# A Multimodal Chain of Tools for Described Object Detection

**Kwanyong Park**[1] **Youngwan Lee**[1,2] **Yong-Ju Lee**[1]

[1]Electronics and Telecommunications Research Institute (ETRI), South Korea
[2]Korea Advanced Institute of Science and Technology (KAIST), South Korea

## Abstract

Described object detection (DOD) is a promising direction for fine-grained and human-interactive visual recognition, where the goal is to detect target objects based on given language descriptions. Despite significant advancements in language-based object detection, current models still struggle with complex descriptions due to limited compositional understanding. To address this issue, we propose a novel multimodal chain-of-tools (MCoTs) framework that seamlessly integrates specialized tools to handle the two core functionalities of the DOD task: localization and compositional reasoning. Specifically, we decompose the complex DOD task into a series of subtasks, with each subtask handled by specialized tools, including detector and multimodal large language model (MLLM). This simple yet effective MCoTs framework demonstrates significant performance improvements on the challenging $D^3$ benchmark without additional training overhead.

## 1 Introduction

Detecting objects of interest has long been a central problem. Among the various task formulations, **Described Object Detection (DOD)** (Xie et al., 2024) has recently emerged as a particularly challenging problem, where the goal is to detect a target object based on complex language descriptions. This nuanced form of object detection is crucial for facilitating detailed human interaction in various downstream applications, such as robotics, interactive image/video editing, and image retrieval. This task involves two critical functionalities: (1) **compositionality**, where the model must faithfully understand complex expressions of visual objects given language queries, and (2) **localization**, where the model must precisely localize the target object in the form of a bounding box.

In recent years, a tremendous number of language-based object detection models (Liu et al., 2023) have been developed, utilizing language queries to specify target objects for tasks such as open-vocabulary object detection (Minderer et al., 2022), visual grounding (Li et al., 2022), and referring expression comprehension (Yu et al., 2016). However, these models have shown limited performance in the specific task of described object detection. Generally, they perform well on short and concise language queries (e.g., category names), but their performance drops significantly when faced with complex descriptions, exhibiting a limited compositional understanding.

To address this issue, recent works (Park et al., 2024; Zhao et al., 2024; Li et al., 2024b) have leveraged generative foundation models (Brown et al., 2020; Achiam et al., 2023) to produce synthetic data, aiming to enhance the compositional understanding of language-based detectors. Among these approaches, WSCL (Park et al., 2024) has proposed generating synthetic triplets consisting of images, descriptions, and bounding boxes using large language models (LLMs)(Brown et al., 2020; Achiam et al., 2023) and diffusion models(Chen et al., 2023; Podell et al., 2023). By combining synthetic data with a tailored learning framework, this approach has significantly improved performance on tasks involving complex and lengthy descriptions.

NeurIPS 2024 Workshop on Compositional Learning: Perspectives, Methods, and Paths Forward.

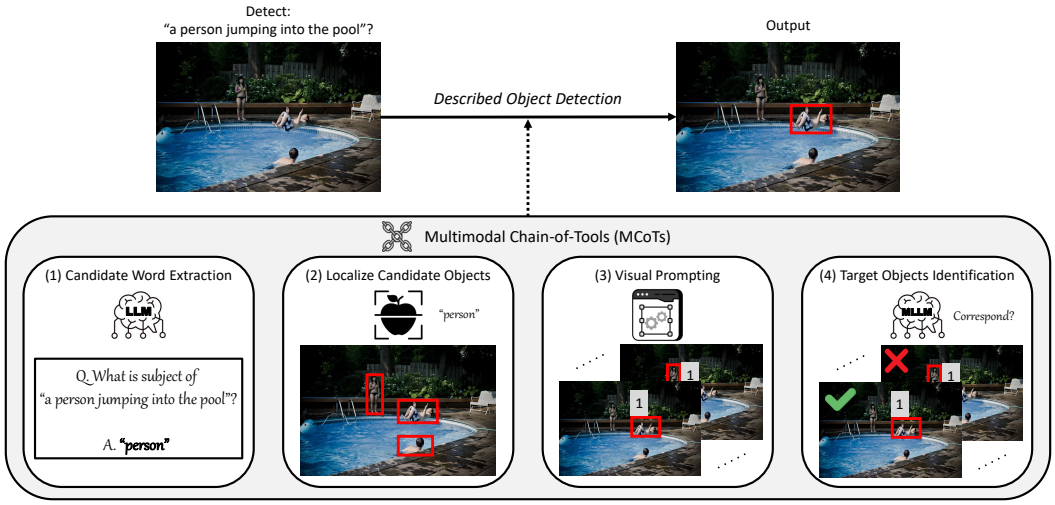


Figure 1: Overview of Proposed Multimodal Chain-of-Tools for Described Object Detection.

