CoTDet: Affordance Knowledge Prompting for Task Driven Object Detection

Jiajin Tang* Ge Zheng* Jingyi Yu Sibei Yang†
School of Information Science and Technology, ShanghaiTech University
{tangjj,zhengge,yujingyi,yangsb}@shanghaitech.edu.cn
https://toneyaya.github.io/cotdet

Abstract

Task driven object detection aims to detect object instances suitable for affording a task in an image. Its challenge lies in object categories available for the task being too diverse to be limited to a closed set of object vocabulary for traditional object detection. Simply mapping categories and visual features of common objects to the task cannot address the challenge. In this paper, we propose to explore fundamental affordances rather than object categories, i.e., common attributes that enable different objects to accomplish the same task. Moreover, we propose a novel multi-level chain-of-thought prompting (MLCoT) to extract the affordance knowledge from large language models, which contains multi-level reasoning steps from task to object examples to essential visual attributes with rationales. Furthermore, to fully exploit knowledge to benefit object recognition and localization, we propose a knowledge-conditional detection framework, namely CoTDet. It conditions the detector from the knowledge to generate object queries and regress boxes. Experimental results demonstrate that our CoTDet outperforms state-of-the-art methods consistently and significantly (+15.6 box AP and +14.8 mask AP) and can generate rationales for why objects are detected to afford the task.

1. Introduction

The traditional object detection task [3, 12, 37, 65], as shown in Figure 1a, aims to detect object instances of given categories in an image, so objects of categories such as the cup, bottle, cake, and knife are detected. In contrast, detection tasks in real-world applications [1, 4], such as intelligent service robots, usually appear in the form of tasks rather than object categories [38]. For example, when an intelligent robot is asked to complete the task of *"opening parcels"*, the robot needs to autonomously locate the tool shown in Figure 1b, *i.e.*, a knife. So the core of this type of task is to detect the object instances in the image that are most suitable for the task [19, 42]. However, the challenges for task driven object detection are multi-fold. Previous methods that directly learn the mapping between objects and tasks from objects' visual context features or categories cannot achieve satisfactory results. As shown in Figure 1c, the context-based approach GGNN [42] wrongly chooses *vegetable stem* for afford task of *"opening parcels"* because it learns that visually slender objects can open parcels. Similarly, the category-based approach TOIST [19] considers that no object in the image can perform the *"opening parcels"* since it could not even recognize an instance of the knife in the image. In contrast, people will intelligently and naturally choose to use the knife to open parcels in the scene of Figure 1b.

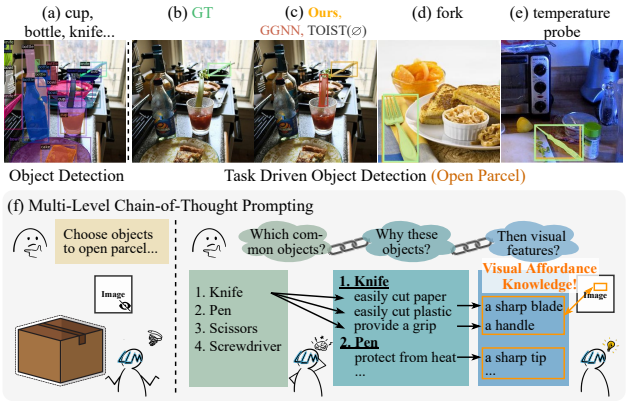


Figure 1. **An example** of (a) traditional object detection, (b)-(e) task driven object detection, and (f) our multi-level chain-of-thought (MLCoT) prompting large language models (LLMs) to generate visual affordance knowledge.

Recently, Large Language Models (LLMs) like GPT-3 [2] and ChatGPT [31] have demonstrated impressive capabilities in encoding general world knowledge from a vast amount of text data [32, 40, 52]. A naive approach is to prompt LLMs directly to return "what objects should we use to *open parcels*" and leverage the returned object categories to simplify task driven object detection to the traditional one. However, we observe that LLMs usually only

*Both authors contributed equally. †Corresponding author.

return a few categories of commonly used objects, such as the knife, pen, paper cutter, and scissors shown in Figure 1f. According to these categories, although the knife in Figure 1b can be identified as the target object, the detection system will miss other objects that also can be used to open parcels, such as the fork in Figure 1d and the temperature probe next to the microwave oven in Figure 1e. In turn, we ask why people can easily lock the fork and temperature probe in Figure 1d and 1e as the target objects? We argue that the reason is that people are not restricted to using specific categories of objects to accomplish a task but instead select objects based on the commonsense knowledge that objects with *"a handle and sharp blade"* or *"a handle and sharp tip"* can *"open parcels"*.

In this paper, we propose to explicitly acquire visual affordance knowledge of the task (*i.e.*, common visual attributes that enable different objects to afford the task) and utilize the knowledge to bridge the task and objects. Figure 1f shows two sets of visual affordance knowledge (marked inside the yellow box) for *opening parcels*. However, it is not trivial to acquire such task-specific visual affordance knowledge. Furthermore, we propose a novel multi-level chain-of-thought prompting (MLCoT) to elicit visual affordance reasoning from LLMs. At the first level (object level), we prompt LLMs to return common objects by the above-mentioned approach. Unlike before, which treats the returned object categories as target categories, we instead treat this query progress as brainstorming to obtain representative object examples. At the second level (affordance level), we generate rationales from LLMs for why object examples can afford the task and cooperate rationales to facilitate LLMs to reason and summarize the visual affordances beyond object examples. As shown in Figure 1f, the rationale and visual affordances that enable the knife to open parcels are *"easily cut through paper and plastic..."* and *"a sharp blade and handle"*, respectively. Our MLCoT can capture the essence of visual affordances behind these object examples without being limited to object categories. Thus we can successfully detect the fork and temperature probe in Figure 1d and 1e as they meet the visual affordances required by the task.

