Spatial Reasoning in Foundation Models: Benchmarking Object-Centric Spatial Understanding

Vahid Mirjalili[*†] **Ramin Giahi**[*] **Sriram Kollipara**[*] **Akshay Kekuda**[*] **Kehui Yao**[*]

Kai Zhao[*‡] **Jianpeng Xu** **Kaushiki Nag** **Sinduja Subramaniam** **Topojoy Biswas**

Evren Korpeoglu **Kannan Achan**

Abstract

Spatial understanding is a critical capability for vision foundation models. While recent advances in large vision models or vision–language models (VLMs) have expanded recognition capabilities, most benchmarks emphasize localization accuracy rather than whether models capture how objects are arranged and related within a scene. This gap is consequential: effective scene understanding requires not only identifying objects, but reasoning about their relative positions, groupings, and depth. In this paper, we present a systematic benchmark for object-centric spatial reasoning in foundation models. Using a controlled synthetic dataset, we evaluate state-of-the-art vision models (e.g., GroundingDINO, Florence-2, OWLv2) and large VLMs (e.g., InternVL, LLaVA, GPT-4o) across three tasks: spatial localization, spatial reasoning, and downstream retrieval tasks. We find a stable trade-off: detectors such as GroundingDINO and OWLv2 deliver precise boxes with limited relational reasoning, while VLMs like SmolVLM and GPT-4o provide coarse layout cues and fluent captions but struggle with fine-grained spatial context. Our study highlights the gap between localization and true spatial understanding, and pointing toward the need for spatially-aware foundation models in the community.

1 Introduction

Understanding and reasoning about spatial relationships between objects is a core challenge for vision–language systems and embodied AI [45]. Despite progress in object detection [42, 43] and open-vocabulary recognition [17, 37], most benchmarks emphasize localization, identifying, and bounding objects [41, 10, 45], rather than contextual spatial reasoning. Yet, applications from embodied interaction to e-commerce recommendation [38, 48, 56, 33] require models to detect objects and interpret how they are arranged and functionally related within a scene. In shopping images, for example, a sofa may appear with a coffee table, rug, and lamp. A useful system must capture relative position, containment, and grouping: a sofa next to a coffee table implies different recommendations than a sofa isolated against a wall. Without such context, results risk incoherence and poor alignment with user intent.

State-of-the-art vision models (e.g., GroundingDINO [44], Florence-2 [49], OWLv2 [36]) and large VLMs (e.g., InternVL [5], LLaVA [31], GPT-4o [18]) broaden what can be localized in complex scenes, but their spatial reasoning ability remains underexplored. A model that detects both a "sofa"

[*]Equal contributions.

[†]All authors are from Walmart Global Tech, USA

[‡]Corresponding author: kai.zhao@walmart.com

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: SPACE in Vision, Language, and Embodied AI.

and a "lamp" may still fail to infer their relative depth, errors that harm multi-item recommendation, scene retrieval, and embodied planning. We present a systematic benchmarking study of spatial reasoning in vision models and VLMs on a synthetic shopping-scene data. Our contribution lies in the evaluation protocol and analysis rather than a new dataset: we probe whether state-of-the-art systems can (i) localize a focal product in clutter, (ii) capture its spatial relations with surrounding items, and (iii) leverage these relations for retrieval and recommendation via a unified detection-to-retrieval pipeline and spatial-localization.

Our results show a consistent gap: task-specific (e.g., object detection) vision models such as GroundingDINO [44] achieve high localization precision, but lack spatial reasoning capability. On the other hand, large VLMs like GPT-4o [18] produce descriptive captions and coarse layouts yet underperform when fine-grained spatial context is required. These results reveal a persistent divide between precise localization and true spatial understanding. Our benchmark surfaces this gap and provides standardized tasks and metrics, establishing a foundation for developing spatially aware foundation models that unify detection accuracy with contextual understanding. We release complete details of the benchmark, covering data generation, evaluation protocol, prompts, metrics, and failure analyses, to ensure reproducibility and extension to other domains.

2 Experiments

2.1 Datasets

To benchmark spatial reasoning capability of vision models and VLMs, we build a unique synthetic dataset where both localization and relevance are known by construction, allowing us to control viewpoint, clutter, and scene context. Our synthetic dataset spans across nine furniture categories: *Bed*, *Chair*, *Cabinet*, *Desk*, *Dresser*, *Planter*, *Shelf*, *Sofa*, and *Vase*. Images are created by rendering 3D products with random rotation, shift, and scale, to increase variety. The rendered images are then composited with background scenes. For each product, we split renders into database (DB) images that contain frontal views and query images that contain more angled views. Details of the data-generation pipeline are given in Appendix A. The synthesized dataset has roughly 11k images, split into 3k DB images and remaining 8k are served as query images. For retrieval, each query is paired with its matching DB image from the same product, which we use as ground-truth relevance. The statistics per category are reported in the Appendix Table 3.

2.2 Benchmark Models

We evaluate 14 models grouped into two families:

- Task-specific (e.g., object detection) vision models include D-FINE[39], OWL-ViT[37], OWLv2[36], Florence-2[53], GroundingDINO[44], and (LISA-7B/13B)[23].
- General-purpose VLMs which include SmolVLM[35], InternVL[5], LLaVA[29], LLaVA-OneVision[26], LLaVA-Next[30], Gemini 2.5[7], and GPT-4o[18].

We run three experiment sets: (i) spatial localization, (ii) spatial reasoning, and (iii) image-based retrieval. For spatial localization, we extract the focal product's bounding box using both vision models and VLMs. Spatial reasoning includes two tasks: predicting the focal object's position on a coarse grid (with 2×2 and 3×3 settings) and predicting coarse depth as front vs. back (VLM-only). For retrieval, we compute VL-CLIP [14] embeddings over crops from each method's predicted boxes on query images, and search an HNSW [34] index built on DB-image embeddings. Full setup and metrics are given in Appendix B.2.

2.3 Main Results

In general, for spatial localization, reasoning, and retrieval experiments, task-specific vision models outperform general-purpose VLMs by a clear margin. GroundingDINO is the strongest overall, leading both spatial localization and image-based retrieval. Among VLMs, InternVL and LLaVA variants are the most competitive, yet they still underperform the best vision models consistently across spatial localization, spatial reasoning, and retrieval.

Table 1 reports spatial localization performance for coarse grid-cell prediction using accuracy, macro-F_1, and MCC metrics. Among task-specific vision models, GroundingDINO is consistently strongest, while LISA-7B ranks lowest in this group. Among the VLMs, LLaVA-OneVision and InternVL are the top performers, whereas GPT-4o and Gemini 2.5 show the lowest performance overall. Comparing across groups, vision models clearly outperform VLMs on both grid settings, and the performance gap widens as the grid becomes smaller.

Figure 1 illustrates typical spatial localization errors for VLMs. Across task-specific vision models, failures are dominated by box-placement issues (shifted or mis-sized extents) in cluttered or low-contrast scenes; while category confusions are comparatively rare. On the other hand, failures in VLMs are more often due to coarse spatial grounding, yielding off-center or overly loose boxes. This observation supports weaker box-level supervision in VLMs relative to vision models.

