I-Perceive: A Foundation Model for Vision-Language Active Perception

Yongxi Huang*, Zhuohang Wang†‡, Wenjing Tang*, Xinyu He†, Cewu Lu*†, Panpan Cai*†
*Shanghai Jiao Tong University {huangyongxi,cai_panpan}@sjtu.edu.cn
†Shanghai Innovation Institute
‡Beihang University

***Abstract*—For a mobile manipulator to act safely and effectively in human environments such as homes, hospitals, and workplaces, it must decide *where to look* to acquire the information a human-specified task demands. We study this problem as *language-conditioned active perception*: given a query image, a set of context images, and a natural language instruction, predict a 6D camera pose that fulfills the specified perception intent. We present I-Perceive, a foundation model that couples a vision-language pathway for semantic grounding with a geometric reasoning pathway for multi-view 3D understanding, connected through multi-layer semantic fusion so that language intent directly shapes geometric viewpoint reasoning. To enable scalable training, we build a large-scale dataset of language–viewpoint pairs from real-world scene scans and simulation through an automated pipeline. Across in-distribution and out-of-distribution indoor benchmarks, I-Perceive substantially outperforms next-best-view, vision-language-navigation, and VLM-prompting baselines in viewpoint accuracy, feasibility, and instruction alignment, while running with the lowest latency. The model generalizes zero-shot to unseen scenes and supports closed-loop active perception, progressively refining viewpoints over sequential interactions.**[1]

I. Introduction

A mobile manipulator operating alongside people must continually decide where to direct its camera in order to gather the information a task requires [1, 4, 5, 23]. When a person asks a robot to "check whether the stove is off" or "look inside the top drawer," the robot must translate an open-ended, human-centric request into a concrete viewpoint in a large, partially observed indoor scene. This capability—*active perception*—is a prerequisite for safe, legible, and effective operation in homes, hospitals, and workplaces.

Existing approaches are fragmented and ill-suited to this setting. Many methods hard-code the perception objective into a task-specific loss for detection, pose estimation, or reconstruction [7, 16, 17, 22], yielding isolated systems that do not generalize across intents. Others assume highly constrained, table-top workspaces [11, 15, 24, 25], far from the room-scale environments a mobile manipulator navigates; navigation-oriented approaches such as vision-language navigation [2, 6] select views only implicitly, optimizing goal-reaching rather than perception quality. We therefore target a foundation model for active perception that (1) generalizes to open-ended perception intents specified in natural language, and (2) scales to large indoor environments suitable for mobile manipulation.

This goal raises several coupled challenges: inferring intent from diverse, underspecified language; recovering scene geometry from sparse views (2D→3D); grounding language into both visual and geometric representations; and mapping the geometry of task-relevant regions to an actionable 6D camera pose. Recent advances offer only partial answers. Vision-Language Models (VLMs) [3, 10, 19] align language with images, and geometric foundation models such as VGGT [14, 20, 21] recover 3D structure and camera poses from sparse views. Yet these capabilities remain *disconnected*: no principled mechanism translates open-ended instructions into viewpoint decisions under partial observability.

To bridge this gap we propose I-Perceive, a foundation model for language-conditioned active perception that performs end-to-end viewpoint decision-making in large indoor scenes. Given a query image, context images, and an instruction, I-Perceive predicts a 6D camera pose fulfilling the intent. It integrates a *vision-language pathway* that extracts semantic intent and a *geometric reasoning pathway* that models multi-view spatial structure, tightly coupled via *cross-pathway fusion*. We cast viewpoint prediction as reasoning over a latent *target frame* whose learnable tokens participate in joint multi-view attention, while *semantic tokens* inject language-grounded features through multi-layer cross-attention. The model thus jointly reasons about *what to observe* and *where to observe it* in 3D. To train it, we construct 162K language–viewpoint samples from real-world scans [8, 13] and 70K from simulation [12] via a fully automated pipeline. Across in- and out-of-distribution indoor benchmarks, I-Perceive significantly outperforms next-best-view, vision-language-navigation, and VLM-prompting baselines in accuracy, feasibility, and task alignment, generalizes zero-shot to unseen scenes, and supports closed-loop refinement over sequential interactions.