Moreover, we claim that visual affordance knowledge not only helps recognize and identify objects suitable for the task but also helps localize objects more precisely because visual attributes such as color and shape are useful in object localization. Therefore, unlike some methods [9, 30, 44, 46, 53] to take knowledge as complementary to the image's visual features, we condition the detector on the visual affordance knowledge to perform knowledge-conditional object detection. Specifically, we follow [19] to use an end-to-end query-based detection framework [3, 65]. But instead of randomly initializing queries, we generate knowledge-aware queries based on image features and visual affordance knowledge. In addition to generating queries, we use visual affordance knowledge to guide the bounding box regression explicitly, inspired by the denoising training [18]. Unlike [18] introduces denoising for accelerating training, our knowledge-conditional denoising training aims to teach the decoder how to utilize visual knowledge to regress the boxes for queries.

Finally, we propose the CoTDet network, which acquires visual affordance knowledge from LLMs via the proposed MLCoT and performs knowledge-conditional object detection to effectively utilize the knowledge. Moreover, our CoTDet can easily be extended to task driven instance segmentation by employing a segmentation head [5, 15]. In summary, our main contributions are:

- We are the first to propose to explicitly acquire visual affordance knowledge and utilize the knowledge to bridge the task and object instances.
- We propose a novel multi-level CoT prompting (MLCoT) to make abstract affordance knowledge concreted, which leverages LLMs to generate and summarize intermediate reasoning steps from object examples to essential visual attributes with rationales.
- We claim that visual affordance knowledge can benefit both object recognition and localization and propose a knowledge-conditional detection framework to condition the detector to generate object queries and guide box regression through denoising training.
- Our CoTDet not only consistently outperforms state-of-the-art methods ($+15.6$ $AP50_{box}$ and $+14.8$ $AP50_{mask}$) by a large margin but also can generate rationales for why objects are detected to afford tasks.

2. Related Work

Task Driven Object Detection/Instance Segmentation aims to detect or segment out the most suitable objects in an image to afford a given task, such as *opening parcel* or *getting lemon out of tea*. Different from traditional object detection or instance segmentation [3, 5, 10, 12, 37, 65], it requires modeling the preference for selecting objects based on a comprehensive understanding of the specific task and the image scene. Although referring expression grounding [6, 15, 41, 45, 54, 56, 57] and segmentation [14, 20, 47, 51, 55, 58, 60, 62] similarly locate the referent according to a natural language description, they rely on the object attributes and relations in the scene for localization without considering the prior knowledge needed to afford the given task. These factors result in the challenge of task driven object detection which requires complex and joint knowledge reasoning on the requirements for one specific task and objects' functional attributes beyond visual recognition and scene understanding.

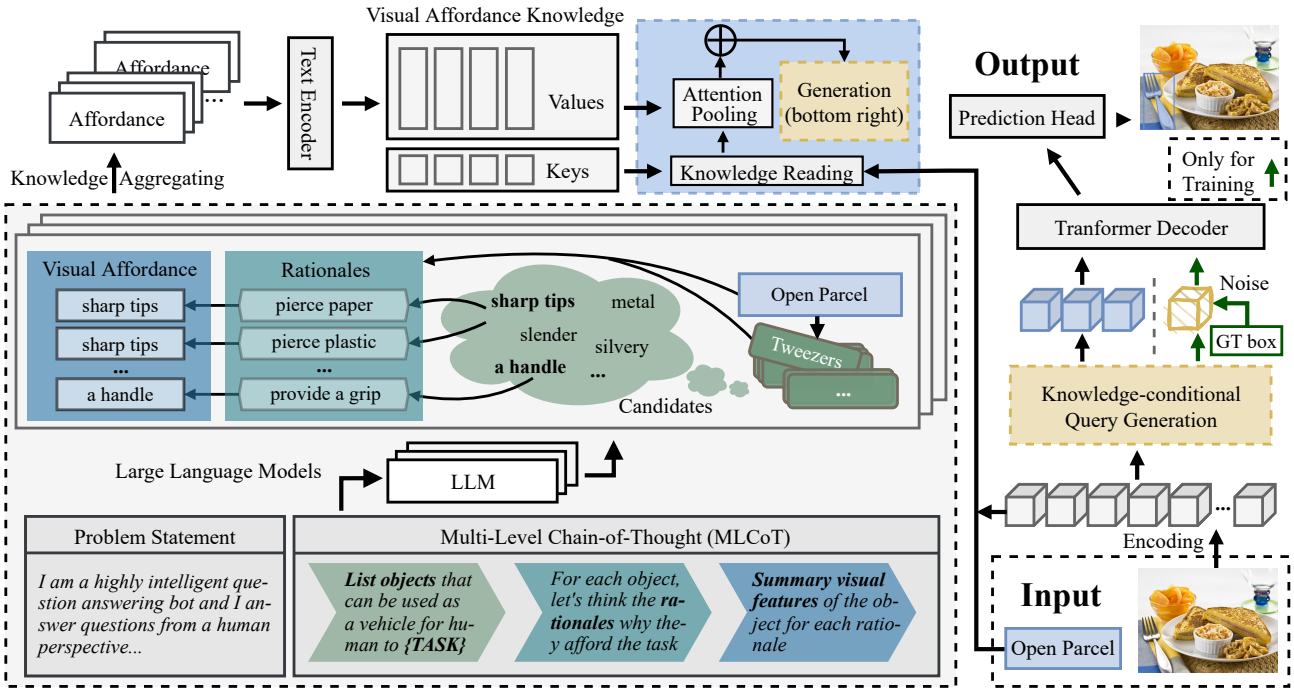


Figure 2. **Overall framework of the proposed CoTDet with multi-level chain-of-thought prompting (MLCoT).** We first generate visual affordance knowledge from LLMs with the proposed novel MLCoT. Next, we perform knowledge-conditional object detection by utilizing the knowledge to generate object queries for the scene as well as guide object localization through denoising training.

