RoboEXP: Action-Conditioned Scene Graph via Interactive Exploration for Robotic Manipulation

Hanxiao Jiang[1] Binghao Huang[1] Ruihai Wu[3] Zhuoran Li[4]
Shubham Garg[2] Hooshang Nayyeri[2] Shenlong Wang[1] Yunzhu Li[1]
[1]University of Illinois Urbana-Champaign [2]Amazon [3]Peking University [4]National University of Singapore

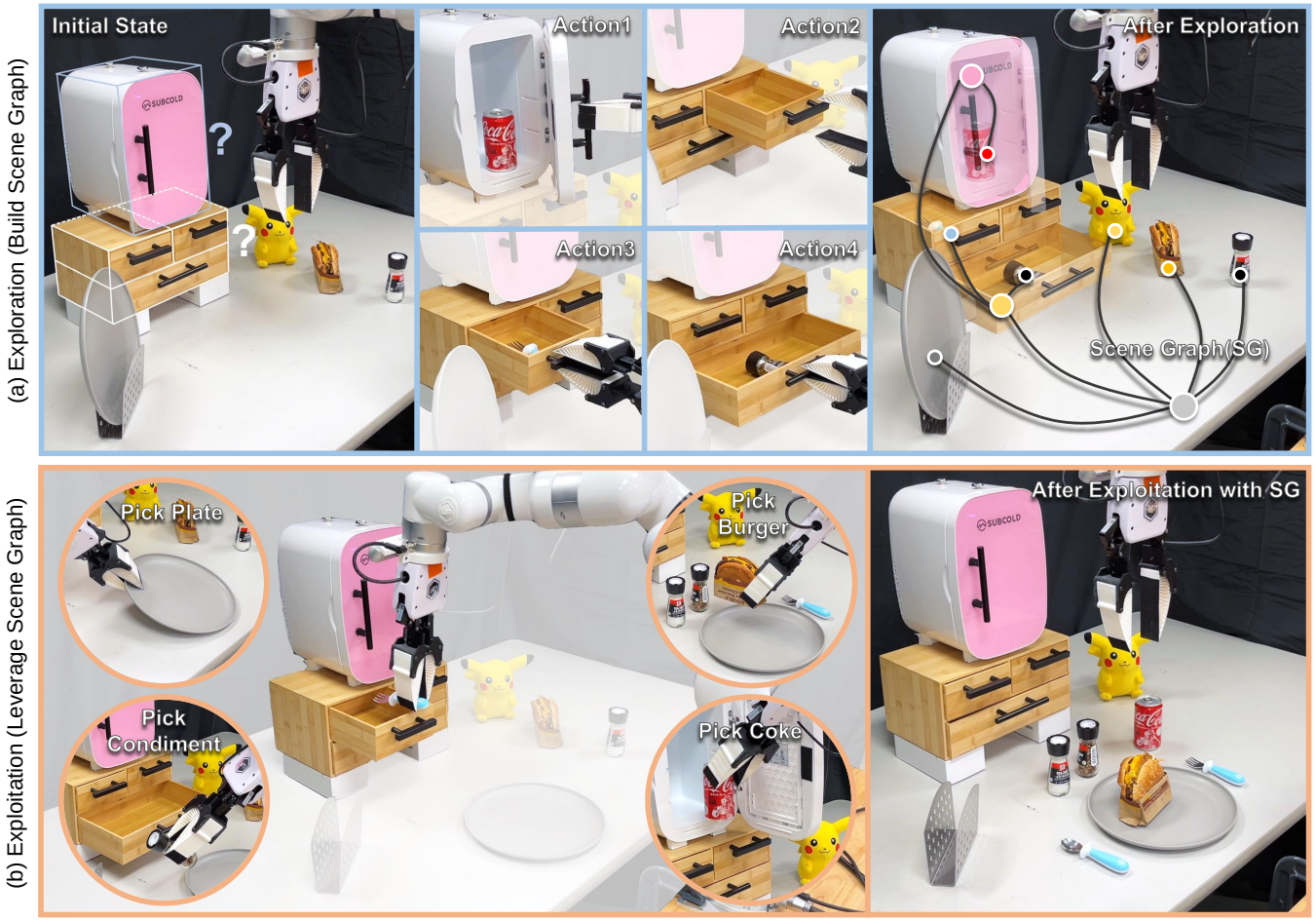

Fig. 1: **Interactive Exploration to Construct an Action-Conditioned Scene Graph (ACSG) for Robotic Manipulation.** (a) **Exploration:** The robot autonomously explores by interacting with the environment to generate a comprehensive ACSG. This graph is used to catalog the locations and relationships of items. (b) **Exploitation:** Utilizing the constructed scene graph, the robot completes downstream tasks by efficiently organizing the necessary items according to the desired spatial and relational constraints.

***Abstract*— Robots need to explore their surroundings to adapt to and tackle tasks in unknown environments. Prior work has proposed building scene graphs of the environment but typically assumes that the environment is static, omitting regions that require active interactions. This severely limits their ability to handle more complex tasks in household and office environments: before setting up a table, robots must explore drawers and cabinets to locate all utensils and condiments. In this work, we introduce the novel task of interactive scene exploration, wherein robots autonomously explore environments and produce an action-conditioned scene graph (ACSG) that captures the structure of the underlying environment. The ACSG accounts for both low-level information, such as geometry and semantics, and high-level information, such as the action-conditioned relationships between different entities in the scene. To this end, we present the Robotic Exploration (RoboEXP) system, which incorporates the Large Multimodal Model (LMM) and an explicit memory design to enhance our system's capabilities. The robot reasons about what and how to explore an object, accumulating new information through the interaction process and incrementally constructing the ACSG. We apply our system across various real-world settings in a zero-shot manner, demonstrating its effectiveness in exploring and modeling environments it has never seen before. Leveraging the constructed ACSG, we illustrate the effectiveness and efficiency of our RoboEXP system in facilitating a wide range of real-world manipulation tasks involving rigid, articulated objects, nested objects like Matryoshka dolls, and deformable objects like cloth. Project Page: https://jianghanxiao.github.io/roboexp-web/**

I. Introduction