II. The I-Perceive Model

a) Problem and overview.: I-Perceive maps multi-view scene observations and a language task to an executable camera pose. The input is a sequence of RGB frames and a natural language instruction ℓ specifying the perceptual intent. The first frame is the *query frame* (the robot's current view); the rest are *context frames* providing auxiliary scene

[1]Project page: https://iperceive1.github.io/

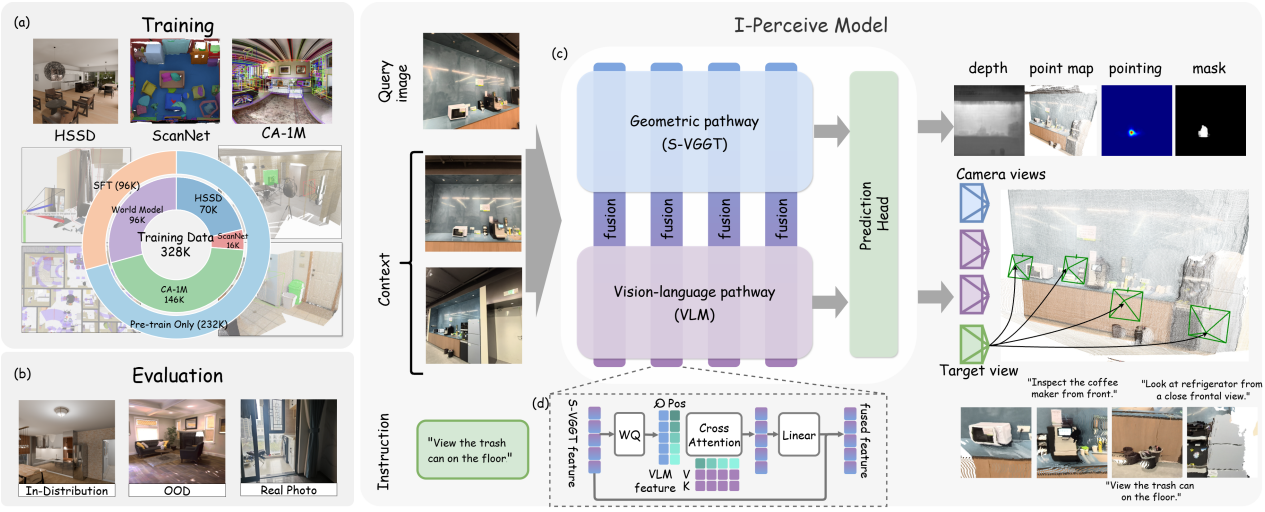


Fig. 1: **I-Perceive** is a foundation model for vision-language active perception. Given context images and a natural language instruction, the model predicts a target camera pose that fulfills the observation intent. By fusing VLM semantic features into a geometric backbone, I-Perceive shows strong zero-shot generalization to unseen real-world environments.

information. The model outputs a target camera pose that is physically feasible and faithful to the instruction. The core difficulty is jointly reasoning over language, visual observations, and 3D geometry under partial observability. I-Perceive addresses this with two complementary pathways whose semantic features are progressively injected into the geometric pathway via multi-layer cross-attention, enabling *language-conditioned geometric reasoning* (Fig. 1).

b) Input and output.: The model takes N RGB frames $\{I_i\}_{i=0}^{N-1}$ and an instruction ℓ, and predicts the target camera parameters $\mathbf{g}_N$ in the coordinate system rooted at I_0. Each camera is parameterized as $\mathbf{g}_i = [\mathbf{q}_i, \mathbf{t}_i, \mathbf{f}_i] \in \mathbb{R}^9$ (rotation quaternion, translation, field of view). During training the model optionally produces dense outputs (depth and point maps) and semantic masks (2D pointing and target masks) as auxiliary tasks; at inference only the camera-pose head is needed.

c) Vision-language pathway.: A pretrained VLM (Qwen3-VL-4B) encodes the images and instruction. Rather than using final-layer features, we extract intermediate key–value (KV) features from selected layers $\mathcal{L} = \{l_j\}_{j=1}^{N_Q}$:

$$K^{(l_j)}, V^{(l_j)} = \text{ExtractKV}(\text{VLM}, l_j, \mathcal{I}, \ell). \quad (1)$$

These multi-layer representations expose the VLM's progressive semantic reasoning, allowing smooth integration with progressive geometric reasoning.