The pioneering work [42] adopts a two-stage framework that first detects objects and then compares among objects to select suitable objects via graph neural networks [43]. The following work TOIST [19] distills target object names to pronouns such as something by conditioning on the language description of the task. However, these works lack to model explicit requirements of tasks and objects' affordances to tasks, which limits their performance and generalization capabilities.

Knowledge Acquisition in Vision-Language Tasks. Integrating external knowledge into computer vision tasks [16, 35,44,46] and vision-language tasks [29,49,50,63] has been found beneficial. Previous methods [16, 28, 30, 35, 53, 61] are interested in acquiring structured knowledge (e.g., ConceptNet [22]), which usually includes commonsense concepts and relations and is presented in fixed data structures such as the graph or triple. Recently, large language models [2, 7, 36] have been demonstrated to learn open-world commonsense knowledge from the large-scale corpus [32, 40]. Some works [9, 59] utilize language models to encode the representations of inputs or directly generate answers conditioning on visual inputs, leveraging the latent knowledge in language models. Unlike previous works, we prompt language model GPT-3 [8] to obtain external knowledge explicitly with the chain of thought (CoT) [25, 52, 64] for better interpretability. To the best of our knowledge, we are the first to explore CoT to acquire visual commonsense knowledge in text form, leveraging the reasoning ability of the language model to filter effective visual functional attributes to afford tasks.

External Knowledge Utilization. Incorporating knowledge reasoning has attracted growing interest in computer vision [11,26,34,35,44,63] and vision-language [49] fields such as object detection [44, 46], visual relationship detection [16, 61] and visual question answering [9, 28, 30, 53]. For tasks that focus on capturing relations among objects, such as scene graph generation and visual relationship detection, extracting knowledge of the interactions between general concepts becomes natural [11, 16, 61, 63]. Other tasks like object detection and image classification rely on category-related knowledge that is retrieved in the knowledge base as definitions or attributes of general concepts [9, 26, 34, 44, 46, 53]. For knowledge utilization, they mainly directly take external knowledge as an expansion to visual content or explicitly constrain the consistency between knowledge and visual content. Different from existing works, we leverage attribute-level commonsense knowledge about requirements for completing tasks and take external knowledge of tasks as the condition to condition the detector for task driven object detection.

3. Method

The framework of our proposed CoTDet is shown in Figure 2. First, we introduce the problem definition and image and text encoders in Section 3.1. Second, we ac-

quire visual affordance knowledge from LLMs by leveraging the multi-level chain-of-thought prompting and aggregation (Section 3.2). Next, we present the knowledge-conditional decoder that conditions acquired knowledge to detect and segment suitable objects (Section 3.3). Finally, we introduce the loss functions in Section 3.4.

3.1. Problem Definition and Encoder

Problem Definition. Given a task in text form S (*e.g.*, *"open parcel"*) and an image I, task driven object detection requires detecting a set of objects most preferred to afford the task. Note that the target objects indicated by the task are not fixed in quantity and category, which may vary with changes in the image scene. In contrast, traditional object detection [3, 10, 65] detects objects of fixed categories while referring image grounding [6, 17, 27, 33] localizes unambiguous objects.

Encoders. For the task S, we leverage the linguistic information from two perspectives. First, the text S is preserved as input for extracting knowledge from LLMs. Besides, we employ RoBERTa [23] as the text encoder to obtain the text feature t_s for the task S, which will be used to query the task-relevant visual content in the image. For the image I, we adopt the ResNet-101 [13] as the visual backbone to extract multi-scale visual feature maps and flatten the maps along the spatial dimension to features V.

3.2. Visual Affordance Knowledge from LLMs

To detect objects to afford a particular task, we naturally consider the task's requirements first and subsequently localize suitable objects that meet the requirements. Nevertheless, the task requirements are abstract and can not directly correspond to the visual content in the image for localizing the objects. Motivated by this, we propose to explicitly extract common visual attributes (*i.e.*, visual affordances) that make different objects can afford the task and use visual affordances to bridge task requirements and object instances in the image. Furthermore, we generate visual affordances from LLMs because LLMs generally contain world knowledge learned from a vast amount of text data.

Specifically, we first design a novel multi-level chain-of-thought prompting to leverage LLMs to generate visual affordances (see Section 3.2.1) and then encode and aggregate them automatically to be utilized for detection (see Section 3.2.2).

3.2.1 Multi-Level Chain-of-Thought Prompting

Our multi-level chain-of-thought prompting (MLCoT) leverages LLMs to generate and summarize intermediate reasoning steps from object examples to essential visual attributes with rationales. MLCoT first brainstorms the object examples to afford the task and then considers rationales why the examples can afford the task and summarize corresponding visual attributes for rationales.

Object-level Prompting as Brainstorming. At the first level, we prompt LLMs to generate daily object examples that afford the input task S. Specifically, we design the following text prompt:

Prompt: What common objects in daily life can be used as a vehicle for the human to [task]? Please list the twenty most suitable objects.
Output: knife, pen, paper cutter, scissors, screwdriver, ...

Where [task] is filled with the task text S. We denote the number of objects returned from LLMs as N_o. For simplicity, we present the most critical parts of the prompts. For complete prompts, please refer to Appendix. One straightforward idea is to perform traditional category-specific object detection with respect to the categories of object examples. However, it is not feasible due to the following observations: (1) object examples are overly limited to partial object categories, resulting in the gap between object categories and the actual task requirements. For instance, the fork in Figure 2 is not among the objects returned by LLMs. (2) a few noisy unsuitable objects may be output. Although the noisy objects are few, relying entirely on the object examples is risky. For example, for the task of extinguishing fire, LLMs return the fire axe, a common firefighting tool, but it cannot be directly used to extinguish the fire.