Imagine a future household robot designed to prepare breakfast. This robot must efficiently perform various tasks such as conducting inventory checks in cabinets, fetching food from the fridge, gathering utensils from drawers, and spotting leftovers under food covers. Key to its success is the ability to interact with and explore the environment, especially to find items that aren't immediately visible. Equipping it with such capabilities is crucial for the robot to effectively complete its everyday tasks.

In this work, we investigate the interactive scene exploration task, where the goal is to efficiently identify all objects, including those that are directly observable and those that can only be discovered through interaction between the robot and the environment (see Fig. 1). Towards this goal, we present a novel scene representation called action-conditioned 3D scene graph (ACSG). Unlike conventional 3D scene graphs that focus on encoding static relations, ACSG encodes both spatial relationships and logical associations indicative of action effects (e.g., opening a fridge will reveal an apple inside). We then show that interactive scene exploration can be formulated as a problem of action-conditioned 3D scene graph construction and traversal.

Tackling interactive scene exploration poses challenges: how can we reason about which objects need to be explored, choose the right action to interact with them, and maintain knowledge about our exploration findings? With these challenges in mind, we propose a novel, real-world robotic exploration framework, the RoboEXP system. RoboEXP can handle diverse exploration tasks in a zero-shot manner, constructing complex action-conditioned 3D scene graph in various scenarios, including those involving obstructing objects and requiring multi-step reasoning (Fig. 2). We evaluate our system across various settings, spanning simple, single-object scenarios to complex environments, demonstrating its adaptability and robustness. The system also effectively manages different human interventions. Moreover, we show that our reconstructed action-conditioned 3D scene graph demonstrates strong capacity in performing multiple complex downstream tasks. Action-conditioned 3D scene graph advances LLM/LMM-guided robotic manipulation and decision-making research [1, 2], extending their operation domain from environments with known or observable objects to complicated environments with unknown or unobserved ones. To our knowledge, this is the first of its kind.