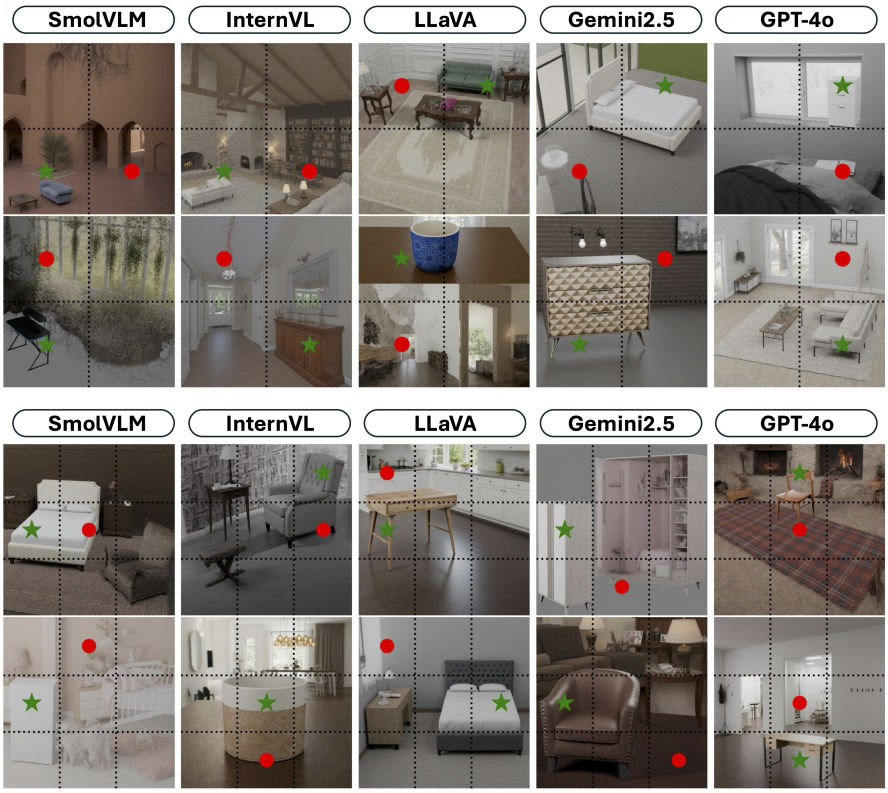

Figure 1: **Spatial localization failures of VLMs.** The GT cell is indicated by a *green star*, and the predicted cell by a *red circle* centered in the chosen cell.

2.4 Spatial Reasoning Capability of VLMs

We further analyze the spatial reasoning capability of VLMs for context-aware scene understanding, by evaluating coarse depth ordering of the focal object relative to its surrounding environment. For this experiment, VLMs are prompted for *front-vs-back* classification (constrained to respond with `front` or `back`). This experiment assesses whether the model can determine if the focal object lies in the foreground (front) or background (back) of the overall scene by considering the relative depth and layering of objects. We report accuracy, precision, recall, and F_1 in Table 2. The results show that InternVL is the best-performing model for coarse-depth prediction, achieving the highest accuracy and F_1 score. While SmolVLM also shows strong performance, other models exhibit significant trade-offs between precision and recall. LLaVA-OneVision achieves near-perfect recall, indicating it correctly identifies almost every background object, though its precision is lower. In contrast, LLaVA-Next has extremely high precision but low recall. Performance drops notably for the last two models, with GPT-4o showing the weakest results, particularly in recall and F_1 score. For visualization purpose, we show failed cases in Figure 2.

Table 1: Spatial localization results for vision models and VLMs, predicting the location of the focal object in an image on a coarse grid

	Model	2×2 grid			3×3 grid		
		Acc ↑	F_1^{macro} ↑	MCC ↑	Acc ↑	F_1^{macro} ↑	MCC ↑
Task-specific vision models	GroundingDINO-1.5	**0.816**	**0.815**	**0.754**	**0.799**	**0.759**	**0.742**
	Florence2-base	0.808	0.808	0.745	0.797	0.754	0.740
	D-FINE	0.799	0.799	0.732	0.789	0.742	0.726
	OWLv2-base-16	0.793	0.792	0.724	0.764	0.715	0.702
	LISA-13B-Llama	0.785	0.785	0.713	0.750	0.690	0.684
	OWL-ViT-base-32	0.759	0.759	0.679	0.728	0.677	0.655
	LISA-7B	0.733	0.732	0.645	0.694	0.628	0.614
General-purpose VLMs	InternVL3-8B	0.643	0.640	0.530	**0.565**	**0.494**	**0.452**
	LLaVA-OneVision	**0.645**	**0.643**	**0.538**	0.548	0.390	0.400
	LLaVA-Next	0.551	0.544	0.415	0.429	0.239	0.244
	LLaVA-1.5-7B	0.492	0.424	0.349	0.055	0.026	0.024
	SmolVLM2-2.2B-Inst.	0.422	0.360	0.253	0.355	0.106	0.071
	Gemini 2.5-Pro	0.262	0.262	0.016	0.241	0.120	0.017
	GPT-4o	0.252	0.199	0.003	0.315	0.088	0.015

Table 2: Assessing spatial reasoning capability of VLMs on predicting coarse-depth ordering of the focal object in an image.

Model	Acc ↑	Prec ↑	Rec ↑	F1 ↑
InternVL3-8B	**0.874**	0.895	0.949	**0.921**
SmolVLM2-2.2B-Instruct	0.854	0.863	0.964	0.911
LLaVA-OneVision	0.832	0.823	**0.998**	0.902
LLaVA-1.5-7B	0.745	0.907	0.748	0.820
LLaVA-Next	0.634	**0.996**	0.530	0.692
Gemini 2.5-Pro	0.677	0.870	0.697	0.774
GPT-4o	0.455	0.931	0.322	0.478

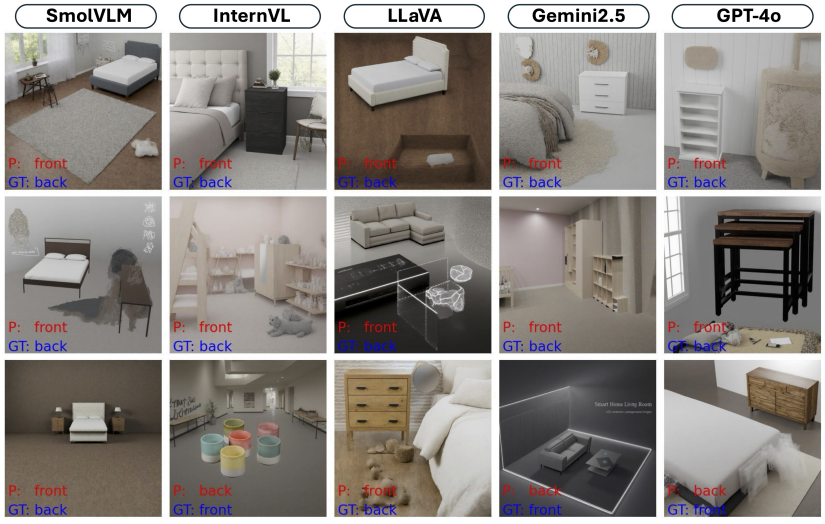

Figure 2: Spatial reasoning failures of VLMs. Annotations: P is what the VLM thinks about the relative position of the item, GT is the ground-truth coarse depth of the item.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[2] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. OpenFlamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

[3] Blender Online Community. Blender - a 3D modelling and rendering package.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[5] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.

[6] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. YOLO-World: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16901–16911, 2024.

[7] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[8] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

[9] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic Head: Unifying object detection heads with attentions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7373–7382, 2021.

[10] Muhammad Sohail Danish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Kartik Kuckreja, Fahad Shahbaz Khan, Paolo Fraccaro, Alexandre Lacoste, and Salman Khan. Geobench-vlm: Benchmarking vision-language models for geospatial tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[12] Zezhong Fan, Xiaohan Li, Luyi Ma, Kai Zhao, Liang Peng, Topojoy Biswas, Evren Korpeoglu, Kaushiki Nag, and Kannan Achan. LayoutAgent: A vision-language agent guided compositional diffusion for spatial layout planning. *arXiv preprint arXiv:2509.22720*, 2025.

[13] Najmeh Forouzandehmehr, Reza Yousefi Maragheh, Sriram Kollipara, Kai Zhao, Topojoy Biswas, Evren Korpeoglu, and Kannan Achan. Cal-rag: Retrieval-augmented multi-agent generation for content-aware layout design. *arXiv preprint arXiv:2506.21934*, 2025.

[14] Ramin Giahi, Kehui Yao, Sriram Kollipara, Kai Zhao, Vahid Mirjalili, Jianpeng Xu, Topojoy Biswas, Evren Korpeoglu, and Kannan Achan. VL-CLIP: Enhancing multimodal recommendations via visual grounding and LLM-augmented CLIP embeddings. *arXiv preprint arXiv:2507.17080*, 2025.

[15] Ross Girshick. Fast R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[16] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[17] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

[18] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[19] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[20] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLOv8, 2023.

[21] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-VLM: Open-vocabulary object detection upon frozen vision and language models. *arXiv preprint arXiv:2209.15639*, 2022.

[22] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025.

[23] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision*

and Pattern Recognition, pages 9579–9589, 2024.

[24] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024.

[25] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37:87874–87907, 2024.

[26] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[28] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975, 2022.

[29] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024.

[30] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, ocr, and world knowledge, January 2024.

[31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[32] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding DINO: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024.