Affordance-level Prompting with Rationales. To address the above challenges and capture the essential visual affordances implied in representative object examples, we propose to generate rationales for why these objects can afford the task and summarize visual affordances from rationales. With the object examples, we prompt LLMs to generate rationales as follows:

Prompt: For each object, let's think about the rationales for why they afford the task from the perspective of visual features.
Output: Knife: They have a sharp blade which can easily cut through paper and plastic; They have a handle which provides a good grip for the user. Pen: ...

By the above prompting, we get a set of rationales for each object. Next, we further prompt LLMs to summarize visual-related rationales to form visual affordances as follows:

Prompt: Summary corresponding visual features of the object for each rationale.
Output: {*A sharp blade and a handle.* }, {...}, ...

Finally, we obtain N_o sets of visual affordances, where each set contains several visual attributes relevant to why objects can afford the task S. And we define each set of visual affordance knowledge as a knowledge unit. Note that although each knowledge unit is derived from the rationales of an object example, the affordance knowledge in that unit is not

limited to that object example or its related categories. For example, the first visual affordance unit comprises *"a sharp blade and a handle"*, which correspond to the returned object *"knife"*. Notably, this visual affordance unit also applies to *"box cutter"* and *"paper cutter"*.

3.2.2 Knowledge Encoding and Aggregating

We further extract a refined knowledge base by filtering out a few knowledge units corresponding to the unsuitable object examples mentioned in Section 3.2.1. For each knowledge unit, we concatenate textual descriptions of its visual affordances into a textual sequence and then utilize RoBERTa [23] to obtain the sentence feature. To filter out unsuitable units, we compute the cosine similarity between each pair of knowledge units and exclude outlier units if their maximum similarity to other units falls below a predetermined threshold. Additionally, for each selected unit, we extract its word representations via RoBERTa. In summary, we aggregate N visual affordance knowledge units, denoted as $\mathcal{K} = \{p_j^k, p_j^v\}_{j=1}^N$, where p_j^k and p_j^v are the sentence feature and word features of j-th unit, respectively.

3.3. Knowledge-conditioned Decoder

We base on the detection architecture of Deformable-DETR [65], a DETR-like detector [3, 5, 18], which uses object queries to capture object-level information for detection (Section 3.3.1). Unlike randomly initializing the object queries, we leverage visual affordance knowledge to generate the object queries (Section 3.3.2) and guide the bounding box regression with denoising training (Section 3.3.3).

3.3.1 Introduction to Deformable-DETR

Deformable-DETR contains a Transformer encoder and a Transformer decoder. The encoder inputs visual features V and outputs the refined visual features $F = \{f_1, f_2, ..., f_i\}$ via multi-scale deformable attention. The decoder randomly initializes queries $Q = \{q_1, q_2, ..., q_k\}$ and predicts a reference point p_k for each object query q_k, and these reference points $L = \{l_1, l_2, ..., l_k\}$ serve as the initial guess of the box centres. Next, the decoder searches for objects O for these queries Q with reference points L via multi-scale deformable cross-attention and self-attention, which is formulated as follows,

$$O = \text{Deformable}([Q, L], F), \tag{1}$$

where $\text{Deformable}(\cdot, \cdot)$ denotes the Transformer decoder of Deformable-DETR.

3.3.2 Knowledge-conditional Query Generation

Instead of randomly initializing the queries, we generate the queries and their reference points based on the visual content of the image, the task, and the visual affordance knowledge. Specifically, we utilize visual affordance knowledge to select visual features and combine them with the knowledge to generate queries, and then the spatial information in these visual features naturally becomes the spatial priors of reference points.

Given visual features $F = \{f_1, f_2, ..., f_i\}$, we first fuse each feature f_i with the task's text feature t_s, and then calculate its relevance to the task's visual affordance knowledge $\mathcal{K} = \{(p_j^k, p_j^v)\}_{j=1}^N$. Since each knowledge unit (p_j^k, p_j^v) in the knowledge base $\mathcal{K}$ is a set of affordances that meet the task requirements, we use the fused feature's largest similarity to each knowledge unit (p_j^k, p_j^v) in the knowledge base $\mathcal{K}$ as the feature's relevance score. And the calculation is formulated as follows,

$$\begin{aligned} s_{i,j} &= \cos(\text{fc}(f_i) + \text{fc}(t_s), p_j^k), \\ r_i &= \max_j(s_{i,j}), d_i = \text{argmax}_j(s_{i,j}), \end{aligned} \tag{2}$$

where $\cos(\cdot, \cdot)$ computes the cosine similarity, $\text{fc}(\cdot)$ represents the fully connected layer, and $s_{i,j}$ is the similarity between i-th visual feature and j-th knowledge unit. Then, r_i and d_i mean the i-th visual feature's relevance score and index of the corresponding knowledge unit, respectively.

Next, we select the visual features with the top-k largest relevance scores $\{r_i\}$ to incorporate their corresponding knowledge $\{(p_{d_i}^k, p_{d_i}^v)\}$ to generate queries Q^{kn} as follows,

$$Q^{\text{kn}} = \text{topk}_{r_i}\{f_i + \text{AttentionPool}(f_i, p_{d_i}^v)\}, \tag{3}$$

where topk_{r_i} means to select the corresponding features with the top-k largest relevance scores r_i. The attention pooling layer [48] $\text{AttentionPool}(f_i, p_{d_i}^v)$ returns the weighted features on $p_{d_i}^v$ based on their similarities to the f_i. Note that, for each knowledge unit (p_j^k, p_j^v), we use its global sentence feature p_j^k to compute its overall similarity to each visual feature in Eq. 2 while adopting word-level features p_j^v to better enhance the query's fine-grained representations in Eq. 3. Similar to Deformable-DETR, we further predict the reference points L^{kn} from the queries Q^{kn}. In addition, to facilitate the learning of Top-k selection, the selected queries Q^{kn} are directly fed into the prediction heads and supervised during training using the same training loss in Section 3.4.