Our contributions are as follows: i) we propose action-conditioned 3D scene graph and introduce the interactive scene exploration task to address the challenging interaction aspect of exploration; ii) we develop the RoboEXP system, capable of exploring complicated environments with unseen objects in a wide range of settings; iii) through extensive experiments, we demonstrate our system's ability to construct complex and complete action-conditioned 3D scene graph, demonstrating significant potential for various manipulation tasks. Our experiments involve rigid and articulated objects, nested objects like Matryoshka dolls, and deformable objects like cloth, showcasing the system's generalization ability across objects, scene configurations, and downstream tasks.

II. Problem Statement

We unfold this section with an introduction of action-conditioned 3D scene graph, a novel scene representation illustrating interactive object relationships (Sec. II-A). We then formulate interactive scene exploration as an action-conditioned 3D scene graph construction and traversal problem (Sec. II-B). Check our Appendix for more details on the formal definition.

A. Action-Conditioned 3D Scene Graph

An action-conditioned 3D scene graph (ACSG) is an actionable, spatial-topological representation that models objects and their interactive and spatial relations in a scene. Fig. 2 depicts a complete action-conditioned 3D scene graph of a tabletop scene. One advantage of our interaction-aware scene graph lies in its simplicity for retrieving and taking actions on an object. Regardless of how complicated the scene is, given our scene graph and a target object, an agent merely needs to sequentially execute all the actions on the paths from the root to the object node in a topological order to retrieve the object. For example, in Fig. 2, to reach the tape inside a cabinet whose door is blocked by a condiment, according to the graph, one simply needs to: 1) pick up the condiment on the table that blocks the cabinet door, and 2) open the cabinet through the door handle.

B. Interactive Exploration

This subsection describes how we can construct a complete action-conditioned scene graph of a real-world scene. This is a challenging problem due to partial observability. For instance, a banana cannot be populated without *opening* the cabinet. To solve this task, we formulate the scene graph construction as an active perception and exploration problem using POMDP-inspired notations. Formally, at each time t, based on our past graph estimation $\mathbf{G}^{t-1}$, and past sensor observations $\mathbf{O}^{t-1}$, our agent takes an action $\mathbf{A}^t$, which causes the environment to transition to a new state, and the agent receives a new observation $\mathbf{O}^t$, which is used to update its current inferred graph $\mathbf{G}^t$. This update might include adding new nodes to the graph or updating the state of an existing node. We will then continue with exploration and keep updating the set of remaining unexplored nodes $\mathbf{U} \subset \mathbf{V}$ (see algorithm in our Appendix).

The goal of the exploration is simple: discover and explore all the nodes of the scene graph in as little time as possible. Towards this, we formulate a reward function with three terms:

$$\mathbf{R}^t = \mathbf{R}^t_{\text{graph}} + \mathbf{R}^t_{\text{explore}} + \mathbf{R}^t_{\text{time}}$$

where $\mathbf{R}^t_{\text{graph}} = |\mathbf{V}^t| - |\mathbf{V}^{t-1}|$ is the graph construction term, which promotes our agent to discover as many nodes as possible to the graph, $\mathbf{R}^t_{\text{explore}} = \max(0, |\mathbf{U}^{t-1}| - |\mathbf{U}^t|)$ gives positive reward to actions that reduce unexplored node set, which prioritize the agent to explore previously unexplored nodes, and immediate reward $\mathbf{R}^t_{\text{time}} = -\lambda, 0 < \lambda < 1$ is a