Despite these advancements, there is still room for improvement in the compositional understanding of language-based object detection models. These models still struggle more with lengthy descriptions than with concise ones. In contrast, the newly emerging multimodal large language models (MLLMs) (Liu et al., 2024; Li et al., 2024a) demonstrate unprecedented compositional understanding in language due to their equipped LLMs. They not only comprehend detailed language instructions but also exhibit high-level reasoning capabilities across the vision and language domains. This is also crucial for the task of described object detection. However, general-purpose MLLMs often fail to precisely localize objects within images. Moreover, fine-tuning these large models to achieve precise localization without compromising other aspects of performance remains a practical challenge.

To address this gap, we propose a novel unified framework that combines the strengths of both object detectors and MLLMs, achieving accurate localization and in-depth compositional understanding. To this end, we decompose the complex DOD task into a sequence of simpler and more tractable problems leveraging detectors and MLLMs. Each sub-task is then tackled by these models, which function as specialized tools. Specifically, we use detectors for obtaining several bounding boxes and then feed the candidate boxes into MLLM for reasoning whether the detected candidates meet the detailed conditions specified in the language query. We refer to this innovative approach as the "**Multimodal Chain-of-Tools (MCoTs)**". On the challenging $D^3$ benchmark (Xie et al., 2024), MCoTs outperforms all baselines, whether based on object detectors or MLLMs. Additionally, our method shows the most robust in handling the complexity of language queries. Notably, this MCoTs framework demonstrates both accurate localization and advanced compositional understanding in a training-free manner.

## 2 A Multimodal Chain of Tools for Described Object Detection

We introduce a novel multimodal chain-of-tools framework that decomposes complex described object detection tasks into several sub-tasks and addresses each one through a sequence of diverse specialized tools (i.e., chain-of-tools), including object detectors and multimodal large language models. First, we provide an overview of the proposed multimodal chain-of-tools framework in Section 2.1, followed by a detailed explanation of how the subtasks are defined and handled by the specialized tools in Section 2.2.

### 2.1 Overview

To design an efficient chain of tools for the DoD task, it is essential to thoroughly understand the challenges of the task as well as the strengths and weaknesses of the specialized tools involved.

The goal of described object detection is to detect target objects based on complex language queries, such as "person right next to a dog." To accomplish this, the system must demonstrate (1) accurate

localization (e.g., providing precise bounding boxes for the target "person") and (2) precise compositional reasoning (e.g., distinguishing the correct "person" from others in the image based on the given description).

From the perspective of these two capabilities, language-based object detectors (Liu et al., 2023) excel in localization but fall short in compositional reasoning. Specifically, these detectors often function like "bags-of-words," detecting all objects mentioned in the description (e.g., "person" and "dog") without considering the relationships or context described. In contrast, multimodal models excel at compositional reasoning but struggle to accurately localize target objects. General MLLMs (Liu et al., 2024; Li et al., 2024a), for instance, can provide reasonable reasoning in the form of language but fail to output precise bounding box coordinates.

With a deep understanding of the tasks and the specialized tools, our high-level approach is simple yet effective: "Let each tool focus on what it does best." First, the detector focuses solely on localization, without considering detailed descriptions (e.g., detecting all "persons" in the image). Then, the MLLM handles the reasoning by checking whether the detected candidates meet the detailed conditions specified in the language query. To implement this high-level concept, we design several subtasks for the DOD task. We explain these subtasks and the effective solutions for each in the following sections.

### 2.2 Detailed Process

The detailed process of the proposed multimodal chain of tools is illustrated in Figure 1.