3.3.3 Knowledge-conditional Decoding

With queries Q^{kn}, reference points L^{kn}, and the refined visual features F, we apply the Deformable decoder to search objects O^{kn} as follows,

$$O^{\text{kn}} = \text{Deformable}([Q^{\text{kn}}, L^{\text{kn}}], F). \tag{4}$$

In addition to utilizing visual affordance knowledge for query generation and providing the decoder with prior

task1: *step on*	task2: *sit comfortably*	task3: *place flowers*	task4: *get potatoes out of fire*	task5: *water plant*
task6: *get lemon out of tea*	task7: *dig hole*	task8: *open bottle of beer*	task9: *open parcel*	task10: *serve wine*
task11: *pour sugar*	task12: *smear butter*	task13: *extinguish fire*	task14: *pound carpet*	

Method	Task(AP_{box}@0.5)														Mean
	task1	task2	task3	task4	task5	task6	task7	task8	task9	task10	task11	task12	task13	task14	
GGNN [42]	36.6	29.8	40.5	37.6	41.0	17.2	43.6	17.9	21.0	40.6	22.3	28.4	39.1	40.7	32.6
TOIST [19]	44.0	39.5	46.7	43.1	53.6	23.5	52.8	21.3	23.0	46.3	33.1	41.7	48.1	52.9	41.3
TOIST† [19]	45.8	40.0	49.4	49.6	53.4	26.9	58.3	22.6	32.5	50.0	35.5	43.7	52.8	56.2	44.1
Ours	**58.9**	**55.0**	**51.2**	**68.5**	**60.5**	**47.7**	**76.9**	**40.7**	**47.4**	**66.5**	**41.9**	**48.3**	**61.7**	**71.4**	**56.9**

Table 1. **Comparison with state-of-the-art models for task driven object detection on COCO-Tasks dataset.** † means the model is with noun-pronoun distillation.

knowledge, we further improve the knowledge utilization by designing a knowledge-based denoising training [18]. As the visual affordance knowledge indicates the target objects' visual attributes, such as shape and size, the knowledge-base denoising guides the decoder in learning how to use this kind of visual knowledge to regress the targets' boxes.

Specifically, during the training stage, we first randomly add noise to ground-truth boxes $O^{\text{gt}} = \{o_m^{\text{gt}}\}_{m=1}^{M}$ to construct the noised objects following DN-DETR [18] and then extract noised boxes' visual features and centers as the noised queries $F^{\text{noise}} = \{f_m^{\text{noise}}\}_{m=1}^{M}$ and reference points L^{noise}, respectively. Notice that the previous denoising training method [18] adds noise to both boxes and categories labels to capture label-box relations better. But we only add noise to boxes because we aim to utilize the knowledge without noise to help denoise boxes. Therefore, we extract the knowledge unit $(p_{d_m}^k, p_{d_m}^v)$ for each ground-truth box o_m^{gt} through Eq. 2. Finally, the knowledge units $\{(p_{d_m}^k, p_{d_m}^v)\}_{m=1}^{M}$ guide the decoder to regress the ground-truth boxes O^{gt} from the noised queries F^{noise}, which is formulated as follows,

$$\begin{aligned} P^{\text{kn}} &= \{\text{AttentionPool}(f_m^{\text{noise}}, p_{d_m}^v)\}_{m=1}^{M} \\ O^{\text{denoise}} &= \text{Deformable}([F^{\text{noise}} + P^{\text{kn}}, L^{\text{noise}}], F,) \end{aligned} \tag{5}$$

where P^{kn} is the visual affordance knowledge of noised queries, and the $\text{Deformable}(\cdot,\cdot)$ in Eq. 4 and Eq. 5 shares the same parameters. And the denoising is only considered in the training stage.

3.4. Loss Functions

Following DETR [3], we use bipartite matching to find the unique predictions for the ground-truth objects and adopt the same bounding box regression loss $\mathcal{L}_{box}$ consisting of L1 loss and GIoU [39] loss. Moreover, we use the binary cross entropy loss as the classification loss $\mathcal{L}_{cl}$. The overall loss is represented as:

$$\mathcal{L}_{cost} = \lambda_{cl}\mathcal{L}_{cl} + \lambda_{box}\mathcal{L}_{box}, \tag{6}$$

where λ_{cl} and λ_{box} are the hyperparameters of the weighted loss. Our method can be easily extended to instance segmentation by adding a segmentation head [5] and replacing the box regression loss with the Dice loss $\mathcal{L}_{mask}$.

4. Experiment

4.1. Dataset and Implementation Details

Dataset. We conduct experiments on the COCO-Tasks dataset [42], which comprises 14 different tasks (see Table 1). This dataset is derived from the COCO dataset [21], but with customized annotations for task driven object detection. Each task contains 3600 training and 900 testing images. Besides, we follow [19] to incorporate mask annotations to the original COCO-Tasks dataset for the instance segmentation benchmark.

Implementation Details. Following previous works [19, 42], we use ResNet-101 [13] as the image encoder and RoBERTa [23] as the text encoder. The model is pre-trained on the COCO dataset but images already part of COCO-Tasks are removed. We train the model for 4000 iterations with the initial learning rate 1e-4 and use AdamW [24] as the optimizer. The hyperparameters λ_{cl} and λ_{box} are 4 and 5. Following [19], we evaluate the segmentation and detection performance of each task using AP_{mask}@0.5 and AP_{box}@0.5, respectively. And we denote their means across all tasks as mAP_{mask} and mAP_{box}. Unless otherwise specified, we leverage the GPT-3 [2] to extract visual affordance knowledge due to its capability to generate rationales [52].

