GroundedPlanBench: Spatially Grounded Long-Horizon Task Planning

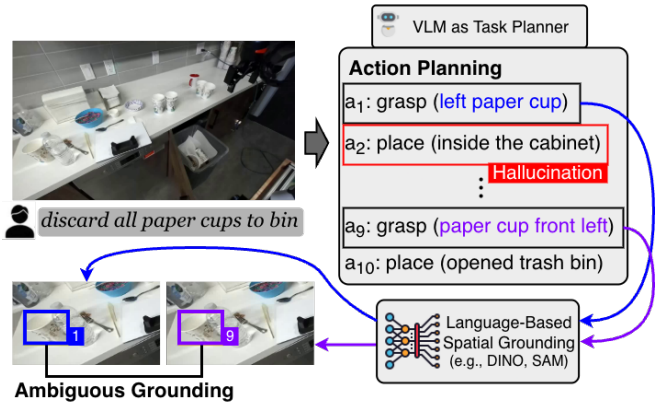
Sehun Jung[1†], HyunJee Song[1†], Dong-Hee Kim[1], Reuben Tan[2], Jianfeng Gao[2], Yong Jae Lee[3], and Donghyun Kim[1✉]

***Abstract*— Recent advances in robot manipulation increasingly leverage Vision–Language Models (VLMs) for high-level reasoning, such as decomposing task instructions into sequential action plans expressed in natural language that guide downstream low-level motor execution. However, current benchmarks do not assess whether these plans are spatially executable, particularly in specifying the exact spatial locations where the robot should interact to execute the plan, limiting evaluation of real-world manipulation capability. To bridge this gap, we define a novel task of `grounded planning` and introduce `GroundedPlanBench`, a newly curated benchmark for spatially grounded long-horizon action planning in the wild. `GroundedPlanBench` jointly evaluates hierarchical sub-action planning and spatial action grounding (*where to act*), enabling systematic assessment of whether generated sub-actions are spatially executable for robot manipulation. Using our benchmark, we evaluate diverse VLMs on `grounded planning` and find that (1) long-horizon and implicit planning remains challenging, (2) spatial grounding is still limited even with strong models.**

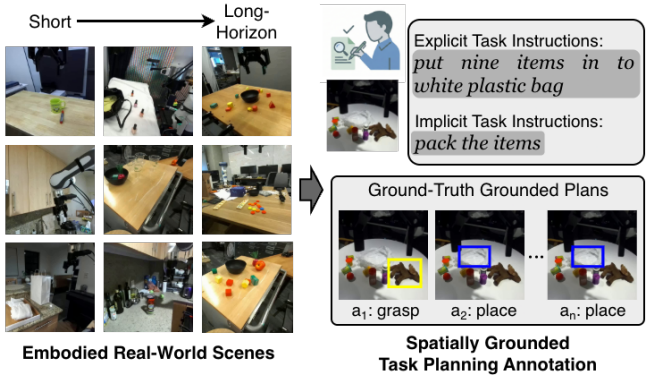
I. Introduction

Recent advances in Vision–Language Models (VLMs), trained on large-scale web data, have driven a paradigm shift in robotics toward general-purpose agents capable of robust perception, reasoning, and generalization for interaction and manipulation in open-world environments. A dominant paradigm uses VLMs to directly map visual observations and language instructions to low-level robot motor signals, known as Vision–Language–Action (VLA) models [1]–[3]. While promising, these end-to-end VLAs often struggle with generalization across diverse scenes and tasks [4], [5]. Consequently, VLM-as-Perception approaches [4], [5] decouple high-level reasoning from low-level motor control, using VLMs for spatial perception to guide downstream execution (*e.g.*, motion planning), thereby improving generalization in robotic manipulation. However, such methods typically lack the hierarchical reasoning capabilities necessary for long-horizon task decomposition and planning. In parallel, the VLM-as-Planner [6]–[8] exploits the reasoning ability of VLMs to decompose complex instructions into sequences of low-level actions expressed in natural language.