[33] Luyi Ma, Wanjia Zhang, Kai Zhao, Abhishek Kulkarni, Lalitesh Morishetti, Anjana Ganesh, Ashish Ranjan, Aashika Padmanabhan, Jianpeng Xu, Jason HD Cho, et al. Grace: Generative recommendation via journey-aware sparse attention on chain-of-thought tokenization. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, pages 135–144, 2025.

[34] Yu A Malkov and Dmitry A Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836, 2018.

[35] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. SmolVLM: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

[36] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36:72983–73007, 2023.

[37] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European conference on computer vision*, pages 728–755. Springer, 2022.

[38] Nikil Pancha, Andrew Zhai, Jure Leskovec, and Charles Rosenberg. PinnerFormer: Sequence modeling for user representation at pinterest. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 3702–3712, 2022.

[39] Yansong Peng, Hebei Li, Peixi Wu, Yueyi Zhang, Xiaoyan Sun, and Feng Wu. D-FINE: Redefine regression task in detrs as fine-grained distribution refinement. *arXiv preprint arXiv:2410.13842*, 2024.

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[41] Rahul Ramachandran, Ali Garjani, Roman Bachmann, Andrei Atanov, Oğuzhan Fatih Kar, and Amir Zamir. How well does GPT-4o understand vision? evaluating multimodal foundation models on standard computer vision tasks. *arXiv preprint arXiv:2507.01955*, 2025.

[42] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[43] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

[44] Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, et al. Grounding DINO 1.5: Advance the" edge" of open-set object detection. *arXiv preprint arXiv:2405.10300*, 2024.

[45] Ilias Stogiannidis, Steven McDonagh, and Sotirios A Tsaftaris. Mind the gap: Benchmarking spatial reasoning in vision-language models. *arXiv preprint arXiv:2503.19707*, 2025.

[46] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural*

Language Processing (EMNLP-IJCNLP), pages 5100–5111, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[47] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[48] Jiahao Tian, Zhenkai Wang, Jinman Zhao, and Zhicheng Ding. Mmrec: Llm based multi-modal recommender system. In *2024 19th International Workshop on Semantic and Social Media Adaptation & Personalization (SMAP)*, pages 105–110. IEEE, 2024.

[49] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829, 2024.

[50] Lewei Yao, Jianhua Han, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, and Hang Xu. DetCLIPv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23497–23506, 2023.

[51] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. DetCLIP: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *Advances in Neural Information Processing Systems*, 35:9125–9138, 2022.

[52] Lewei Yao, Renjie Pi, Jianhua Han, Xiaodan Liang, Hang Xu, Wei Zhang, Zhenguo Li, and Dan Xu. DetCLIPv3: Towards versatile generative open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 27391–27401, 2024.

[53] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

[54] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

[55] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. GLIPv2: Unifying localization and vision-language understanding. *Advances in Neural Information Processing Systems*, 35:36067–36080, 2022.

[56] Tao Zhang, Kehui Yao, Luyi Ma, Reza Yousefi Maragheh, Jiao Chen, Kai Zhao, Jianpeng Xu, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. No-human in the loop: Agentic evaluation at scale for recommendation. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*.

[57] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. RegionCLIP: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16793–16803, 2022.

[58] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European conference on computer vision*, pages 350–368. Springer, 2022.

Spatial Reasoning in Foundation Models: Benchmarking Object-Centric Spatial Understanding APPENDIX

A Data Generation

A.1 Data Generation Pipeline

Existing public datasets rarely pair precise box-level ground truth with product-level relevance for the same scenes. To address this, we construct a synthetic dataset in which both localization and relevance are known by construction, allowing controlled variation in viewpoint, clutter, and context.

Recently, LayoutAgent [12] proposed using vision-language models (VLMs) for layout planning and spatial realism [13]. Inspired by LayoutAgent, we designed a data-generation pipeline to synthesize product–in–scene images by compositing 3D product renders into text-conditioned backgrounds (see Fig. 3). For each environment type (e.g., *living room*, *bedroom*, *patio*), we generate short textual scene descriptions using GPT-4 and split them so that database (DB) and query images use disjoint descriptions. Each product's 3D asset is rendered in Blender [3] under random rotations to obtain a clean foreground mask; we apply random scale/shift augmentations and export the exact 2D GT bounding box from the transformed mask. The augmented foreground and a sampled scene description are fed to FLUX-Kontext [22] to generate photorealistic composites.

DB-query pairs that are synthesized from the same source item form ground-truth correspondences (see examples in Fig. 4).

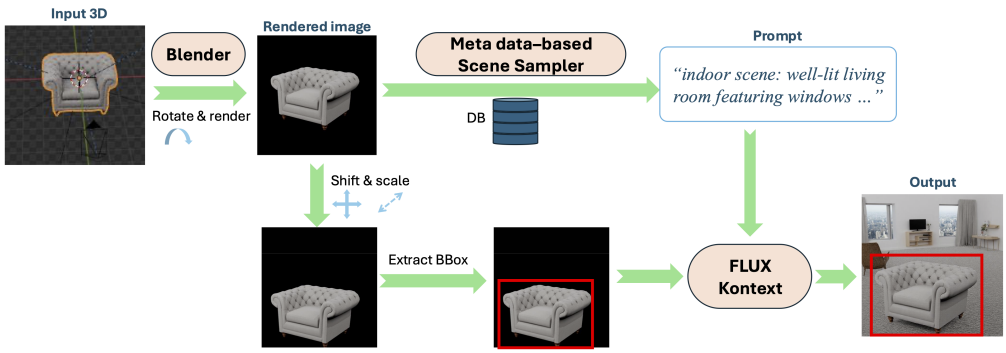

Figure 3: Data generation pipeline: generating synthetic product-in-scene images from a product's 3D asset based on a textual scene description sampled from a database.

A.2 Dataset Statistics

Table 3 reports per-category counts of unique items (*samples*) and resulting images, partitioned into database (DB) and query sets produced by our data-generation pipeline.

A.3 Synthetic Data Samples

In this section, we show a few examples generated through our designed pipeline. Figures 5, 6, 7 and 8 show data generated from different 3D objects in a multitude of everyday environments.

Table 3: Dataset summary by category. Each item is generated via our data-generation pipeline and split into database (DB) and query images.

Category	Samples	Images	DB images	Query images
Bed	22	2939	1128	1811
Cabinet	100	1007	201	806
Chair	100	1005	201	804
Desk	100	1001	200	801
Dresser	62	624	272	352
Planter	135	1349	667	682
Shelf	100	1001	200	801
Sofa	22	1652	630	1022
Vase	50	501	250	251
Total	**691**	**11079**	**3749**	**7330**

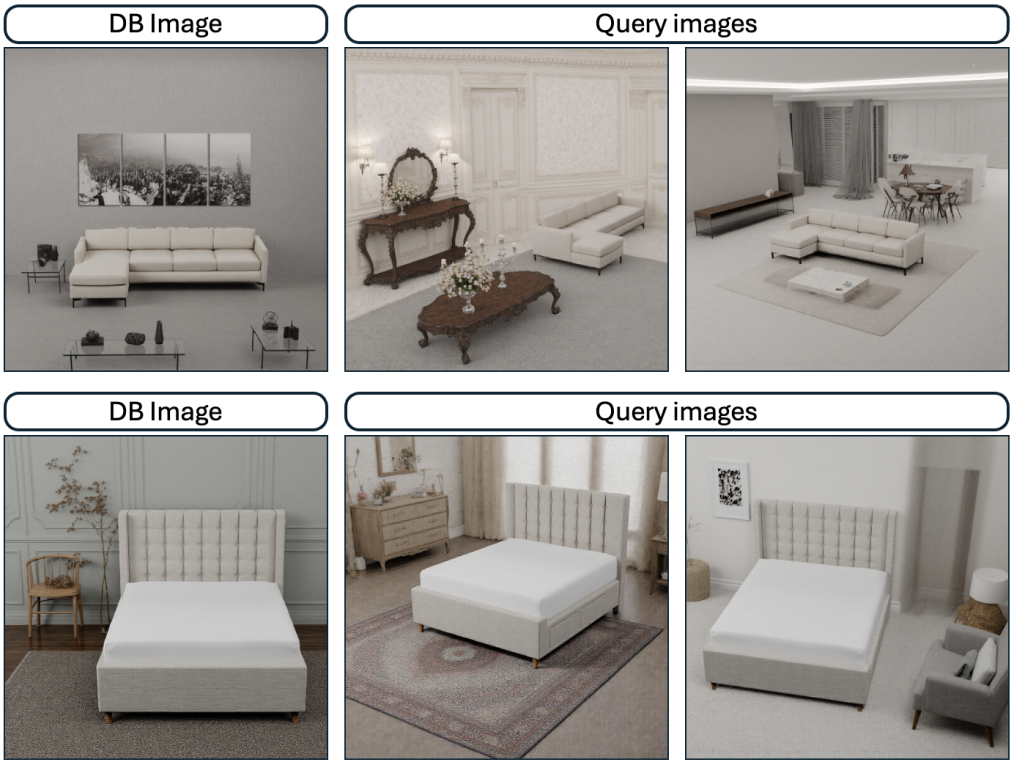

Figure 4: Generated samples are split into DB images (left) and query images (right).

