0.65	+0.37
ProcTHOR-10K	CUT3R	+0.53	+0.21	+0.50	−0.72	+0.66
	TTT3R	+0.60	+0.15	+0.54	−0.70	+0.60
	π^3	+0.00	+0.36	+0.11	−0.83	+0.25
	VGGT-Ω	−0.05	+0.47	+0.03	−0.87	+0.37
MP3D	CUT3R	+0.49	+0.49	+0.56	−0.71	+0.54
	TTT3R	+0.51	+0.39	+0.58	−0.71	+0.49
	π^3	+0.20	+0.61	+0.41	−0.81	+0.28
	VGGT-Ω	+0.16	+0.43	+0.35	−0.79	+0.12

Coverage and observed-region geometry. Across the 19 policies in Tables 2 and 7, Coverage AUC and observed-region completion error, averaged over four reconstruction models, have Pearson correlations of 0.82, 0.89, and 0.86 on iGibson, ProcTHOR-10K, and MP3D (Figure 3, left). The correlations are stronger for CUT3R and TTT3R ($r = 0.79$–0.95) than for π^3 and VGGT-Ω ($r = 0.30$–0.72). On MP3D, GPT-5.6-luna, InternVL3-8B, and the greedy oracles have CUT3R completion errors of 2.76–2.97 m, compared with 2.05 m for the random walk. Conversely, Kimi-VL-A3B has the lowest completion error on iGibson under all four models while covering only 0.18 of the scene. These comparisons include different observed regions and, for view selectors, different acquisition constraints. *Low reconstruction error does not necessarily indicate effective navigation: a policy may achieve it while covering only a small part of the scene.*

Whole-scene and observed-region evaluation. Table 3 reports within-scene Spearman correlations, with 95% confidence intervals from 2000 bootstrap resamples of whole scenes. On all three datasets, coverage correlates negatively with whole-scene completion error under every model

Table 4: CUT3R Chamfer distance (m) on iGibson for the frames a planner kept and for those a ground-truth coverage oracle selects from the same walk. All datasets and models: Table 17.

Planner	Kept	Oracle	Change
Qwen2.5-VL-3B	0.535	0.430	−20%
Ovis2-8B	0.705	0.645	−9%
InternVL3-8B	0.855	0.685	−20%

Table 5: Novel-view synthesis with CUT3R on iGibson, on a held-out view set fixed per scene (Section A.5). All datasets and models: Table 15.

Policy	PSNR ↑	SSIM ↑	LPIPS ↓
Random walk	10.20	0.422	0.538
Qwen2.5-VL-7B	9.26	0.345	0.545
InternVL3-8B	10.57	0.473	0.537
GPT-5.6-luna	10.96	0.495	0.532
Greedy (mass gain)	10.59	0.466	0.538
OmniNav	8.94	0.293	0.550
VLN-Analogical	10.06	0.413	0.543
AREA3D	**11.03**	**0.516**	**0.527**

($\rho = -0.46$ to -0.87), and positively with accuracy error against the whole surface. Per-frame depth error also correlates positively; its interval includes zero only for VGGT-Ω on MP3D. Observed-region completion has positive coefficients, although the intervals include zero for π^3 and VGGT-Ω on ProcTHOR-10K. Chamfer distance has positive correlations for CUT3R and TTT3R; the intervals include zero for both global-attention models. *Broader coverage is associated with better whole-scene completion alongside larger accuracy and depth errors.*

4.4 Which capture properties are associated with geometric error?

We examine frame overlap, spatial spread, selection, and ordering using the recorded trajectories and controlled changes to their inputs. Within-scene correlations and confidence intervals follow the procedure above. These analyses distinguish changes to the frame set from changes to its order, while spatial extent and observed area can still vary together.