In this paper, we examine a key limitation of VLM-as-Planner: the generated sub-actions, expressed in natural language, are often ambiguous or hallucinated, resulting in plans that are not physically or spatially executable. Existing benchmarks for VLM-based task planning [9]–[12] evaluate the ability to decompose high-level instructions into low-level action plans. However, while effective for abstract planning, these benchmarks are often simulator-based, limiting exposure to diverse real-world scenes and tasks, and frequently lack explicit evaluation of spatial grounding during planning. Consequently, they measure what actions to perform but not where to execute them, making it challenging to verify spatial feasibility. As illustrated in Fig. 1a, the natural language sub-actions generated by the VLM-as-Planner paradigm often fail to accurately specify where to act due to inherent ambiguities



(a) Limitation of prior decoupled task planning & spatial grounding.



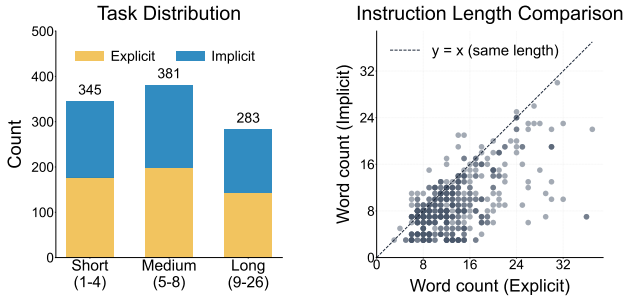
(b) Construction of `GroundedPlanBench`.

Fig. 1: Motivation of `grounded planning`. (a) VLM-as-Planner decomposes high-level instructions into natural language sub-actions, which are then grounded by separate perception modules. However, the lack of explicit spatial specification often leads to ambiguous action grounding. (b) Our `GroundedPlanBench` jointly annotates and evaluates hierarchical sub-action planning and spatial grounding under both explicit and implicit instructions.

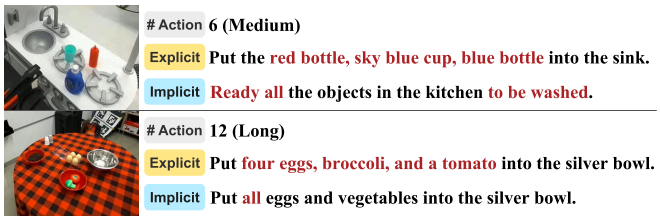
† indicates first authors with equal contributions.

[1]Korea University, [2]Microsoft Research [3]University of Wisconsin-Madison

* The work was supported the Institute of Information &communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. RS-2025-25439490). (✉ Corresponding author: d_kim@korea.ac.kr).



(a) (Left) Dataset statistics. (Right) Word counts between explicit and implicit instructions. Each point represents the paired implicit and explicit instruction length for a given video.



(b) Examples of paired explicit and implicit instructions.

Fig. 2: Task distribution and instruction types in `GroundedPlanBench`.

or hallucination in linguistic descriptions.

We define a novel task of `grounded planning`, where the goal is to perform **spatially grounded and long-horizon action planning** for task completion in the wild. To bridge this gap, we introduce `GroundedPlanBench`, a new benchmark designed to rigorously assess two critical dimensions: (1) hierarchical action planning, which requires deriving appropriate actions across diverse scenes and tasks, spanning short- to long-horizon objectives based on both explicit and implicit user instructions, and (2) spatial action grounding, which requires the model to precisely identify where to interact with objects for each action step. Unlike prior benchmarks, by jointly evaluating action planning and spatial grounding, our benchmark assesses whether VLM-generated plans correspond to physically feasible robot manipulation behaviors (Fig. 1b).

Using our benchmark, we evaluate closed-source and open-source VLMs on `grounded planning`. We find that (1) planning under implicit or long-horizon instructions in diverse real-world scenes remains challenging and (2) spatial grounding of actions ("where to act") in current VLMs is limited even with strong grounding models.