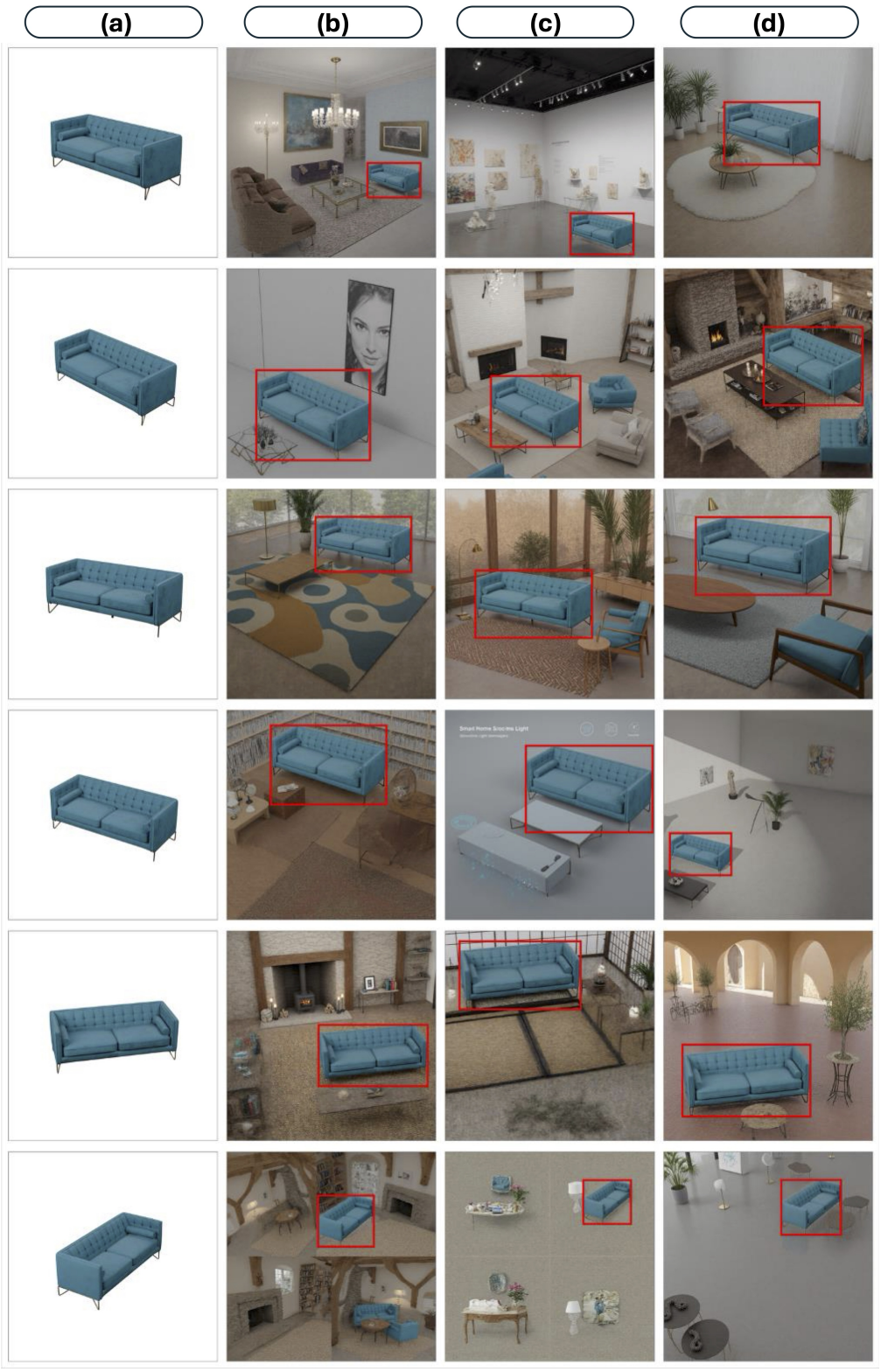

Figure 5: Generated data for a teal upholstered sofa with button-tufted back cushions and metal legs. Column (a) shows the original object in different rotations in isolation, while columns (b), (c) and (d) present the same object in the corresponding rotation (same row) scaled, shifted and integrated into various indoor scenes. The scenes include diverse residential and commercial environments such as modern living rooms, art galleries, minimalist spaces, rustic interiors with fire places, and contemporary office settings. Columns (b), (c), (d) also have the ground truth bounding box shown in red.