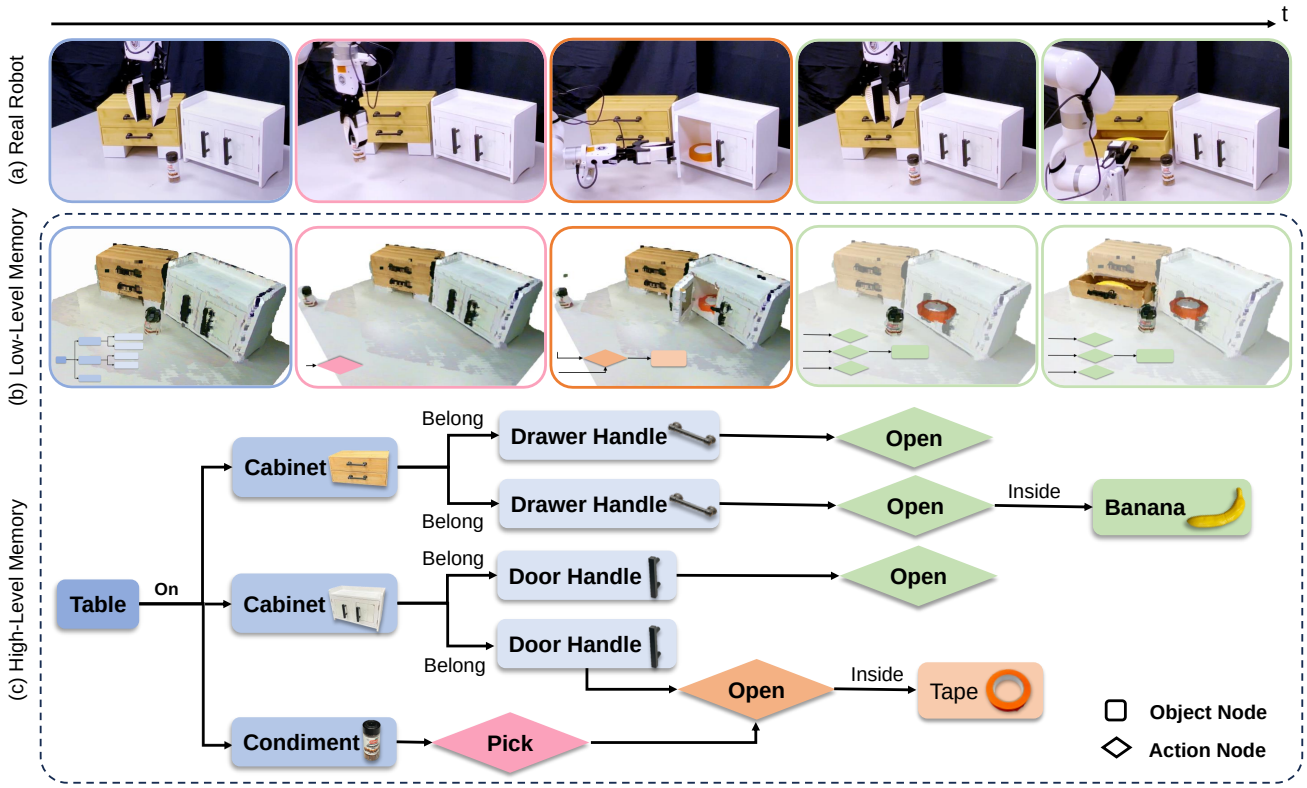

Fig. 2: **Action-Conditioned 3D Scene Graph from Interactive Scene Exploration.** To illustrate the construction process of our ACSG in the interactive scene exploration, we depict a scenario wherein a robot arm explores a tabletop scene containing two cabinets and a condiment obstructing the left door. (a) The robot arm actively interacts with the scene, completing the interactive scene exploration process. (b) We showcase the corresponding low-level memory in our ACSG, which represents the geometry and semantic information of the scene. The small graph within each visualization represents a segment of the final scene graph. (c) We present the high-level memory of our action-conditioned scene graph. The graph reveals that picking up the condiment serves as a precondition for opening the door, and opening the bottom drawer allows the observation of the concealed tape and banana.

negative time reward that optimizes the time efficiency and allows the exploration to terminate when there is no more node to explore.

III. Method

To tackle the task outlined in Section Sec. II, we present our RoboEXP system, designed to autonomously explore unknown environments by observing and interacting with them.

At the core of our system is a large foundational model-powered instantiation of ACSG. Specifically, our framework consists of four modules: perception, memory, decision-making, and action, as shown in Fig. 3. To address the challenge of perceiving what is present in the scene, our **perception module** (Fig. 3a) utilizes Grounding-DINO ([3]), Segment Anything in High Quality (SAM-HQ) [4, 5], and CLIP [6] to detect objects or parts and extract their language-embedded semantic features. Our **decision-making module** (Fig. 3c) employs the rich commonsense knowledge contained in large multimodal models, such as GPT-4V [7, 8], to assist in selecting which objects to explore and what actions to take, and in validating their plausibility. Once the decision-making module has chosen a skill, our **action module** (Fig. 3d) is then activated to follow the plans formulated by the prior modules. During the entire physical interaction process, our **memory model** (Fig. 3a), which maintains the action-conditioned scene graph, will be continuously updated to preserve the scene's knowledge for future exploration and exploitation. Despite its strong capacity, our hardware system is simple—it requires only a single RGB-D wrist camera as sensor input and uses a single robot arm for actions (see our appendix for more details).