Overlap between frames. Among 20 descriptors of frame poses and overlap, mean pairwise co-visibility, measured by the IoU of visible surface cells, correlates negatively with Chamfer distance on every dataset and under every model ($\rho = -0.27$ to -0.48). The number of disconnected groups of overlapping frames correlates positively ($\rho = +0.18$ to $+0.56$). Both associations remain after a Benjamini–Hochberg correction across descriptors. *Frame sets with greater shared visibility are associated with lower reconstruction error.*

Spatial extent at a fixed budget. We hold the frame count and ordering rule fixed and sample frames from progressively wider pools around a random centre (Figure 3, right). From the narrowest pool to the whole scene, mean pairwise camera distance rises from 1.2–3.6 m to 4.7–10.6 m, and Chamfer distance increases by a factor of 1.6–3.1 across models and datasets. Within-scene correlations between camera distance and Chamfer distance are positive in all 12 model–dataset combinations ($\rho = +0.23$ to $+0.78$), with intervals excluding zero in 10. The wider pools also cover more of the scene. *Spreading a fixed frame budget over a larger region is associated with higher error here, but this experiment does not separate spatial spread from observed area.*

Frame selection along a fixed walk. We replay the trajectories of three VLM planners spanning the reported Chamfer-distance rankings. A coverage oracle uses ground truth to select the same number of frames from all headings at the visited waypoints. Its selections have lower Chamfer distance in all 36 planner–dataset–model combinations, with reductions of 4%–48% (median 23%). Table 4 shows iGibson with CUT3R; every dataset and model is in Table 17. This oracle is an analysis reference, since ground-truth surfaces are unavailable to benchmark policies. *Changing frame selection can improve reconstruction without changing the visited waypoints.*

Order of the captured frames. Shuffling a fixed frame set increases CUT3R's Chamfer distance in all nine planner–dataset pairs, with intervals excluding zero in seven (Figure 4). For TTT3R, the shuffle intervals include zero in eight of nine pairs. A nearest-neighbour ordering changes Chamfer distance by at most 0.117 m, and its intervals include zero in 15 of 18 cases across the two models.

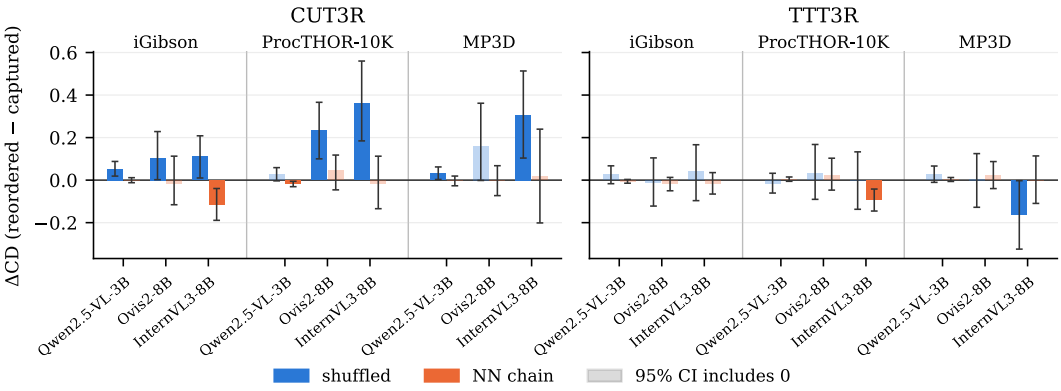

Figure 4: Effect of frame order on reconstruction. Bars show the change in Chamfer distance relative to capture order when the same retained frames are uniformly shuffled or arranged into a nearest-neighbour (NN) chain from the first frame. Positive values indicate higher reconstruction error. Error bars are 95% confidence intervals; grey bars have intervals that include zero.

CUT3R is sensitive to shuffling in these experiments, while nearest-neighbour reordering offers no consistent improvement over capture order.

4.5 Does novel-view synthesis agree with geometric evaluation?

We evaluate eight capture policies on a held-out view set fixed per scene, using PSNR, SSIM, and LPIPS. Table 5 shows iGibson with CUT3R; Table 15 covers every dataset and reconstruction model, and Appendix A.11 shows example renders. This evaluates rendering beyond the captured views, whereas observed-region completion measures only surfaces seen by the retained frames.