II. Benchmark Overview

We introduce `GroundedPlanBench`, a benchmark designed to evaluate the spatially grounded action planning capabilities of VLM-as-Planner in unconstrained environments. This benchmark encompasses varied task horizons and both implicit and explicit instructions across a wide array of scenes and objects, as illustrated in Fig. 2.

Task Definition: `Grounded Planning`. Given an image I and corresponding high-level instruction H, the grounded plan $a^{(I,H)}$ for executing the task can be decomposed into a sequence of N sub-steps, $a^{(I,H)} = [a_1, a_2, \ldots, a_N]$, where each a_i denotes a primitive action at the i-th substep. Motivated by [13], we treat the 'grasp' and 'place' as the essential atomic primitives for general-purpose robotic manipulation, and additionally define 'open' and 'close' as separate primitives for functional interactions with articulated objects (*e.g.*, drawers and cabinets). In contrast to prior task planning literature, we explicitly ground each sub-action (a_i) in the plan with spatial signals (*i.e.*, bounding boxes and points), thereby removing the localization ambiguity inherent in natural language instructions. To this end, we prompt the VLM task planner to generate spatially grounded action plans using prompts of the following form (simplified here and further optimized for each VLM) in Fig. 3:

```
Input:
- One observation image {I}
- One high-level instruction {H}
Goal: Generate a spatially grounded action plan using
ONLY: {open, grasp, place, close}.
Each action must operate on exactly one grounded
entity.
Primitive spec:
- open(target_text, bbox)
- close(target_text, bbox)
- grasp(target_text, bbox)
- place(target_text, point)
```

Fig. 3: An example of a simplified instruction prompt for spatially grounded planning.

Data Selection and Annotation. Fig. 1b illustrates an overview of the overall pipeline. To evaluate the ability of VLMs to perform spatially grounded action planning in real-world settings, we build upon the DROID dataset [14], a large-scale embodied robotics resource containing demonstrations with substantial object diversity across 564 scenes and 86 tasks. To construct our test benchmark, we sample 308 videos spanning short-, medium-, and long-horizon demonstrations. From the first frame of each video, human annotators define a scene-executable task and provide both explicit (concrete and detailed) and implicit (abstract) instructions, along with the corresponding spatially grounded action plan $[a_1, a_2, \ldots, a_n]$. Each grasp, open, and close action is grounded using a bounding box over the target object, while the place action is grounded with a bounding box indicating the intended destination or receptacle region. We then group episodes by annotated action horizon: plans with 1–4 actions are categorized as Short, 5–8 as Medium, and 9–26 as Long, totaling 1,009 evaluation episodes.

Evaluation Metric. We adopt metrics that jointly measure sequential planning accuracy and spatial precision. We assess each action step, requiring both the correct ordering of primitives and precise spatial grounding. Since a grasp primitive is always followed by a place primitive in the ground-truth (GT) plan, we evaluate grasp–place as a single atomic pair. Specifically, we verify whether the Intersection over Union (IoU) between predicted and GT grasp bounding boxes exceeds an IoU threshold ($\tau_g = 0.5$), and verify whether the predicted placement point ($\hat{p}_i$) falls within the corresponding GT placement bounding box (B_i^p). For open and close primitives, we verify whether the IoU between

TABLE I: Evaluation results on `GroundedPlanBench` across varying task horizons (short (1–4), medium (5–8), long (9–26)) and instruction types (explicit and implicit instructions).