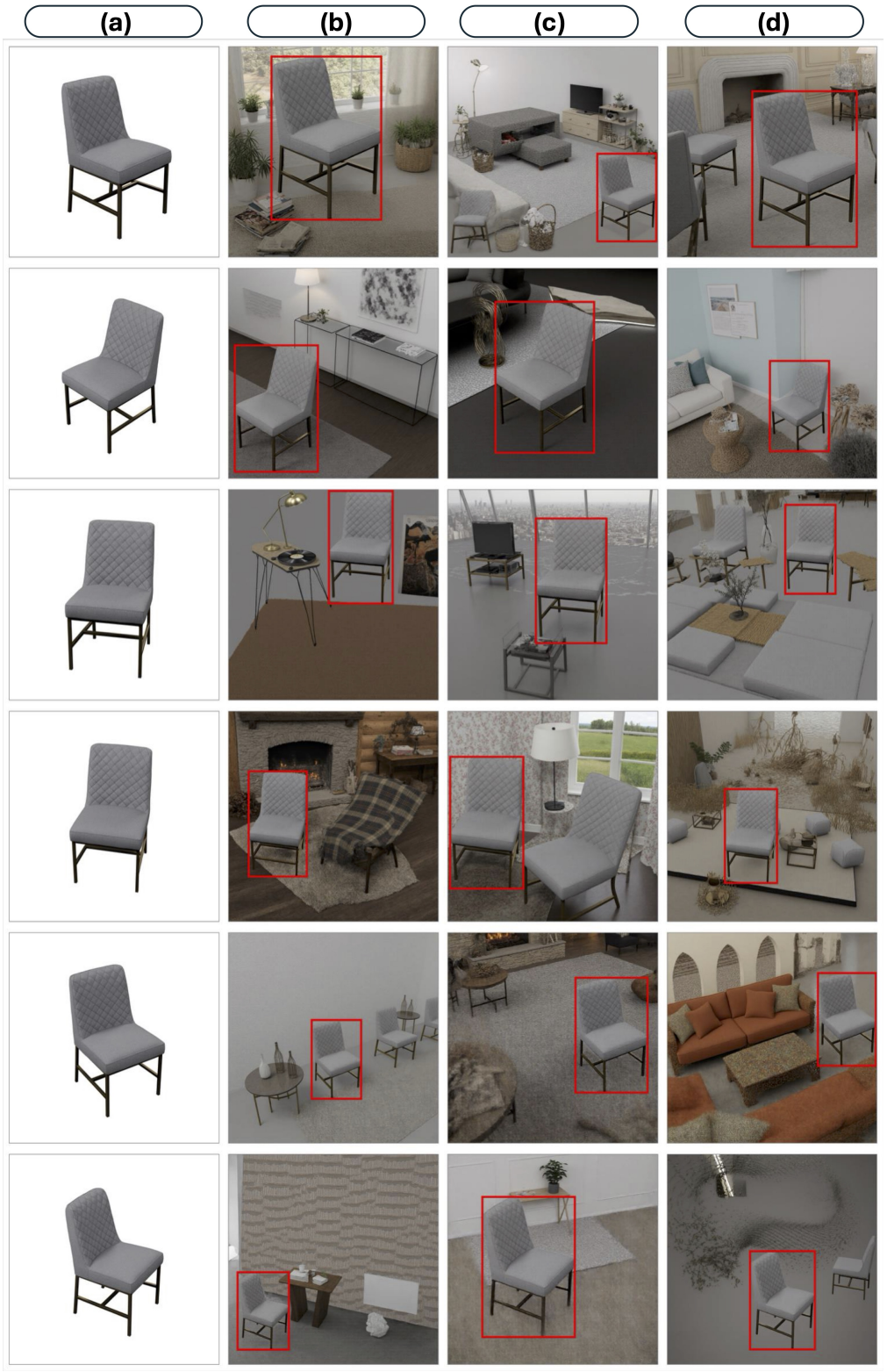

Figure 6: Generated data for a grey upholstered dining chair with quilted diamond-pattern backrest and black metal frame. Column (a) shows the original object in different rotations in isolation, while columns (b), (c) and (d) present the same object in the corresponding rotation (same row) scaled, shifted and integrated into various indoor scenes. The scenes include diverse residential and commercial environments such as modern living rooms, minimalist offices, rustic interiors with fireplaces, contemporary dining spaces, and gallery-like settings. Columns (b), (c), (d) also have the ground truth bounding box shown in red.

Figure 7: Generated data for a wooden sideboard with three drawers and black metal geometric frame base. Column (a) shows the original object in different rotations in isolation, while columns (b), (c) and (d) present the same object in the corresponding rotation (same row) scaled, shifted and integrated into various indoor scenes. The scenes include diverse residential and commercial environments such as modern living rooms, home offices with bookshelves, gallery spaces, contemporary bedrooms, minimalist interiors, and traditional spaces with brick walls. Columns (b), (c), (d) also have the ground truth bounding box shown in red.

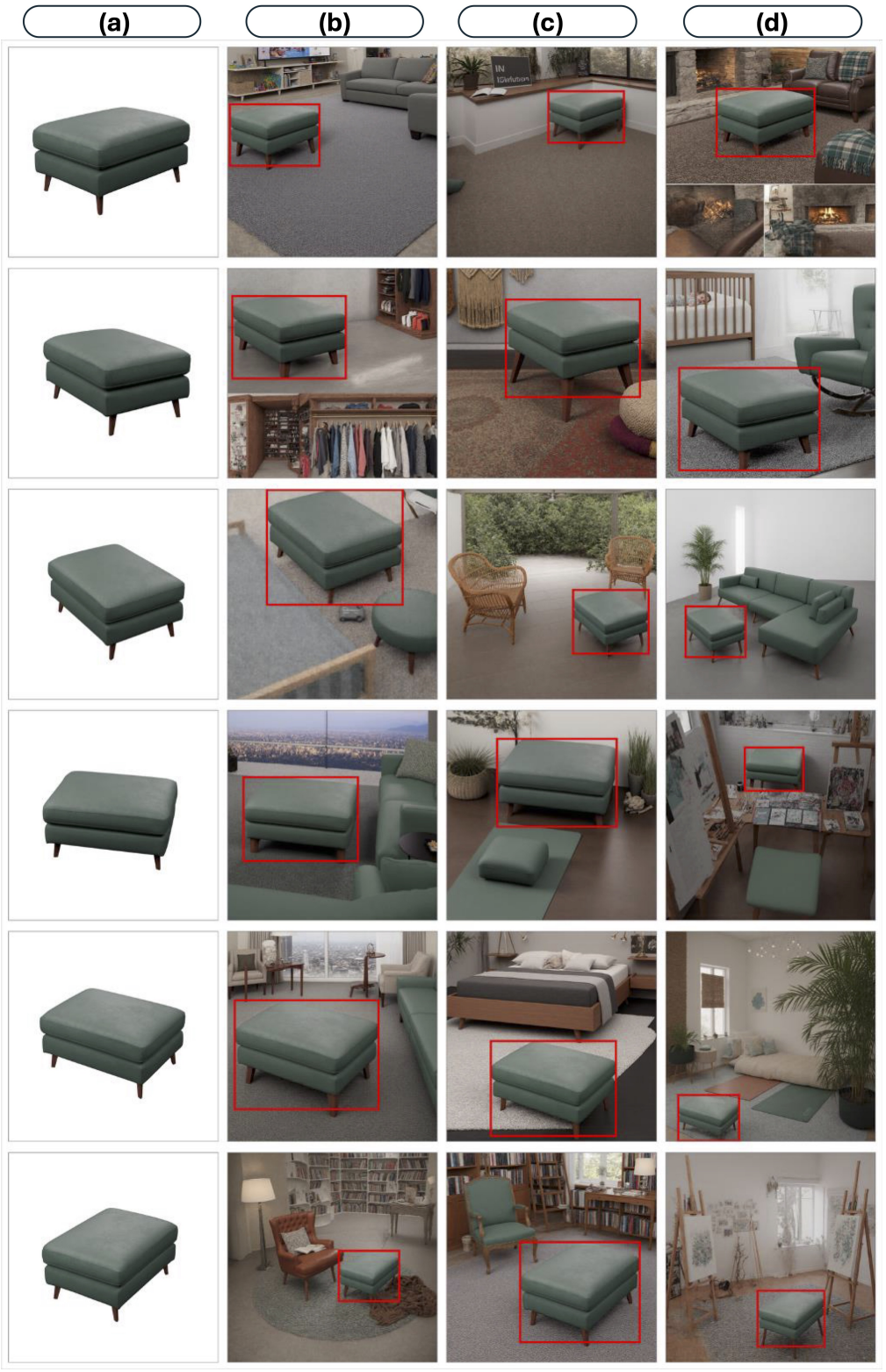

Figure 8: Generated data for a sage green upholstered ottoman with wooden legs. Column (a) shows the original object in different rotations in isolation, while columns (b), (c) and (d) present the same object in the corresponding rotation (same row) scaled, shifted and integrated into various indoor scenes. The scenes include diverse residential environments such as modern living rooms, cozy bedrooms, contemporary home offices, outdoor terraces, bohemian-style spaces with macramé decor, and traditional libraries with bookshelves. Columns (b), (c), (d) also have the ground truth bounding box shown in red.

B Methodology

B.1 Model Groups

We evaluate two model families distinguished by output type and prompting interface: Task-specific vision models, and general-purpose VLMs. Table 4 provides a summary of model properties and prompting inference for the selected models.

Table 4: Models evaluated in this study. (REC = referring expression comprehension).

	Model	Family / Style	Inference	Year
Task-specific vision models	D-FINE	Strong non-OVD detector	non-REC	2024
	OWL-ViT-base-32	Promptable OVD	Word	2022
	OWLv2-base-16	Promptable OVD (self-training)	Word	2023
	Florence-2-base	Foundation model used for detection	Sentence	2024
	Grounding DINO	Grounding-style pretraining	Word	2023
	LISA-7B	Reasoning Segmentation	Sentence	2023
	LISA-13B-Llama	Reasoning Segmentation	Sentence	2023
General-purpose VLMs	LLaVA-1.5-7B	Open LVLM	Sentence	2023
	LLaVA-Next-7B	Open LVLM	Sentence	2024
	LLaVA-OneVision-7B-si	Open LVLM	Sentence	2024
	SmolVLM2-2.2B-Instruct	Open LVLM (efficient)	Sentence	2024
	InternVL3-8B	Open LVLM	Sentence	2023
	Gemini 2.5-Pro	API LVLM	Sentence	2025
	GPT-4o	API LVLM	Sentence	2024

B.2 Evaluation Protocol and Metrics

We designed the following complementary tasks for benchmarking the performance of task-specific vision models and VLMs:

- **Localization (box-based):** Given an image and target name (e.g., *sofa*), predict one box for the main product; task-specific models directly output class-conditioned bounding box, while VLMs are prompted in REC style to output a box (first valid output is used). Evaluation metrics include $mIoU$, $AP_{0.5}$, and $AP_{0.75}$, $AP_{0.5:0.95}$.
- **Localization (coarse grid):** We overlay a 2×2 or 3×3 grid and define the GT cell by majority overlap with the GT box. VLMs return a cell index/tag; For task=specific vision models, top-1 predicted boxes are converted by assigning the cell with maximum overlap. We use multiclass metrics: accuracy, macro-F_1, and multiclass MCC.
- **Spatial reasoning – coarse depth ordering:** VLMs are prompted to answer with {`front`, `back`} under a constrained prompt whether the focal object is in front or back of the scene; we report accuracy, precision, recall, and F_1.
- **Patch-based retrieval (downstream):** We crop the predicted box (clipped to bounds), embed the crop with VL-CLIP [14], and query an HNSW index [34] built over DB embeddings to return top-k candidates. Models are compared using Precision@k and Hit@k for $k \in \{1, 2, 3\}$, based on the DB–query GT correspondences obtained through the data synthesis pipeline.