d) Geometric reasoning pathway.: We adopt a multi-view transformer inspired by VGGT [20], which recovers 3D structure and camera poses from observations. Unlike VGGT, which estimates poses of *observed* views, our goal is to infer an *unseen*, task-driven viewpoint under partial observability. Each input frame is mapped to vision tokens v and camera tokens c, processed by attention layers alternating intra- and inter-frame attention to build an implicit 3D understanding. We add two key elements: *(i) Target-frame tokens*—a latent target frame I_T with learnable tokens that participate in the same alternating attention as observed frames, casting viewpoint decision-making as extrapolation from seen to unseen views; and *(ii) Semantic tokens* s, per-frame carriers of language-grounded information that, in intra-frame attention, interact with vision tokens and VLM features and then propagate globally via inter-frame attention.

e) Cross-pathway fusion.: Language information is repeatedly injected into semantic tokens via cross-attention:

$$s_i' = s_i + \gamma \cdot \text{Proj}\Big(\text{Attn}\Big(W_Q s_i, K^{(l)}, V^{(l)}\Big)\Big), \ \forall l \in \mathcal{L}. \quad (2)$$

Compared with shallow or late fusion, injecting semantics at multiple layers lets language intent progressively modulate spatial reasoning throughout the network. Final token representations are decoded by regression and dense/semantic heads; at inference, only camera-pose prediction is required (Fig. 2, left).

f) Training and dataset.: No public dataset directly supports language-conditioned active perception, so we build one via an automated pipeline that turns indoor scene datasets into active-perception tasks with minimal human intervention. Each sample is a sequence of $N+1$ frames—N observed and one target $\mathcal{F}_N = (I_N, D_N, M_N, \mathbf{g}_N)$ satisfying the intent—paired with an instruction ℓ. The pipeline (a) samples target frames with object-centric heuristics from dataset annotations; (b) samples query/context frames to emulate partial observation; (c) generates free-form instructions with a VLM conditioned on the target view and verbalized viewpoint attributes; and (d) disambiguates references to similar objects with a 2D point hint. A world-model step further refines ∼30% of CA-1M scenes for diversity. Steps (a)–(d) yield 162K real-world samples from ScanNet [8] and CA-1M [13] and 70K from HSSD [12]; the world-model step adds 96K post-training samples. We pre-train for 5 epochs with the VLM frozen and

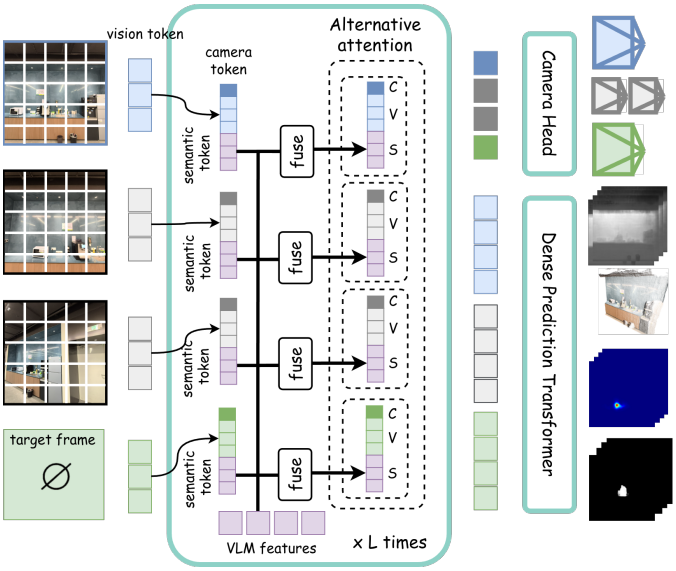


Fig. 2: A zoom-in view of I-Perceive's geometric-reasoning pathway: semantic tokens carry language-grounded VLM features that are fused into multi-view geometric attention to predict the latent target viewpoint.

all targets enabled, then fine-tune for 3 epochs, unfreezing the VLM with LoRA [9] and focusing on instruction following (30 hours on 32 H200 GPUs).