4.2. Comparison with State-of-the-Art Methods

Table 1 and Table 2 show the comparison of our CoTDet with state-of-the-art models (SOTAs) on detection and segmentation benchmarks. Our model consistently outperforms the SOTAs [19, 42] on all benchmarks and tasks.

Comparison with SOTAs. Compared to TOIST [19], our CoTDet achieves significant performance improvement (15.6% mAP_{box} and 14.8% mAP_{mask}), which demonstrates the effectiveness of our task-relevant knowledge acquisition and utilization. Compared to the two-stage

Method	Task(AP_{mask}@0.5)														Mean
	task1	task2	task3	task4	task5	task6	task7	task8	task9	task10	task11	task12	task13	task14	
GGNN [42]	31.8	28.6	45.4	33.7	46.8	16.6	37.8	15.1	15.0	49.9	24.9	18.9	49.8	39.7	32.4
TOIST [19]	37.0	34.4	44.7	34.2	51.3	18.6	40.5	17.1	23.4	43.8	29.3	39.9	46.6	42.4	35.2
TOIST† [19]	40.8	36.5	48.9	37.8	43.4	22.1	44.4	20.3	26.9	48.1	31.8	34.8	51.5	46.3	38.8
Ours	**55.0**	**51.6**	**51.2**	**57.7**	**60.1**	**43.1**	**65.9**	**40.4**	**45.4**	**64.8**	**40.4**	**48.7**	**61.7**	**64.4**	**53.6**

Table 2. **Comparison with state-of-the-art models for task driven instance segmentation on COCO-Tasks dataset.** † means the model is with noun-pronoun distillation.

Ablation	task2	task6	task9	**mAP_{box}**	**mAP_{mask}**
"objects"	25.4	16.5	21.0	31.9	31.3
"visual"	50.4	30.3	38.3	48.1	44.7
w/o rationales	52.0	40.7	41.2	52.4	49.0
MLCoT	55.0	47.7	47.5	56.9	53.6
MLCoT(ChatGPT)	50.6	48.1	50.3	57.0	54.0
Def+GGNN [42]	38.6	24.7	23.4	38.8	35.8
Def+TOIST [19]	43.4	21.0	29.0	40.3	37.6
Init w/ MLCoT	42.2	35.9	35.6	48.7	46.4
Fuse w/ MLCoT	44.0	42.3	41.2	50.6	47.7
Select w/ MLCoT	50.0	47.2	43.7	55.3	51.7
Full Decoder	55.0	47.7	47.5	56.9	53.6

Table 3. **Ablation study** about knowledge acquisition, detection framework, and knowledge utilization of our CoTDet.

method GGNN [42], we achieve 24.3% mAP_{box} and 21.2% mAP_{mask} performance gain, which demonstrates the importance of leveraging the visual affordance knowledge rather than purely visual context information.

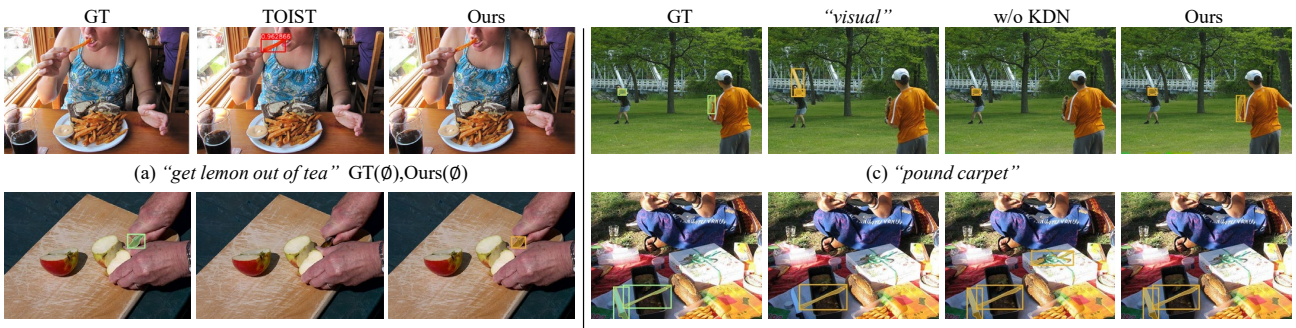
Comparison on Sub-tasks. The following comparisons on sub-tasks further demonstrate that the affordance-level knowledge is capable of bridging tasks and objects. Our CoTDet significantly improves the detection and segmentation performance on task4 (*get potatoes out of fire*), task6 (*get lemon out of tea*) and task7 (*dig hole*), achieving approximately 20% mAP improvement on both benchmarks. These tasks face the common challenge of the wide variety of targets' categories and visual appearances, which is hardly dealt with by methods like [19, 42] that merely learn the mapping between tasks and objects' categories and visual features. In contrast, our method explicitly acquires the visual affordance knowledge of tasks to detect rare objects and avoid overfitting to common objects, outperforming significantly in these tasks. In addition, for those less challenging tasks with a few ground-truth object categories, we still achieve approximately 8% mAP improvement, demonstrating the effectiveness of conditioning on visual affordances to object localization.

4.3. Ablation Study

We evaluate seven variants of our CoTDet and two SOTAs with the same backbones as ours to validate the effectiveness of the proposed knowledge acquisition and utilization. The results are shown in Table 3. In addition to mAP, we report AP_{box}@0.5 on relatively easy task2 (*sit comfortably*) and challenging task6 (*get lemon out of tea*) and task9 (*open parcel*) for reference, and the full results and analysis on sub-tasks are provided in Appendix.