C Results

C.1 Localization

Table 5 provides the box-based localization performance for both task-specific vision models and general-purpose VLMs. Among task-specific models, GroundingDINO leads, followed by Florence-2 and OWLv2. Notably, while D-FINE attains strong mIoU, its AP is the lowest in this group, indicating a calibration gap between box quality and confidence scores under our setup.

Among general-purpose VLMs, InternVL is the strongest, followed by LLaVA; while other VLMs fall off rapidly, especially at stricter IoU thresholds where AP nearly collapses.

Table 5: Localization results (box-based) for task-specific vision models and general-purpose VLMs.

	Model	mIoU ↑	$AP_{0.5}$ ↑	$AP_{0.75}$ ↑	$AP_{0.5:0.95}$ ↑
Task-specific vision models	GroundingDINO-1.5	**0.773**	**0.821**	**0.711**	**0.695**
	Florence2-base	0.767	<u>0.755</u>	0.617	<u>0.602</u>
	OWLv2-base-16	0.735	0.743	<u>0.631</u>	0.592
	LISA-13B-Llama	0.711	0.658	0.513	0.505
	LISA-7B	0.649	0.535	0.427	0.409
	OWL-ViT-base-32	0.706	0.436	0.374	0.338
	D-FINE	<u>0.768</u>	0.340	0.320	0.309
General-purpose VLMs	InternVL3-8B	**0.550**	**0.454**	<u>0.079</u>	<u>0.160</u>
	LLaVA-Next-7B	**0.550**	0.342	**0.158**	**0.173**
	LLaVA-1.5-7B	0.487	<u>0.389</u>	0.022	0.111
	LLaVA-OneVision-7B-si	0.388	0.148	0.001	0.029
	Gemini 2.5-Pro	0.210	0.010	0.000	0.002
	GPT-4o	0.181	0.003	0.000	0.000
	SmolVLM2-2.2B-Instruct	0.136	0.005	0.000	0.000

Based on results in Table 5, task-specific models substantially outperform general-purpose VLMs on both mIoU and AP across different thresholds. Practically, this suggests a simple hybrid strategy for applications that need both reasoning and high-accuracy boxes: first leverage a reasoning-capable VLM to resolve the referent via language (e.g., infer the product/object name or category), then feeding the VLM output to a specialized open-vocabulary detector (OVD) for precise localization. This "reason-then-localize" pipeline preserves the VLM's scene understanding while delegating box regression and confidence calibration to models optimized for detection.

C.2 Spatial Localization Analysis

Figure 1 illustrates typical errors from five VLMs when predicting coarse grid locations of the focal product. We show two examples per model for both a 2×2 grid and a 3×3 grid. The ground-truth and model's prediction cells are marked with green and red symbols. Errors are often adjacent-cell mistakes near cell boundaries. Moving from 2×2 to 3×3 grid increases difficulty and error frequency due to finer spatial quantization and ambiguity in cluttered scenes.

C.3 Retrieval Analysis

We evaluate ANN retrieval after cropping each query image to the predicted bounding box from each model and embedding the crop with VL-CLIP. We report retrieval performance using Precision@k and Hit@k for $k \in 1, 5, 10$ compared against a full-image baseline (Table 6).

GroundingDINO yields the strongest overall retrieval among the task-specific models. All models in this group improve over the full-image baseline, although D-FINE slightly underperforms the baseline on Hit rate.

LLaVA-Next is the top performer across all metrics among general-purpose VLMs, while Gemini and GPT-4o show the lowest performance, falling below the no-crop baseline.

Retrieval failure cases Figures 10 and 11 visualize top-5 ANN retrieval results using a single query image across models. Row 1 in each figure shows the full-image baseline (i.e., retrieval on full image without cropping). Subsequent rows use query crops from the predicted bounding boxes of each model, embedded with VL-CLIP for retrieval. Columns display retrieved candidates left-to-right (rank 1 to 5); correct matches are outlined with green and incorrect ones in red borders. Small differences in the predicted crop can induce significant changes in downstream retrieval. Models that better preserve salient, object-specific features in their boxes tend to yield better matching retrievals.

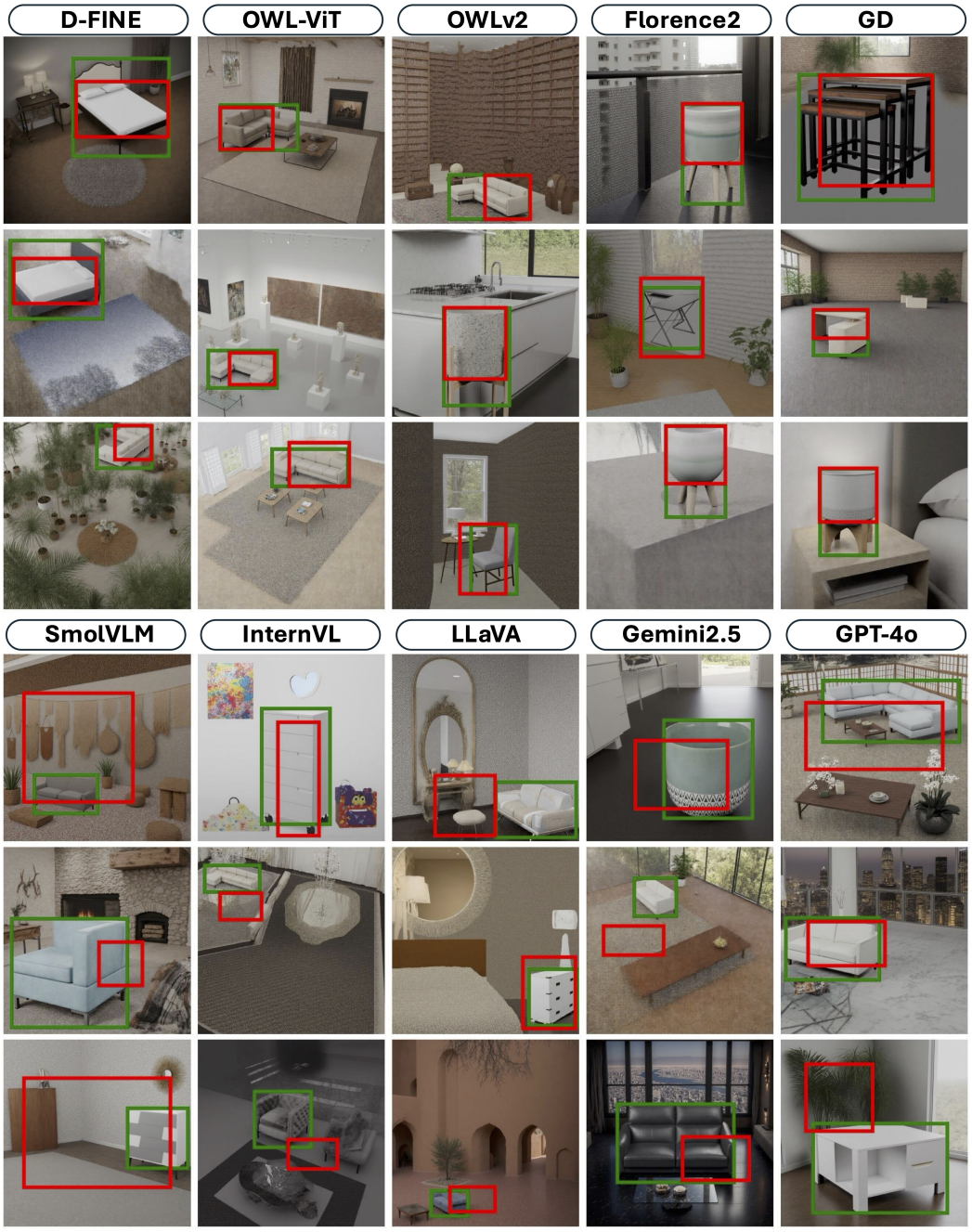

Figure 9: **Spatial localization failure cases across task-specific vision models (top) and general-purpose VLMs (bottom).** Green boxes denote ground truth; red boxes are predictions.