III. Experiments

a) Benchmarks.: We evaluate on both in-distribution and out-of-distribution indoor settings. *HSSD (in-distribution):* 52 novel tasks with scene-level splits where all test scenes are unseen during training; target viewpoints are annotated by human experts rather than the data pipeline. *Replica (out-of-distribution):* 89 novel language-conditioned tasks on reconstructed real-world scenes [18] for zero-shot evaluation, also expert-annotated.

b) Baselines.: We compare against three strong baselines. *(a) Language-Conditioned Next-Best-View (NBV+L):* samples candidate viewpoints around the target region and uses a VLM to select the best view by instruction alignment. *(b) Vision-Language Navigation with view selection (VLN+P):* navigates toward the target with NaVILA [6], then applies heuristic viewpoint orientation. *(c) VLM prompting:* prompts a general-purpose VLM to output a camera location and an image-space look-at point, leveraging VLMs' 2D pointing strength.

c) Metrics.: We report complementary metrics with 95% non-parametric bootstrap intervals over test cases: *Constraint Violation (CV)*, the fraction of predicted views that miss the target or collide; *Pose Error* (e_t, e_r), translation and rotation error against the annotated pose; *Voxel-Matching F1*, a tolerant content-centric score (a ground-truth voxel is matched if a predicted-view voxel lies within $\sim$10 cm); and *Judge Ranking*, a list-wise protocol where GPT-5.4 and human experts rank candidate views from different methods.

d) Absolute performance.: Fig. 3 summarizes performance on HSSD and Replica. Across both settings, I-Perceive achieves the best results at the lowest runtime: lower constraint violation and pose error, and markedly higher voxel-matching F1. Among baselines, VLN+P performs worst as its heuristic view selection ignores language intent, NBV+L is hampered by assuming fully observable targets, and VLM prompting gives reasonable task success but lower view quality at much higher latency. I-Perceive instead produces views highly consistent with human annotations and near-zero constraint violations.

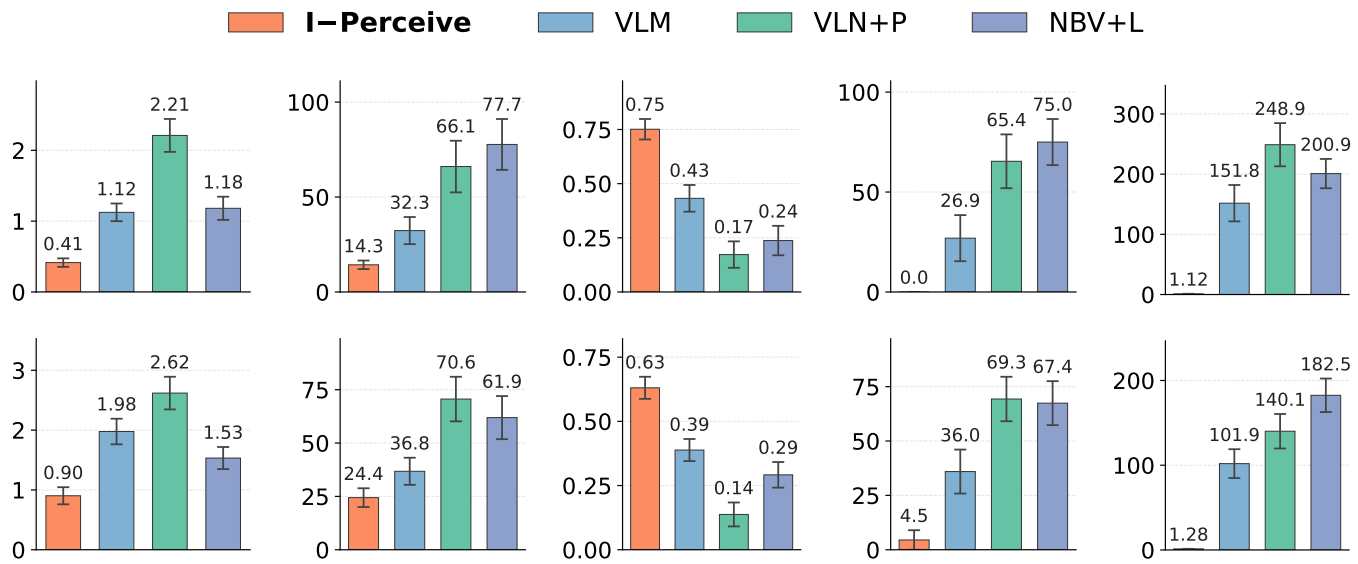
e) Relative performance.: Fig. 4 reports the list-wise judge-ranking protocol. I-Perceive is ranked first far more often than any baseline in both in- and out-of-distribution settings (Fig. 4a,b), and achieves pairwise winning rates of 87.5%–94.1% over all baselines (Fig. 4c,d). GPT-5.4 and human rankings are highly consistent (Spearman 0.875), indicating that the automatic judge aligns well with human preference. Because pose error over-commits to a single annotated pose under multi-modal valid viewpoints, this preference-based measure offers a more reliable relative assessment of viewpoint quality.