Model	Task Success Rate (TSR) ↑						Action Recall Rate (ARR) ↑					
	Explicit			Implicit			Explicit			Implicit		
	short	medium	long	short	medium	long	short	medium	long	short	medium	long
Proprietary VLMs												
Decoupled Task Planning + Spatial Grounding												
GPT 5.2 + ER1	41.8	13.6	8.4	31.0	10.4	3.6	54.8	48.9	46.7	45.8	42.4	38.4
GPT 5.2 + SAM3 + ER1	36.2	10.1	4.9	28.0	8.7	1.4	47.6	42.4	36.9	41.9	36.2	32.1
End-to-End Spatially Grounded Planning												
GPT 5.2	3.4	0.0	0.0	0.6	0.0	0.0	7.2	3.8	3.7	3.0	3.0	2.3
Gemini-2.5-Flash	20.9	11.6	5.6	10.7	6.0	5.0	28.3	28.2	29.0	21.5	20.6	22.8
Gemini-3-Flash	**67.2**	**52.0**	**42.7**	**57.1**	**31.7**	**17.9**	**73.1**	**71.6**	**75.1**	**67.2**	**57.8**	**55.9**
Open-Source VLMs												
Decoupled Task Planning + Spatial Grounding												
Qwen3-VL-4B + SoM	14.7	2.0	0.0	4.2	0.5	0.0	19.5	18.3	14.6	9.4	9.9	8.1
Qwen3-VL-4B + ER1	37.3	9.1	2.8	18.5	4.4	0.7	49.7	39.8	32.5	33.0	26.4	21.0
Qwen3-VL-4B + SAM3 + ER1	35.6	6.1	2.8	18.5	2.2	0.7	48.6	33.4	30.6	31.8	23.7	20.2
Qwen3-VL-32B + ER1	40.1	12.6	8.4	29.8	7.1	2.9	52.0	47.0	45.4	45.5	36.3	35.6
Qwen3-VL-32B + SAM3 + ER1	36.7	11.6	4.9	29.2	6.0	0.7	48.8	41.6	42.4	43.7	31.6	33.7
End-to-End Spatially Grounded Planning												
InternVL3.5-8B	0.0	0.0	0.0	0.6	0.0	0.0	0.3	0.3	0.6	0.6	0.3	0.0
Qwen3-VL-4B	39.5	18.2	5.6	22.6	6.0	1.4	50.9	43.3	36.7	35.0	27.4	23.0
Qwen3-VL-32B	42.9	30.8	14.0	23.8	12.0	2.9	55.6	55.7	54.2	39.9	37.7	33.4

the predicted and GT bounding boxes exceeds the threshold ($\tau_d = 0.5$). In summary, we define the sub-action success as:

$$\text{Succ}_i = \begin{cases} \mathbf{1}\left[\text{IoU}_i \geq \tau_g \ \wedge \ \hat{p}_i \in B_i^p\right], & \text{grasp–place,} \\ \mathbf{1}\left[\text{IoU}_i \geq \tau_d\right], & \text{open/close.} \end{cases}$$

Additionally, while task planning often requires a specific sequence of actions (*e.g.*, "move cup first and then move a spoon"), certain tasks allow for multiple valid execution sequences, such as "move all objects into the pot." To account for this flexibility, we define an Unordered Action Group, in which the relative ordering of actions does not affect overall task validity. Within each unordered block, any permutation of the GT actions is considered correct. Based on these criteria, we define the following evaluation metrics: (1) Action Recall Rate (ARR): ARR measures the proportion of generated actions that match the GT sub-actions, without considering their sequential orders. (2) Task Success Rate (TSR): A task is considered successful if and only if all actions in the sequence are correctly planned and spatially grounded.

III. Experiments

A. Experimental Settings

Baselines. We evaluate baselines across the following two paradigms: (i) *End-to-End Spatially Grounded Planning*: In this configuration, VLMs directly generate action sequences alongside their corresponding spatial grounding (i.e., specific bounding boxes and points). We evaluate several state-of-the-art models in this category, including Gemini-3-Flash, Gemini-2.5-Flash [15], Qwen3-VL [16], and InternVL3.5 [17]. (ii) *Decoupled Task Planning and Spatial Grounding*: This scenario evaluates VLMs as high-level task planners that output natural language plans. These plans are then spatially grounded by SAM3 [18] and the specialized VLM, Embodied-R1 [5], where SAM3 provides segmentation-based bounding boxes and Embodied-R1 predicts interaction points. This two-stage design leverages complementary strengths: general-purpose VLMs for task planning and grounding models for spatial localization. We investigate whether this decoupled approach leads to improved task planning that remains spatially grounded, compared to direct grounded planning. To assess modular performance, we employ GPT 5.2 and Qwen3-VL as high-level task planners. We evaluate three distinct grounding configurations: (1) employing Set-of-Mark (SoM) [19] to provide visual context to the planner, (2) using SAM3 for grasp and Embodied-R1 for placement (denoted as "SAM3 + ER1"), and (3) using Embodied-R1 for both grasp and placement to translate the planners' natural language plans into spatially grounded coordinates (denoted as "ER1").