**Chain 1: Candidate Word Extraction.** To ensure that the detector focuses on the subtask of localization, we need to provide a concise language query. For this purpose, we extract the essential word from the complex description that precisely represents the target objects. Specifically, we extract the subject noun from the given description. To accomplish this, we utilize an open-source large language model, Llama3 (Touvron et al., 2023), by querying it like, "`Please extract the subject noun in the following phrase.`".

**Chain 2: Localize Candidate Objects.** Once the subject noun is extracted, the detector's role is to locate all candidate objects in the image. By narrowing the detector's task to focus solely on detecting objects based on a single word (similar to understanding category names in open-vocabulary detection), we significantly simplify the task. It is important to note that the primary goal of this step is not to uniquely detect the target object but to ensure that the target object is included among the detected candidates.

**Chain 3: Visual Prompting for MLLMs.** The next subtask is to identify the correct target objects from the pool of detected candidates. For this, we leverage the compositional reasoning capabilities of MLLMs across the visual and language domains. One challenge is how to integrate these reasoning capabilities into the context of the DoD task. Inspired by set-of-mark prompting (Yang et al., 2023), we overlay bounding box on the image and mark them with number. This allows MLLMs, with their visual representation and OCR capabilities, to easily reference specific objects in the image. To reduce ambiguity and complexity in subsequent steps, we overlay one box at a time, marking it with "1". Examples of this visual prompting process are shown in Figure 1.

**Chain 4: Target Objects Identification with MLLMs.** Given the image with the marked candidate object, we prompt the MLLMs by asking whether the marked object corresponds to the original description. For this step, we use a rationalizing reasoning (Camburu et al., 2018) prompt, such as, "`Please begin by responding with yes or no, followed by a detailed explanation.`" We parse the MLLMs' answers to determine whether the marked object matches the given description. Only the candidates that correspond to the description are output as the final detection results. For this process, we leverage LLaVA-Onevision (Li et al., 2024a), a state-of-the-art open-source MLLM.

## 3 Experiments

**Baselines.** We compare our MCoTs framework with the following methods: (1) A multimodal large language model (**MLLM**), SPHINX (Lin et al., 2023), a recently proposed MLLM designed for region-level understanding; (2) language-based object detectors (**Detector**) originally designed for

Table 1: Performance comparison with state-of-the-art methods on $D^3$ (Xie et al., 2024) benchmark.

| Method | Type | Full | Pres | Abs | AP-S | AP-M | AP-L | AP-XL |
|---|---|---|---|---|---|---|---|---|
| SPHINX-7B | MLLM | 10.6 | 11.4 | 7.9 | - | - | - | - |
| OWL-ViT-L | Detector | 9.6 | 10.7 | 6.4 | 20.7 | 9.4 | 6.0 | 5.3 |
| Grounding-DINO | | 20.7 | 20.1 | 22.5 | 22.6 | 22.5 | 18.9 | 16.5 |
| OFA-DOD | | 21.6 | 23.7 | 15.4 | 23.6 | 22.6 | 20.5 | 18.4 |
| FIBER-B | | 22.7 | 21.5 | 26.0 | 30.1 | 25.9 | 17.9 | 13.1 |
| GENNEG-FIBER | Det+Syn | 26.0 | 25.2 | 28.1 | **35.5** | 29.7 | 20.5 | 14.2 |
| DESCO-FIBER | | 28.1 | 27.2 | 30.5 | 35.3 | 30.2 | 24.8 | 19.5 |
| WSCL-FIBER | | 30.8 | 31.0 | 30.4 | 33.9 | 33.7 | 27.8 | 22.8 |
| MCoTs (Ours) | MCoTs | **39.8** | **39.6** | **40.1** | 35.2 | **43.5** | **38.4** | **35.7** |

open-vocabulary detection or referring expression comprehension, including OWL-ViT (Minderer et al., 2022), Grounding-DINO (Liu et al., 2023), OFA-DOD (Xie et al., 2024), and FIBER (Dou et al., 2022); and (3) synthetic data-based frameworks (**Det+Syn**) that enhance compositional understanding in detectors through the use of generated synthetic data, including GENNEG (Zhao et al., 2024), DESCO (Li et al., 2024b), and WSCL (Park et al., 2024). For synthetic data-based methods, we benchmark their best model, which fine-tunes a language-based detector, FIBER.