f) Inference time.: I-Perceive is also the most efficient method, requiring only a single forward pass for viewpoint prediction. Its mean inference time is 1.1 s on HSSD and 1.3 s on Replica, which involve on average 2.6 and 3.2 context frames per case, respectively—an order of magnitude faster than baselines that rely on repeated VLM calls, candidate evaluation, or online navigation rollouts.

g) Closed-loop active perception.: Fig. 5 evaluates I-Perceive with no initial context (a single start image) and targets initially out of view, so context must be gathered through closed-loop execution. I-Perceive steadily improves over successive steps and nearly matches the oracle performance with full context, despite never being explicitly trained for closed-loop operation—an emergent benefit of accepting variable-length context. This sequential, instruction-driven behavior is directly relevant to mobile manipulators that must explore human environments to satisfy a request.

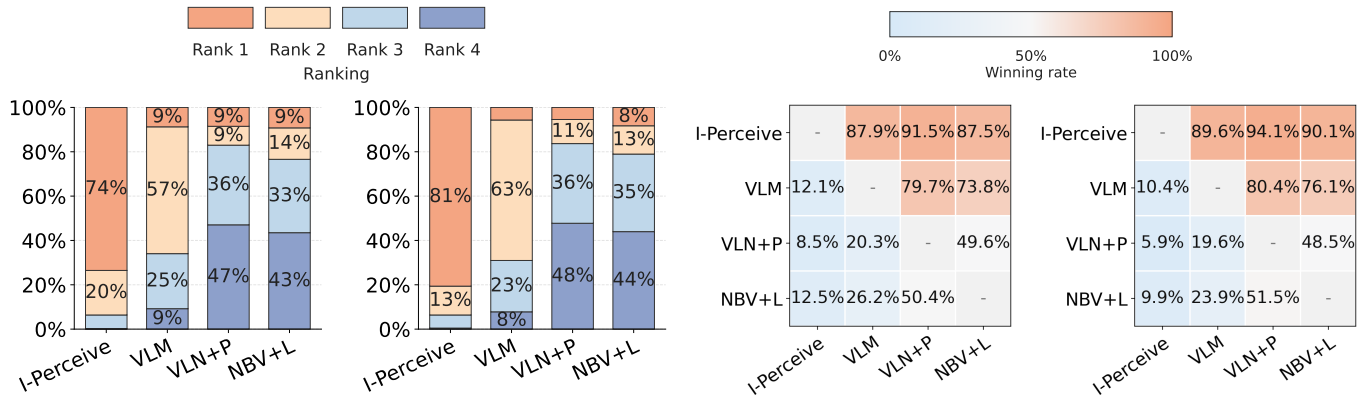
IV. Conclusion

We presented I-Perceive, a foundation model for language-conditioned active perception in large indoor environments. By injecting semantic grounding from vision-language models into multi-view geometric reasoning, I-Perceive predicts task-conditioned camera poses from visual context and natural language, outperforming next-best-view, VLN, and VLM-prompting baselines in accuracy, constraint satisfaction, preference, and runtime, and supporting closed-loop refinement. By turning open-ended human instructions into feasible viewpoints in room-scale scenes, I-Perceive is a step toward mobile manipulators that perceive human environments on demand. Our current constraint checking detects collisions but does not correct them; future work will incorporate collision-aware viewpoint optimization and tighter



(a) $e_t(m) \downarrow$ (b) $e_r(deg) \downarrow$ (c) $F1 \uparrow$ (d) $CV(\%) \downarrow$ (e) Runtime $(s) \downarrow$

Fig. 3: Absolute performance of I-Perceive and baselines on HSSD (upper) and Replica (lower). I-Perceive attains the best accuracy, feasibility, and F1 at the lowest latency across both settings.