B. Experimental Results

Overall Performances. Table I summarizes the performance on `GroundedPlanBench`. Among the evaluated models, Gemini-3-Flash leads in end-to-end spatially grounded planning, achieving particularly high Task Success Rates (TSR) in short-horizon tasks when provided with explicit instructions. In comparison, models such as GPT-5.2 and Gemini-2.5 exhibit a significant performance gap compared to Gemini-3-Flash. Furthermore, open-source models like InternVL3.5 and Qwen3-VL struggle even with short-horizon tasks, resulting in substantially lower success rates. These further degrades with implicit instructions, where VLMs struggle to infer missing sub-actions due to limited reasoning and intent understanding.
Are action plans spatially grounded? To investigate this, we visualize the spatial grounding of plans generated by

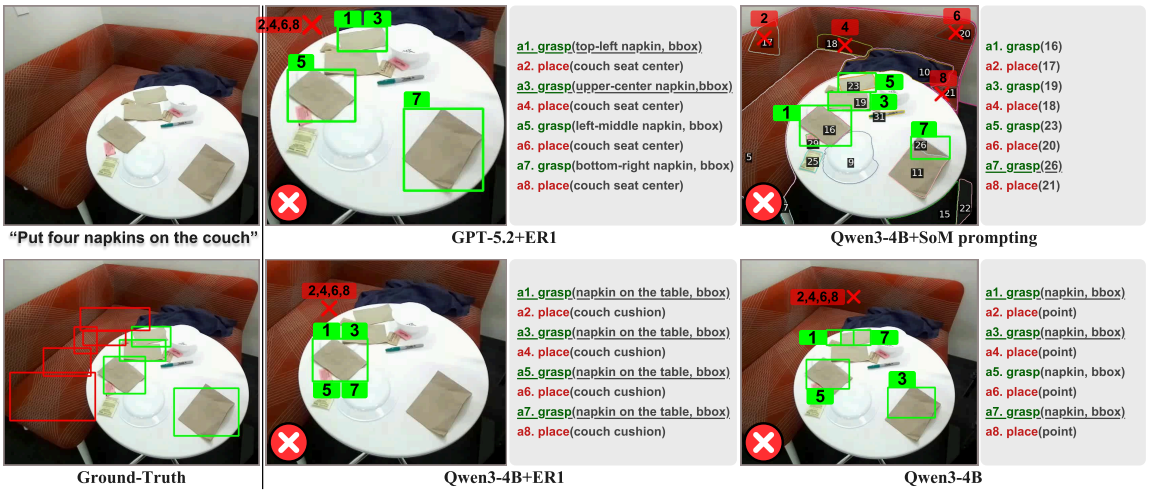


Fig. 4: Visualization of grounded planning that jointly generates sub-actions and their corresponding spatial locations, where underlined actions are incorrectly grounded due to semantically similar objects (*e.g.*, identical napkins).

decoupled approaches in Fig. 4. We observe that when using decoupled task planning and spatial grounding (e.g., GPT-5.2+ER-1, Qwen-3+ER-1), the natural language action plans often specify objects ambiguously or redundantly. This misalignment between linguistic description and visual entities prevents the model from correctly grounding all objects required for the task. In the case of SoM (Set-of-Mark) prompting, the task planner frequently fails to select the correct objects, likely due to the high density of visual noise and overlapping prompts in complex real-world settings.

IV. Conclusion and Future Work

This paper addresses the gap between long-horizon task planning (what to do) and spatial grounding (where to act) in existing VLMs. We introduce `GroundedPlanBench`, a benchmark for jointly evaluating high-level planning and spatial grounding in real-world environments. Experiments with both closed- and open-source VLMs reveal significant challenges in spatial grounding for action planning. Future work will explore stronger world models for verifying and correcting grounded plans to improve physical executability.