MLCoT Prompting for Knowledge Acquisition. To evaluate the impact of core designs in MLCoT, we replace the MLCoT pipeline with the following approaches and utilize the acquired knowledge as the condition to guide detection: (1) We encode the object categories returned by the object level of MLCoT as the knowledge to perform the knowledge-conditional object detection. The results (31.9% mAP_{box} and 31.3% mAP_{mask}) demonstrate that simply extracting the object categories from LLMs cannot achieve satisfactory performance. (2) We attempt to acquire affordance-level rather than object-level knowledge. Specifically, we prompt LLMs by asking *"what visual features can we use to determine the suitability of an object for {TASK}?"* to generate visual affordance knowledge directly. The attempt improves the above object-level model by 16.2% mAP_{box} and 13.4% mAP_{mask}, showing the necessity of exploring the essential visual affordances behind the object categories. However, this model still underperforms by approximately 9% mAP compared to our full model. It is difficult to summarize a unified description of widely varying objects without priors, resulting in only one set of visual attributes being returned from LLMs. (3) To increase the diversity of visual affordances, we prompt LLMs to generate visual features for each object retrieved, which leads to a significant improvement to 52.4% mAP_{box} and 49.0% mAP_{mask}. (4) Finally, we further add rationales to filter out the misleading and irrelevant attributes, achieving a 4.5% and 4.6% increase in mAP_{box} and mAP_{mask}, respectively. (5) We also evaluate the effect of using different LLMs to extract visual affordance knowledge. Our MLCoT with ChatGPT [31] has a similar mAP to MLCoT with GPT-3.

Knowledge-conditional Object Detection. To validate the effectiveness of our proposed knowledge-conditioned decoder, we conduct ablation studies with two baselines and three variants based on Deformable-DETR [65] framework: (1) We develop GGNN [42] on the Deformable-DETR detection framework. Def+GGNN simply learns the rela-



(a) *"get lemon out of tea"* GT(∅),Ours(∅)

(b) *"open bottle of beer"* TOIST(∅),

Rationale for (b): sharp blade with a pointed end to insert into the bottle cap.

(c) *"pound carpet"*

(d) *"get lemon out of tea"*

Rationale for (d): handle of the spatula is long enough to reach the bottom of the cup; flat edge is designed to scoop up the lemon pieces.

Figure 3. **Visualization for prediction results** of our CoTDet, its variants, and the existing best-performing TOIST [19].

tions between objects and identifies objects based on their contexts, limiting its performance. (2) Besides, similar to TOIST [19], we initialize queries with the task's textual feature based on our framework. The performance gap (16.6% mAP_{box} and 14.1% mAP_{mask}) between Def+TOIST and our final model. (3) We introduce the visual affordance knowledge extracted by MLCoT but simply use it to initialize the queries of the decoder (Init w/MLCoT). The model achieves significant performance gain compared to the two baselines. (4) We further fuse knowledge with the image's visual feature map to construct a multi-modal feature map (Fuse w/MLCoT), which jointly understands the two modalities and improves performance (1.9% mAP_{box} and 1.3% mAP_{mask}) compared to the last model. (5) Our proposed knowledge-conditional query generation, generating based on the visual content of the image, the task, and the visual affordance knowledge, helps the decoder better localize the objects, resulting in average improvements of 4.7% mAP_{box} and 4.0% mAP_{mask}. (6) Finally, the knowledge-conditional denoising training improves AP_{box} and AP_{mask} by 1.6% and 1.9%, respectively.

4.4. Visualization

Figure 3 visualizes qualitative results for several examples. For (a), no objects in the image should be selected to *"get lemon out of tea"*. Our model can successfully return the empty set, while TOIST detects the french fry that is one of the salient objects in the image as the tool. Similarly, as knives are uncommon for *"opening bottle of beer"*, the knife in (b) is challenging for TOIST to identify and locate. Guided by the visual affordance of "sharp blade with a pointed end", our model correctly localizes and selects the sharp knife. The (c) and (d) demonstrate effectiveness without MLCoT or knowledge-conditional denoising training (KDN). With visual affordance knowledge obtained by directly asking LLMs, our model relies solely on matching with the single knowledge unit, which incorrectly detects the trunk in (c) and misses the knife in (d). The former trunk is easily confused with objects that are "flat, broad with a handle", while the latter knife is ignored because its visual attributes of straight mismatch the single knowledge unit that includes "curved or angled". Furthermore, without KDN, our detector lacks explicit guidance, leading to inaccurate detection in challenging scenes. Specifically, the glove in (c) and the knife in (d) are not detected successfully, and the packing line in (d) is mistakenly detected.

5. Conclusion

In this paper, we focus on challenging task driven object detection, which is practical in the real world yet underexplored. To bridge the gap between abstract task requirements and objects in the image, we propose to explicitly extract visual affordance knowledge for the task and detect objects having consistent visual attributes to the visual knowledge. Furthermore, our CoTDet utilizes visual affordance knowledge to condition the decoder in localizing and recognizing suitable objects.

Limitations: While acknowledging the disparity between the COCO-Task dataset and real-world application scenarios, attributed to its limited task variety and preference for images and annotations, our approach has the potential to extend beyond these confines. Notably, our knowledge acquisition and utilization are flexible and generalizable, granting it the capacity to transcend specific dataset, specific tasks, object categories, or tools. We leave this to future works. Furthermore, with the incorporation of LLM, our approach inherits potential social biases from LLM, which could potentially be reflected in the preference for selecting frequently used tools.

Acknowledgment: This work was supported by the National Natural Science Foundation of China (No.62206174), Shanghai Pujiang Program (No.21PJ1410900), Shanghai Frontiers Science Center of Human-centered Artificial Intelligence (ShangHAI), MoE Key Laboratory of Intelligent Perception and Human-Machine Collaboration (ShanghaiTech University), and Shanghai Engineering Research Center of Intelligent Vision and Imaging.