Coverage and held-out rendering. Using the tabulated means for all eight policies, Spearman correlations between Coverage AUC and PSNR range from 0.83 to 1.00 across datasets and reconstruction models. AREA3D has the highest CUT3R and TTT3R PSNR on every dataset, while GPT-5.6-luna has the highest VGGT-Ω PSNR. AREA3D selects from a viewpoint pool without navigating, so these rankings describe the resulting capture sets rather than equivalent navigation policies. The MP3D comparison also retains the incomplete GPT-5.6-luna evaluation noted above. *Across the evaluated capture sets, higher coverage is associated with better held-out PSNR, although policy rankings vary across rendering metrics.*

Agreement with completion error. On iGibson, Qwen2.5-VL-7B has lower CUT3R observed-region completion error than the random walk (0.982 vs. 1.144 m), but lower PSNR (9.26 vs. 10.20). Its PSNR is also lower than the random walk's for every other dataset and reconstruction model. *Observed-region completion and held-out rendering can rank the same capture policies differently, supporting their joint evaluation.*

5 Conclusion

We introduced NAV3R, a benchmark that scores the RGB frames an embodied agent keeps by the reconstruction and rendering they support. Three conclusions follow. First, reconstruction models should be tested on frames an agent would actually capture: more views help when sampled uniformly but hurt along a walk, most for the streaming models CUT3R and TTT3R (Section 4.3). Second, coverage alone is not enough when the goal is geometry. The policies that cover the most reconstruct the observed region less accurately (Section 4.3), yet coverage still decides how well unseen views can be rendered (Table 5), so a capture policy must be judged on both. Third, open VLMs are weak explorers: only InternVL3-8B beats a random walk on all three datasets, and VLN methods help only when they act at the level of waypoints (Section 4.2). Natural next steps are continuous control with motion blur and pose noise, as in VLN-CE, and capture policies that trade coverage against co-visibility between views.

AI USE STATEMENT

We used an LLM to improve paper writing (grammar and wording); all authors reviewed the edits and take full responsibility for the content.

ETHICS STATEMENT

This work evaluates reconstruction and navigation on three established indoor scene datasets — Matterport3D, iGibson and ProcTHOR-10K — used under their respective licences. Matterport3D consists of scans of real buildings; we use only the standard test split and the geometry and imagery as distributed, and we introduce no new capture of real environments and no human subjects. The remaining two sources are synthetic or reconstructed. We see no direct pathway from this benchmark to harmful application beyond the general dual-use character of indoor mapping, and we note that the benchmark's central finding — that exploring for coverage recovers more of a scene but reconstructs it less accurately — argues against, rather than for, indiscriminate capture.

REPRODUCIBILITY STATEMENT

The benchmark is designed so that results can be reproduced without re-running the policies. Trajectory generation and scoring are separated: a policy writes a frame list, and a manifest-driven sweep replays that list through each reconstruction model (Appendix A.8), so a published frame list is sufficient to reproduce every geometric and synthesis metric we report. The pose graph is offline and deterministic given a scene; the entrance waypoint is computed once per scene and fixed across trajectories; stochastic policies are seeded per trajectory. All four reconstruction models are run from their original repositories at their published settings, and no generation parameter of any vision-language model is tuned (Appendix A.7). One known source of irreproducibility remains: per-trajectory reconstruction metrics under CUT3R are not bit-deterministic across runs. Stochastic policy sweeps contain 75 or 90 trajectories (81 completed trajectories for GPT-5.6-luna on MP3D), while deterministic policies have one trajectory per scene, giving 15 or 18. The random-trajectory ladders use 90 trajectories each; uniform sampling uses 18 scenes per dataset. Tables and figure captions specify their aggregation units.

REFERENCES

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 3674–3683. IEEE, 2018.

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.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. In *International Conference on 3D Vision*, pp. 667–676, 2017.

Timothy Chen, Adam Dai, Maximilian Adang, Grace Gao, and Mac Schwager. Coverage optimization for camera view selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19443–19451, 2026a.

Xiao Chen, Quanyi Li, Tai Wang, Tianfan Xue, and Jiangmiao Pang. Gennbv: Generalizable next-best-view policy for active 3d reconstruction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16436–16445. IEEE, 2024a.

Xiao Chen, Tai Wang, Quanyi Li, Tao Huang, Jiangmiao Pang, and Tianfan Xue. Gleam: Learning generalizable exploration policy for active mapping in complex 3d indoor scenes. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5558–5568. IEEE, 2025.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. TTT3R: 3d reconstruction as test-time training. In *International Conference on Learning Representations*, 2026b.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.