Results reinforce that crops which preserve the correct visual features (pose, discriminative parts) enable consistent instance-level retrieval, whereas suboptimal boxes cause retrieval to drift toward contextually similar but incorrect items.

Table 6: ANN retrieval performance over VL-CLIP embeddings using predicted-box crops. The first row shows the full-image (no crop) as a baseline.

		Precision@k ↑			Hit@k ↑	
	Model	@1	@5	@10	@5	@10
	Full image (no crop)	0.400	0.285	0.236	0.576	0.659
Task-specific vision models	GroundingDINO-1.5	**0.554**	**0.426**	**0.270**	**0.736**	**0.790**
	OWL-ViT-base-32-base-32	0.551	**0.426**	<u>0.266</u>	0.721	0.768
	Florence-2-base	<u>0.539</u>	0.419	0.264	<u>0.728</u>	<u>0.788</u>
	OWLv2-base-16-base-16	0.517	0.404	0.256	0.699	0.759
	LISA-13B-Llama	0.499	0.388	0.248	0.686	0.737
	LISA-7B	0.467	0.363	0.231	0.643	0.693
	D-FINE	0.447	0.310	0.251	0.581	0.633
General-purpose VLMs	LLaVA-Next-7B	**0.498**	**0.391**	**0.244**	**0.671**	**0.725**
	LLaVA-1.5-7B	<u>0.448</u>	0.344	0.225	<u>0.634</u>	<u>0.714</u>
	InternVL3-8B	0.445	<u>0.346</u>	<u>0.228</u>	0.625	0.700
	LLaVA-OneVision-7B-si	0.345	0.279	0.185	0.526	0.605
	SmolVLM2-2.2B-Instruct	0.277	0.209	0.147	0.421	0.491
	Gemini 2.5-Pro	0.173	0.135	0.096	0.254	0.308
	GPT-4o	0.129	0.112	0.081	0.233	0.281

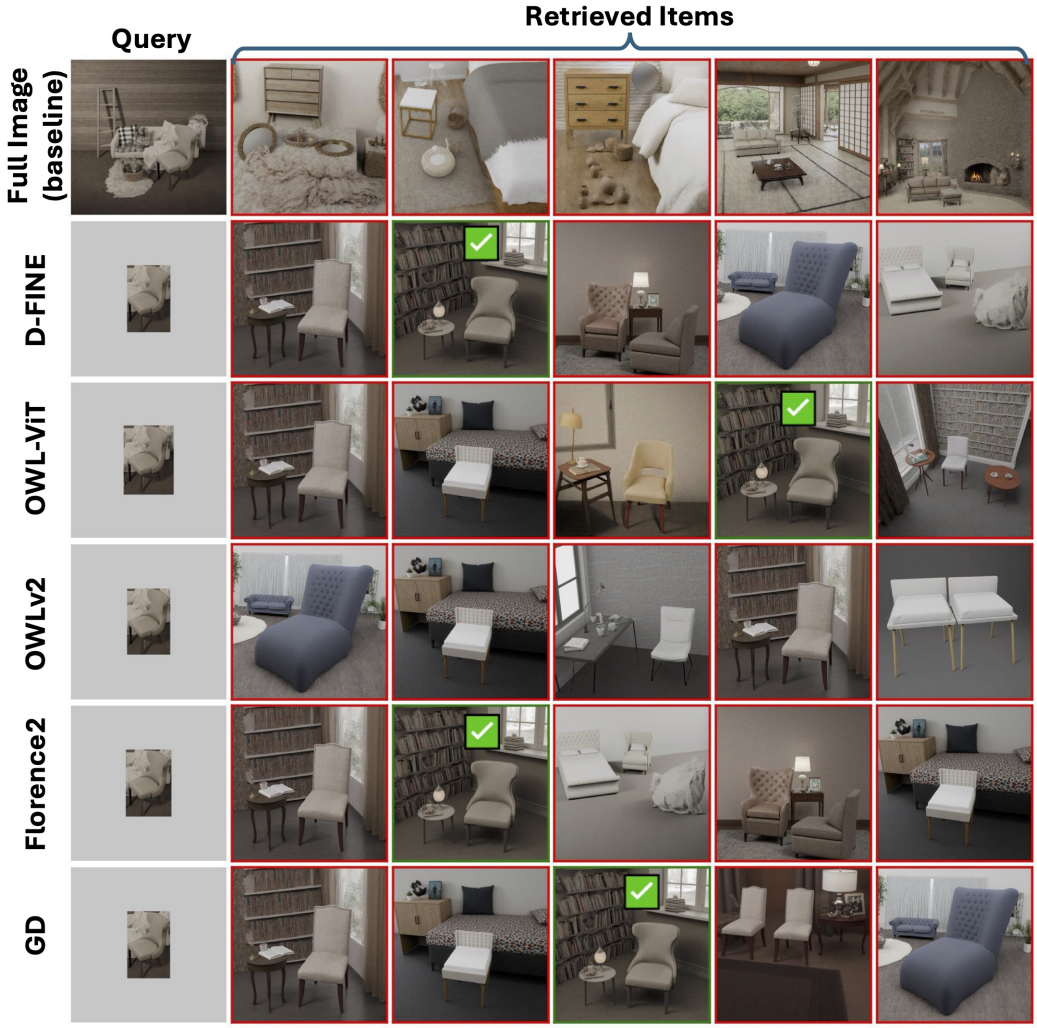

Figure 10: **Retrieval results with crops made from predicted boxes of task-specific vision models (top-5 shown).** Row 1: full-image baseline (no crop). Rows 2–6: crops from D-FINE, OWL-ViT, OWLv2, Florence-2, and GroundingDINO, respectively. The *same query image* is used across all rows; only the crop differs by model. In this example, the baseline and OWLv2 produce *no* correct matches in the top-5; *D-FINE* and Florence-2 retrieve a correct match at rank 2; OWL-ViT and GroundingDINO retrieve a correct match at rank 3. This highlights how modest crop shifts (e.g., tighter/looser boxes or slight offsets) can substantially alter retrieval outcomes.