References

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022. 1

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 3, 6

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020. 1, 2, 4, 5, 6

[4] Dian Chen, Brady Zhou, Vladlen Koltun, and Philipp Krähenbühl. Learning by cheating. In *Conference on Robot Learning*, pages 66–75. PMLR, 2020. 1

[5] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 2, 5, 6

[6] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779, 2021. 2, 4

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[8] Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020. 3

[9] Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, and Prem Natarajan. Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5067–5077, 2022. 2, 3

[10] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 2, 4

[11] Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. Scene graph generation with external knowledge and image reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1969–1978, 2019. 3

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1, 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 6

[14] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 108–124. Springer, 2016. 2

[15] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 2

[16] Keizo Kato, Yin Li, and Abhinav Gupta. Compositional learning for human object interaction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 234–251, 2018. 3

[17] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 4

[18] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13619–13627, 2022. 2, 5, 6

[19] Pengfei Li, Beiwen Tian, Yongliang Shi, Xiaoxue Chen, Hao Zhao, Guyue Zhou, and Ya-Qin Zhang. Toist: Task oriented instance segmentation transformer with noun-pronoun distillation. *arXiv preprint arXiv:2210.10775*, 2022. 1, 2, 3, 6, 7, 8

[20] Liang Lin, Pengxiang Yan, Xiaoqian Xu, Sibei Yang, Kun Zeng, and Guanbin Li. Structured attention network for referring image segmentation. *IEEE Transactions on Multimedia*, 24:1922–1932, 2021. 2

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[22] Hugo Liu and Push Singh. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226, 2004. 3

[23] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 4, 5, 6

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[25] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *International Conference on Learning Representations (ICLR)*, 2023. 3

[26] Chengzhi Mao, Revant Teotia, Amrutha Sundar, Sachit Menon, Junfeng Yang, Xin Wang, and Carl Vondrick. Doubly right object recognition: A why prompt for visual rationales. *arXiv preprint arXiv:2212.06202*, 2022. 3

[27] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 4

[28] Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14111–14121, 2021. 3

[29] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 3

[30] Medhini Narasimhan, Svetlana Lazebnik, and Alexander Schwing. Out of the box: Reasoning with graph convolution nets for factual visual question answering. *Advances in neural information processing systems*, 31, 2018. 2, 3

[31] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. 1, 7

[32] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019. 1, 3

[33] Bryan A Plummer, Arun Mallya, Christopher M Cervantes, Julia Hockenmaier, and Svetlana Lazebnik. Phrase localization and visual relationship detection with comprehensive image-language cues. In *Proceedings of the IEEE international conference on computer vision*, pages 1928–1937, 2017. 4

[34] Sarah Pratt, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. *arXiv preprint arXiv:2209.03320*, 2022. 3

[35] Mengshi Qi, Yunhong Wang, Jie Qin, and Annan Li. Kegan: Knowledge embedded generative adversarial networks for semi-supervised scene parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5237–5246, 2019. 3

[36] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 3

[37] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1, 2

[38] Allen Z Ren, Bharat Govil, Tsung-Yen Yang, Karthik R Narasimhan, and Anirudha Majumdar. Leveraging language for accelerated learning of tool manipulation. In *Conference on Robot Learning*, pages 1531–1541. PMLR, 2023. 1

[39] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 6

[40] Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*, 2020. 1, 3

[41] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 817–834. Springer, 2016. 2

[42] Johann Sawatzky, Yaser Souri, Christian Grund, and Jurgen Gall. What object should i use?-task driven object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7605–7614, 2019. 1, 3, 6, 7

[43] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008. 3

[44] Sheng Shen, Chunyuan Li, Xiaowei Hu, Yujia Xie, Jianwei Yang, Pengchuan Zhang, Anna Rohrbach, Zhe Gan, Lijuan Wang, Lu Yuan, et al. K-lite: Learning transferable visual models with external knowledge. *arXiv preprint arXiv:2204.09222*, 2022. 2, 3

[45] Cheng Shi and Sibei Yang. Spatial and visual perspective-taking via view rotation and relation reasoning for embodied reference understanding. In *European Conference on Computer Vision*, pages 201–218. Springer, 2022. 2

[46] Krishna Kumar Singh, Santosh Divvala, Ali Farhadi, and Yong Jae Lee. Dock: Detecting objects by transferring common-sense knowledge. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 492–508, 2018. 2, 3

[47] Jiajin Tang, Ge Zheng, Cheng Shi, and Sibei Yang. Contrastive grouping with transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23570–23580, 2023. 2

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[49] Duc Minh Vo, Hong Chen, Akihiro Sugimoto, and Hideki Nakayama. Noc-rek: Novel object captioning with retrieved vocabulary from external knowledge. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17979–17987. IEEE, 2022. 3

[50] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427, 2017. 3

[51] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-

driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022. 2

[52] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022. 1, 3, 6

[53] Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4622–4630, 2016. 2, 3

[54] Sibei Yang, Guanbin Li, and Yizhou Yu. Cross-modal relationship inference for grounding referring expressions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4145–4154, 2019. 2

[55] Sibei Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4644–4653, 2019. 2

[56] Sibei Yang, Guanbin Li, and Yizhou Yu. Propagating over phrase relations for one-stage visual grounding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pages 589–605. Springer, 2020. 2

[57] Sibei Yang, Guanbin Li, and Yizhou Yu. Relationship-embedded representation learning for grounding referring expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8):2765–2779, 2020. 2

[58] Sibei Yang, Meng Xia, Guanbin Li, Hong-Yu Zhou, and Yizhou Yu. Bottom-up shift and reasoning for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11266–11275, 2021. 2

[59] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089, 2022. 3

[60] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022. 2

[61] Dongran Yu, Bo Yang, Qianhao Wei, Anchen Li, and Shirui Pan. A probabilistic graphical model based on neural-symbolic reasoning for visual relationship detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10609–10618, 2022. 3

[62] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1307–1315, 2018. 2

[63] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Bridging knowledge graphs to generate scene graphs. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 606–623. Springer, 2020. 3

[64] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. 3

[65] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1, 2, 4, 5, 7