An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Xueyan Zou, Jan Kautz, Erdem Biyik, Hongxu Yin, Sifei Liu, and Xiaolong Wang. NaVILA: Legged robot vision-language-action model for navigation. In *Robotics: Science and Systems*, 2025.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, and Roozbeh Mottaghi. ProcTHOR: Large-scale embodied AI using procedural generation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 5982–5994, 2022.

Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, et al. InstantSplat: Sparse-view Gaussian splatting in seconds, 2024.

Wen Jiang, Boshu Lei, and Kostas Daniilidis. FisherRF: Active view selection and uncertainty quantification for radiance fields using fisher information. In *European Conference on Computer Vision*, 2024.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023.

Mukul Khanna, Ram Ramrakhya, Gunjan Chhablani, Sriram Yenamandra, Theophile Gervet, Matthew Chang, Zsolt Kira, Devendra Singh Chaplot, Dhruv Batra, and Roozbeh Mottaghi. Goat-bench: A benchmark for multi-modal lifelong navigation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16373–16383. IEEE, 2024.

Kimi Team. Kimi-VL technical report. *arXiv preprint arXiv:2504.07491*, 2025.

Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*, pp. 104–120, 2020.

Shiyao Li, Antoine Guédon, Clémentin Boittiaux, Shizhe Chen, and Vincent Lepetit. Nextbest-path: Efficient 3d mapping of unseen environments. In *International Conference on Learning Representations*, volume 2025, pp. 49464–49482, 2025.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*, 2024.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul McVay, Oleksandr Maksymets, Sergio Arnaud, et al. OpenEQA: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16488–16498, 2024.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. OpenAI, 2026. URL `https://openai.com/index/gpt-5-6/`.

Bokui Shen, Fei Xia, Chengshu Li, Roberto Martin-Martin, Linxi Fan, Guanzhi Wang, Claudia P'erez-D'Arpino, Shyamal Buch, Sanjana Srivastava, Lyne P. Tchapmi, Micael E. Tchapmi, Kent Vainio, Josiah Wong, Li Fei-Fei, and Silvio Savarese. iGibson 1.0: A simulation environment for interactive tasks in large realistic scenes. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2021.

Chih-Jung Tsai, Hwann-Tzong Chen, and Tyng-Luh Liu. Creat3r: Confidence reaggregation for exploration-aware active 3d reconstruction. In *Forty-third International Conference on Machine Learning*, 2026.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D perception model with persistent state. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3d vision made easy. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations*, volume 2026, pp. 10481–10497, 2026b.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Tianling Xu, Shengzhe Gan, Leslie Gu, Yuelei Li, Fangneng Zhan, and Hanspeter Pfister. AREA3D: Active reconstruction agent with unified feed-forward 3d perception and vision-language guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Xinda Xue, Junjun Hu, Minghua Luo, Shichao Xie, Jintao Chen, Zixun Xie, Kuichen Quan, Wei Guo, Mu Xu, and Zedong Chu. OmniNav: A unified framework for prospective exploration and visual-language navigation. In *International Conference on Learning Representations*, 2026.

Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, et al. EmbodiedBench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. In *Proceedings of the International Conference on Machine Learning*, 2025a.

Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, Yi Lin, and Hengshuang Zhao. Visual spatial tuning. *arXiv preprint arXiv:2511.05491*, 2025b.

Naoki Yokoyama, Ram Ramrakhya, Abhishek Das, Dhruv Batra, and Sehoon Ha. Hm3d-ovon: A dataset and benchmark for open-vocabulary object goal navigation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5543–5550. IEEE, 2024.

Shuang Zeng, Dekang Qi, Xinyuan Chang, Feng Xiong, Shichao Xie, Xiaolong Wu, Shiyi Liang, Mu Xu, Xing Wei, and Ning Guo. JanusVLN: Decoupling semantics and spatiality with dual implicit memory for vision-language navigation. In *International Conference on Learning Representations*, 2026.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Yue Zhang, Tianyi Ma, Zun Wang, Yanyuan Qiao, and Parisa Kordjamshidi. Vision-and-language navigation with analogical textual descriptions in llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15028–15036, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.