**Evaluation Benchmark.** We compare our proposed framework with baseline methods on the $D^3$ (Xie et al., 2024) dataset, which provides an in-depth evaluation for detecting objects specified by complex descriptions. Unlike traditional referring expression benchmarks, $D^3$ introduces scenarios where descriptions either refer to no object or multiple instances in an image, allowing for a more comprehensive assessment of the models' compositional understanding. $D^3$ also offers a suite of sub-metrics for detailed analysis. Descriptions are classified into ABS ("absence") and PRES ("presence") categories, based on whether the description includes expressions of absence (e.g., "without"). In addition to the overall evaluation metric, which covers all descriptions (referred to as FULL), $D^3$ provides distinct metrics for ABS and PRES. The AP-S/M/L/XL sub-metrics categorize performance based on the length of descriptions (short, medium, long, and very long), offering insights into detection performance relative to the complexity of the descriptions.

**Experimental Results.** The quantitative results are summarized in Table 1. Our MCoTs framework significantly outperforms all existing baselines by a large margin, demonstrating the effectiveness of our approach in both localization and compositional understanding. The performance improvements are consistent, regardless of whether the descriptions include expressions of absence (as seen in the PRES and ABS scores). More importantly, MCoTs excels in handling both concise and complex, lengthy descriptions, as indicated by the AP-S/M/L/XL scores. In contrast, all previous methods showed a monotonic degradation in performance as the length and complexity of descriptions increased. These results highlight the unique advantage of MCoTs by seamlessly combining the strengths of both detector and MLLM: precise localization and advanced compositional understanding.

## 4 Conclusion

In this paper, we introduce a novel multimodal chain-of-tools framework for described object detection (DOD). We first identify two key functionalities essential for the task: precise localization and advanced compositional understanding. To address these, we leverage specialized tools, including object detectors and multimodal large language models (MLLMs), for each functionality. Specifically, we break down complex DOD tasks into a chain of subtasks and allow specialized tools to focus on individual tasks. This simple yet effective multimodal chain-of-tools framework achieves impressive performance without additional training overhead. In future work, we will focus on automating the multimodal chain-of-tools framework for a wider range of real-world applications.

**Acknowledgement** This work was supported by the Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2022-0-00871, Development of AI Autonomy and Knowledge Enhancement for AI Agent Collaboration).

## References


Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Camburu, O.-M., Rocktäschel, T., Lukasiewicz, T., and Blunsom, P. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31, 2018.

Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

Dou, Z.-Y., Kamath, A., Gan, Z., Zhang, P., Wang, J., Li, L., Liu, Z., Liu, C., LeCun, Y., Peng, N., et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. *Advances in neural information processing systems*, 35:32942–32956, 2022.

Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Li, Y., Liu, Z., and Li, C. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Li, L., Dou, Z.-Y., Peng, N., and Chang, K.-W. Desco: Learning object recognition with rich language descriptions. *Advances in Neural Information Processing Systems*, 36, 2024b.

Li, L. H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N., et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10965–10975, 2022.

Lin, Z., Liu, C., Zhang, R., Gao, P., Qiu, L., Xiao, H., Qiu, H., Lin, C., Shao, W., Chen, K., et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., et al. Simple open-vocabulary object detection. In *European Conference on Computer Vision*, pp. 728–755. Springer, 2022.

Park, K., Saito, K., and Kim, D. Weak-to-strong compositional learning from generative models for language-based object detection. *arXiv preprint arXiv:2407.15296*, 2024.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Xie, C., Zhang, Z., Wu, Y., Zhu, F., Zhao, R., and Liang, S. Described object detection: Liberating object detection with flexible expressions. *Advances in Neural Information Processing Systems*, 36, 2024.

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

Yu, L., Poirson, P., Yang, S., Berg, A. C., and Berg, T. L. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 69–85. Springer, 2016.

Zhao, S., Zhao, L., Suh, Y., Metaxas, D. N., Chandraker, M., Schulter, S., et al. Generating enhanced negatives for training language-based object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13592–13602, 2024.