IV. Experiments

To assess our system's efficacy across various exploration scenarios, we compared it with a strong baseline by augmenting GPT-4V with ground truth actions. We designed five types of experiments, each with 10 different settings varying in object number, type, and layout. Our quantitative analysis reveals that our RoboEXP system consistently surpasses the baseline across various tasks. Furthermore, we validate the performance of our system in constructing ACSG through qualitative demonstrations. Check the Appendix and our supplementary video for more details on the qualitative and quantitative results.

Evaluation. To thoroughly assess the efficacy of our system compared to the baseline, we have designed five key metrics (Success, Object Recovery, State Recovery, Unexplored Space, Graph Edit Distance) to measure its performance. It is crucial to note that the output of our task, represented by ACSG,

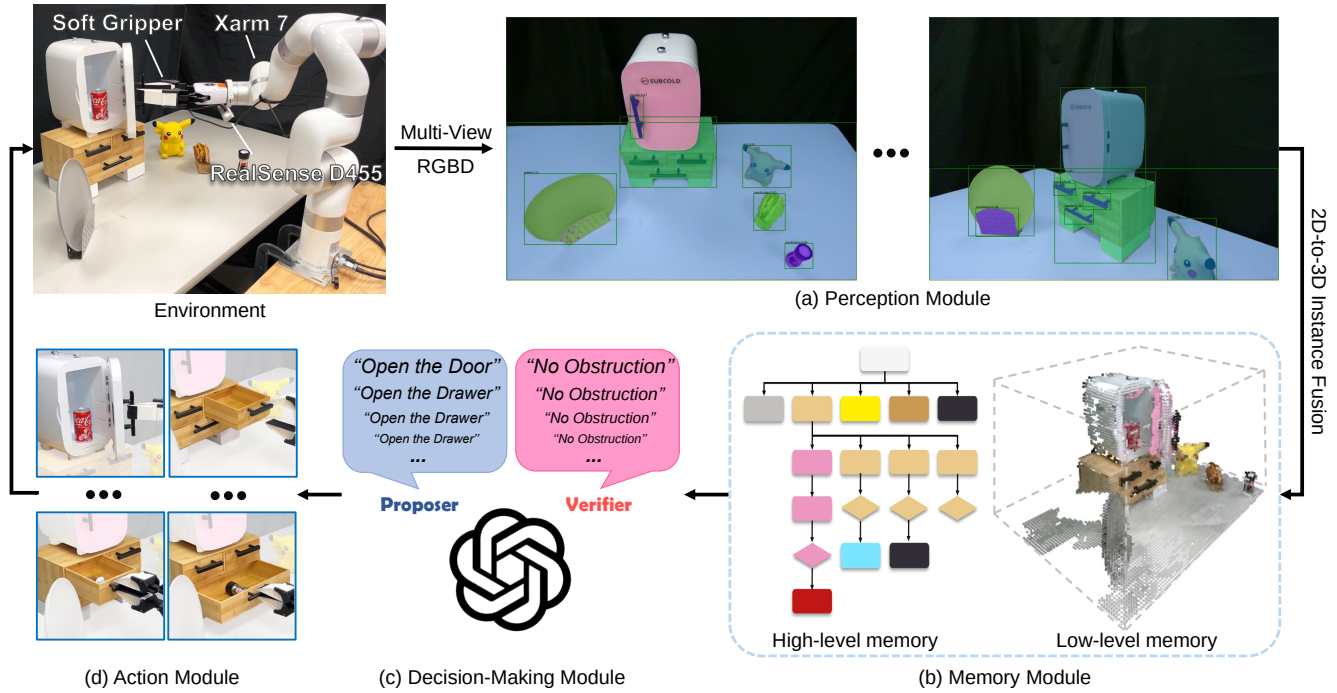

Fig. 3: **Overview of Our RoboEXP System.** We present a comprehensive overview of our RoboEXP system, comprised of four modules. (a) Our **perception module** takes RGBD images as input and produces the corresponding 2D bounding boxes, masks, object labels, and associated semantic features as output. (b) The **memory module** seamlessly integrates 2D information into the 3D space, achieving more consistent 3D instance segmentation. Additionally, it constructs the high-level graph of our ACSG through the merging of instances. (c) Our **decision-making module** serves dual roles as a proposer and verifier. The proposer suggests various actions, such as opening doors and drawers, while the verifier assesses the feasibility of each action, considering factors like obstruction. (d) The **action module** executes the proposed actions, enabling the robot arm to interact effectively with the environment.

aligns precisely with the format of ACSG for our system. Conversely, for the baseline, we manually construct ACSG based on its actions and the new observations it uncovers. Due to the unstructured nature of the raw scene graph from the baseline, we carefully refine it according to the observable objects, providing an upper-bound baseline for comparison during evaluation.

Comparison. The quantitative findings underscore the superior performance of our system compared to the baseline method. Our approach showcases a notable enhancement across all metrics, outperforming the baseline by a considerable margin. The collective assessment of success rate, object recovery, and unexplored space metrics unequivocally validates the efficacy of our system in exploring unfamiliar scenes through interactive processes. It is essential to highlight that in the case of object recovery, the baseline method may occasionally choose to randomly open certain drawers or doors to unveil objects. This randomness contributes to a seemingly higher object recovery rate for the baseline, which may not necessarily correlate with its overall success. The unexplored space metric shows that our system is much more stable in exploring all need-to-explore spaces.