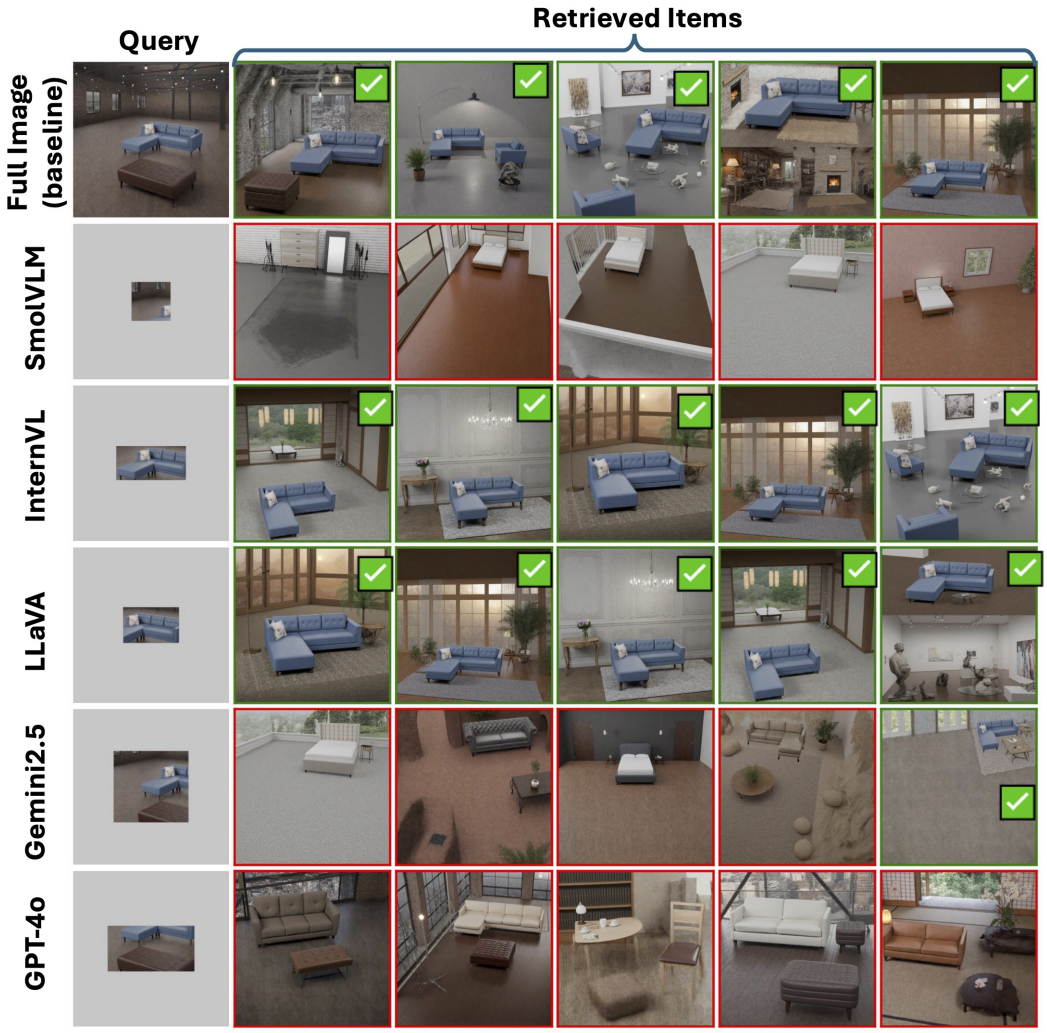

Figure 11: **Retrieval results using crops made from predicted boxes of general-purpose VLMs (top-5 shown).** Row 1: full-image baseline. Rows 2–6: crops from SmolVLM, InternVL, LLaVA, Gemini 2.5, and GPT-4o, respectively. The *same query image* is used across all rows. Here, the full-image baseline, InternVL, and LLaVA achieve *all* top-5 correct; SmolVLM and GPT-4o have *no* correct matches; Gemini 2.5 produces a single correct match at rank 5.

D VLM Prompts

D.1 Object position on a 2×2 grid

The prompt used in the 2×2 grid-cell localization experiment; the VLM selects the dominant quadrant (A–D) containing the target object.

```
Divide this image into 4 regions. Where is the {object} most
    prominently located in this image? If the object spans multiple
    regions, choose the region where the majority or most prominent
    part of the object is located.

Options:
A) top-left
B) top-right
C) bottom-left
D) bottom-right

Chose from one of the above options.
```

D.2 Object position on a 3×3 grid

The prompt used for the 3×3 grid-cell localization experiment; the VLM selects the dominant cell (A–I) where the target object lies.

```
Divide this image into 9 regions. Where is the {object} most
    prominently located in this image? If the object spans multiple
    regions, choose the region where the majority or most prominent
    part of the object is located.

Options:
A) top-left
B) top-center
C) top-right
D) middle-left
E) middle-center
F) middle-right
G) bottom-left
H) bottom-center
I) bottom-right

Chose from one of the above options.
```

D.3 Object Depth Estimation

The prompt used in the front/back depth-order classification experiment; the VLM indicates whether the target object is in the foreground or background.

```
Look at the {object} in this image. Is it positioned in the
    foreground (front) or background (back) of the scene?
Consider the relative depth and layering of objects in the image.

A) Front (foreground)
B) Back (background)

Chose from one of the above options.
```

D.4 Object Bounding Box

```
Look at the {object} in this image. Please identify the bounding
    box coordinates that tightly enclose this object.

Provide the coordinates as absolute pixel values based on the image
     dimensions:
- Image width: {width} pixels
- Image height: {height} pixels
- (0,0) is the top-left corner of the image

Format: (x1, y1, x2, y2) where:
- x1, y1 = top-left corner of the bounding box in pixels
- x2, y2 = bottom-right corner of the bounding box in pixels

Response format: Provide only the coordinates in the specified
    format.
```

E Related Works

E.1 Object Detection: From Closed-set to Open-vocabulary

Early object detectors were based on a fixed category list, often referred to as the closed-set detectors, where a model is trained and evaluated on the same finite taxonomy. This limitted inference applications: adding a new class required retraining.[16, 15, 43, 42, 20, 4, 54] To relax the fixed-label constraint, open-vocabulary detection (OVD) brings text supervision into the loop. Early OVD lines either distilled knowledge from vision–language models into detectors or learned region–text alignment at scale.[17, 57, 28, 55]

Grounding DINO [32] adopts a dual-encoder, single-decoder architecture with an image encoder, a language encoder, and a cross-modality decoder for box refinement. Its 1.5 variant [44] employs a larger ViT-L backbone [11] and is pre-trained on over 20M grounded image–text pairs, reporting 54.3 AP on COCO and 55.7 AP on the LVIS *minival* zero-shot transfer benchmark. OWL-ViT [37] pre-trains vision and text encoders with image–text contrastive learning (as in CLIP [40] and ALIGN [19]) and adapts them for open-vocabulary detection with text prompts. OWL-ST and OWL-v2 [36] further scale this approach via self-training, using an existing detector to generate pseudo-boxes on web-scale image–text pairs, yielding substantial gains on rare LVIS categories.

Florence [53] pre-trains language and image encoders with contrastive objectives and adapts to detection by attaching a Dynamic Head adapter [9]. Its successor, Florence-2 [49], comprises an image encoder and a multimodal encoder–decoder trained under a unified multi-task paradigm. While not language-grounded, D-FINE [39] revisits bounding-box regression with fine-grained distribution refinement and global localization self-distillation, offering competitive AP at high FPS.[39] Complementary to box-based OVD, LISA [23] performs reasoning segmentation, predicting language-conditioned masks with an LLM-guided planner.[23] Broader efforts further expand OVD with larger corpora, using semi/self-training, and efficiency-oriented designs.[51, 50, 52, 58, 21, 6]

E.2 Vision–Language Foundation Models for Object Localization and Spatial Reasoning

Modern vision–language models (VLMs) pair a visual encoder with a language model and learn from large image–text corpora. Common design choices include lightweight connectors (projection or cross-attention), instruction tuning for task following, and support for multi-image or video inputs. This line of work established a general recipe for multimodal reasoning and flexible prompting.[1, 2, 27, 8, 46, 25, 24]

LLaVA (Large Language and Vision Assistant) [31, 29] and its successors illustrate the rapid progress in open multimodal assistants. LLaVA connects a CLIP vision encoder to a Vicuna-based LLM via a simple projection and is trained end-to-end through a two-stage instruction-tuning pipeline [31], achieving state-of-the-art accuracy on ScienceQA and demonstrating strong visual dialogue abilities.

SmolVLM [35] is a compact model that can run on-device capable of performing tasks such as visual question answering, captioning, and visual storytelling.

Recent VLM research also explores localization and dense description. InternVL [5], Gemini [7], GPT-4o [18], and Gemma 3 [47] all support localization via bounding boxes or segmentation. As VLMs continue to grow in scale and multimodality, they increasingly unify tasks such as visual question answering, open-ended captioning, object localization, and more, moving the field beyond simple image captioning toward general-purpose vision–language understanding.

F Discussion and Future Works

We introduced a unified benchmark and evaluation protocol for product-centric retrieval that bridges detection and instance-level matching. Using a synthetic data pipeline, each product yields database images with more frontal views and query images with more angled views, enabling controlled tests of view/pose robustness. We index database embeddings with VL-CLIP and evaluate localization performance of task-specific vision models (e.g., OWL-ViT, GroundingDINO, D-FINE) alongside general purpose VLM (e.g., LLaVA, SmolVLM, InternVL). Localization quality (mIoU, AP) and retrieval quality (Precision@k, Hit@k for $k \in$1,5,10) are measured under a common setup. Across experiments, precise crops are a primary driver of retrieval success: using whole-image queries amplifies background bias, while missed/imprecise boxes and severe view changes are the dominant failure modes.

Overall, our analysis demonstrates that task-specific vision models consistently outperform general-purpose VLMs accross all experiments. Future directions of this work includes (i) end-to-end training that jointly optimizes localization and retrieval embeddings, (ii) stronger view- and pose-invariant representations (e.g., 3D/geometry cues or multi-view augmentation), (iii) spatial reasoning over multi-object scenes (compositional relations and complements), and (iv) scaling to richer real-world catalogs with harder negatives and human-in-the-loop evaluation.