(a) GPT ranking (b) Human ranking (c) GPT winning rate (d) Human winning rate

Fig. 4: Judge-ranking results of I-Perceive and baselines on HSSD and Replica. (a,b) rank distributions assigned by GPT and human judges (rank 1 is best); (c,d) pairwise winning-rate matrices (row vs. column).

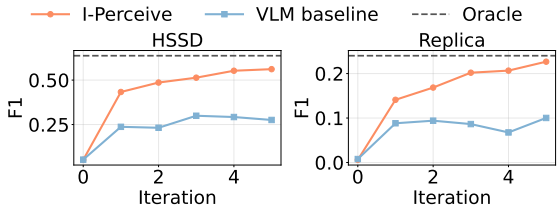


Fig. 5: Closed-loop execution progressively improves viewpoint quality across steps, approaching the oracle that observes full context. Starting from a single image with the target out of view, I-Perceive gathers context through successive predictions.

integration with downstream mobile-manipulation planning and control.

References

[1] John Aloimonos, Isaac Weiss, and Amit Bandyopadhyay. Active vision. *International journal of computer vision*, 1(4):333–356, 1988.

[2] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683, 2018.

[3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei

Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[4] Ruzena Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):966–1005, 1988.

[5] Ruzena Bajcsy, Yiannis Aloimonos, and John K Tsotsos. Revisiting active perception. *Autonomous Robots*, 42(2): 177–196, 2018.

[6] A.-C. Cheng et al. Navila: Legged robot vision-language-action model for navigation. *arXiv:2412.04453*, 2024.

[7] Cl Connolly. The determination of next best views. In *Proceedings. 1985 IEEE international conference on robotics and automation*, volume 2, pages 432–435. IEEE, 1985.

[8] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

[10] Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic vlms: Investigating the design space of visually-conditioned language models. In *Forty-first International Conference on Machine Learning*, 2024.

[11] Justin Kerr, Kush Hari, Ethan Weber, Chung Min Kim, Brent Yi, Tyler Bonnen, Ken Goldberg, and Angjoo Kanazawa. Eye, robot: Learning to look to act with a bc-rl perception-action loop. *arXiv preprint arXiv:2506.10968*, 2025.

[12] Mukul Khanna*, Yongsen Mao*, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X. Chang, and Manolis Savva. Habitat Synthetic Scenes Dataset (HSSD-200): An Analysis of 3D Scene Scale and Realism Tradeoffs for ObjectGoal Navigation. *arXiv preprint*, 2023.

[13] Justin Lazarow, David Griffiths, Gefen Kohavi, Francisco Crespo, and Afshin Dehghan. Cubify anything: Scaling indoor 3d object detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22225–22233, 2025.

[14] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024.

[15] Mengzhen Liu, Enshen Zhou, Cheng Chi, Yi Han, Shanyu Rong, Liming Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Sapave: Towards active perception and manipulation in vision-language-action models for robotics. *arXiv preprint arXiv:2603.12193*, 2026.

[16] Miguel Mendoza, J Irving Vasquez-Gomez, Hind Taud, L Enrique Sucar, and Carolina Reta. Supervised learning of the next-best-view for 3d object reconstruction. *Pattern Recognition Letters*, 133:224–231, 2020.

[17] Juil Sock, S Hamidreza Kasaei, Luis Seabra Lopes, and Tae-Kyun Kim. Multi-view 6d object pose estimation and camera motion planning using rgbd images. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2228–2235, 2017.

[18] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

[19] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[20] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025.

[21] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024.

[22] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.

[23] Rui Zeng, Yuhui Wen, Wang Zhao, and Yong-Jin Liu. View planning in robot active vision: A survey of systems, algorithms, and applications. *Computational Visual Media*, 6(3):225–245, 2020.

[24] Muzhi Zhu, Hao Zhong, Canyu Zhao, Zongze Du, Zheng Huang, Mingyu Liu, Hao Chen, Cheng Zou, Jingdong Chen, Ming Yang, et al. Active-o3: Empowering multimodal large language models with active perception via grpo. *arXiv preprint arXiv:2505.21457*, 2025.

[25] Yanwen Zou, Chenyang Shi, Wenye Yu, Han Xue, Jun Lv, Ye Pan, Chuan Wen, and Cewu Lu. Activeglasses: Learning manipulation with active vision from ego-centric human demonstration. *arXiv preprint arXiv:2604.08534*, 2026.