Moreover, both the success rate and graph edit distance underscore the close alignment of our system with human actions, highlighting the efficiency of our approach across diverse scenarios. The state recovery metric assesses whether the final state post-exploration resembles the initial state. Our system consistently shows effective state recovery; however, the baseline may trick this metric by opting not to take any action, resulting in an artificially high score in this aspect.

Our results also underscore our system's ability to achieve robust and efficient exploration throughout the exploration process. Our system excels in efficiently discovering all concealed objects, whereas the baseline fails either due to a lack of early-stage actions or an inability to explore all need-to-explore spaces even upon completion. The analysis of errors in both our system and the baseline reveals the specific failure cases encountered by the baselines. In contrast, our system demonstrates enhanced robustness in both perception and decision-making.

Our qualitative results further illustrates various exploration scenarios along with their corresponding ACSG. These scenarios encompass ACSG with varying width or depth, highlighting our system's adaptive capability across diverse objects such as rigid, articulated objects, nested objects, and deformable objects. In addition, the scenario in Fig. 2 shows that our system is able to deal with the scenario with obstruction.

V. Conclusion

We introduced RoboEXP, a foundation-model-driven robotic exploration framework capable of effectively identifying all objects in a complex scene, both directly observable and those revealed through interaction. Central to our system is action-conditioned 3D scene graph, an advanced 3D scene graph that goes beyond traditional models by explicitly modeling interactive relations between objects. Experiments have shown RoboEXP's superior performance in interactive scene exploration across various challenging scenarios, significantly outperforming a strong GPT4V-based baseline. Notably, the reconstructed action-conditioned 3D scene graph is crucial for guiding complex downstream manipulation tasks, like preparing breakfast in a mock-kitchen environment with fridges, cabinets, and drawer sets.

REFERENCES

[1] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023. 2

[2] Yingdong Hu, Fanqi Lin, Tong Zhang, Li Yi, and Yang Gao. Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning. *arXiv preprint arXiv: 2311.17842*, 2023. 2

[3] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3

[4] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. *arXiv preprint arXiv: 2306.01567*, 2023. 3

[5] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 3

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *ICML*, 2021. 3

[7] OpenAI. Gpt-4v(ision) system card. *https://cdn.openai.com/papers/GPTV System Card.pdf*, 2023. 3

[8] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v(ision). *arXiv preprint arXiv: 2309.17421*, 2023. 3