References

[1] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[2] O. Mees, D. Ghosh, K. Pertsch, K. Black, H. R. Walke, S. Dasari, J. Hejna, T. Kreiman, C. Xu, J. Luo *et al.*, "Octo: An open-source generalist robot policy," in *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024.

[3] J. Yang, R. Tan, Q. Wu, R. Zheng, B. Peng, Y. Liang, Y. Gu, M. Cai, S. Ye, J. Jang *et al.*, "Magma: A foundation model for multimodal ai agents," in *Proceedings of the computer vision and pattern recognition conference*, 2025, pp. 14 203–14 214.

[4] Y. Yuan, H. Cui, Y. Chen, Z. Dong, F. Ni, L. Kou, J. Liu, P. Li, Y. Zheng, and J. Hao, "From seeing to doing: Bridging reasoning and decision for robotic manipulation," *arXiv preprint arXiv:2505.08548*, 2025.

[5] Y. Yuan, H. Cui, Y. Huang, Y. Chen, F. Ni, Z. Dong, P. Li, Y. Zheng, and J. Hao, "Embodied-r1: Reinforced embodied reasoning for general robotic manipulation," *arXiv preprint arXiv:2508.13998*, 2025.

[6] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as policies: Language model programs for embodied control," *arXiv preprint arXiv:2209.07753*, 2022.

[7] M. Zawalski, W. Chen, K. Pertsch, O. Mees, C. Finn, and S. Levine, "Robotic control via embodied chain-of-thought reasoning," *arXiv preprint arXiv:2407.08693*, 2024.

[8] F. Liu, K. Fang, P. Abbeel, and S. Levine, "Moka: Open-world robotic manipulation through mark-based visual prompting," *arXiv preprint arXiv:2403.03174*, 2024.

[9] P. Sermanet, T. Ding, J. Zhao, F. Xia, D. Dwibedi, K. Gopalakrishnan, C. Chan, G. Dulac-Arnold, S. Maddineni, N. J. Joshi *et al.*, "Robovqa: Multimodal long-horizon reasoning for robotics," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 645–652.

[10] R. Yang, H. Chen, J. Zhang, M. Zhao, C. Qian, K. Wang, Q. Wang, T. V. Koripella, M. Movahedi, M. Li *et al.*, "Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents," *arXiv preprint arXiv:2502.09560*, 2025.

[11] S. Zhang, Z. Xu, P. Liu, X. Yu, Y. Li, Q. Gao, Z. Fei, Z. Yin, Z. Wu, Y.-G. Jiang *et al.*, "Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 11 142–11 152.

[12] Y. Ji, H. Tan, J. Shi, X. Hao, Y. Zhang, H. Zhang, P. Wang, M. Zhao, Y. Mu, P. An *et al.*, "Robobrain: A unified brain model for robotic manipulation from abstract to concrete," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 1724–1734.

[13] B. Wang, J. Zhang, S. Dong, I. Fang, and C. Feng, "Vlm see, robot do: Human demo video to robot action plan via vision language model," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025, pp. 17 215–17 222.

[14] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis *et al.*, "Droid: A large-scale in-the-wild robot manipulation dataset," *arXiv preprint arXiv:2403.12945*, 2024.

[15] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. 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.

[16] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge *et al.*, "Qwen3-vl technical report," *arXiv preprint arXiv:2511.21631*, 2025.

[17] W. Wang, Z. Gao, L. Gu, H. Pu, L. Cui, X. Wei, Z. Liu, L. Jing, S. Ye, J. Shao *et al.*, "Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency," *arXiv preprint arXiv:2508.18265*, 2025.

[18] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang *et al.*, "Sam 3: Segment anything with concepts," *arXiv preprint arXiv:2511.16719*, 2025.

[19] J. Yang, H. Zhang, F. Li, X. Zou, C. Li, and J. Gao, "Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v," *arXiv preprint arXiv:2310.11441*, 2023.