

Afford-X: Generalizable and Slim Affordance Reasoning for Task-oriented Manipulation

Xiaomeng Zhu^{*} , Yuyang Li^{*} , Leiyao Cui , Pengfei Li ,
Huan-ang Gao , Yixin Zhu , and Hao Zhao 

Abstract—Object affordance reasoning, the ability to infer object functionalities based on physical properties, is fundamental for task-oriented planning and activities in both humans and Artificial Intelligence (AI). This capability, required for planning and executing daily activities in a task-oriented manner, relies on commonsense knowledge of object physics and functionalities, extending beyond simple object recognition. Current computational models for affordance reasoning from perception lack generalizability, limiting their applicability in novel scenarios. Meanwhile, comprehensive Large Language Models (LLMs) with emerging reasoning capabilities are challenging to deploy on local devices for task-oriented manipulations. Here, we introduce LVIS-Aff, a large-scale dataset comprising 1,496 tasks and 119k images, designed to enhance the generalizability of affordance reasoning from perception. Utilizing this dataset, we develop Afford-X, an end-to-end trainable affordance reasoning model that incorporates Verb Attention and Bi-Fusion modules to improve multi-modal understanding. This model achieves up to a 12.1% performance improvement over the best-reported results from non-LLM methods, while also demonstrating a 1.2% enhancement compared to our previous conference paper. Additionally, it maintains a compact 187M parameter size and infers nearly 50 times faster than the GPT-4V API. Our work demonstrates the potential for efficient, generalizable affordance reasoning models that can be deployed on local devices for task-oriented manipulations. We showcase Afford-X’s effectiveness in enabling task-oriented object grasping for robots across various tasks and environments, underscoring its efficiency and potential for manipulation tasks in physical environments.

Index Terms—Affordance reasoning, task-oriented manipulation, slim, generalizable.

I. INTRODUCTION

EFFECTIVE interaction with the world demands more than object recognition; it requires understanding how objects can be used. This concept, known as *affordance reasoning* [1], [2], transcends the traditional “what is where” paradigm [3] of object detection and classification systems [4]–[8]. Through affordance reasoning, agents infer potential functions from physical properties—a fundamental capability that enables both task-oriented manipulation and adaptive problem-solving in complex environments [9], [10].

This reasoning capability enables humans to naturally select appropriate tools for specific tasks [11] and devise creative solutions in unfamiliar or resource-constrained environments [12]. Consider, for example, repurposing a hollowed-out bell pepper as a water container when conventional containers are unavailable, as shown in Fig. 1(c). Such adaptability, rooted in understanding object properties and their potential functions, exemplifies the flexibility required for effective interaction across diverse environments [13].

The importance of affordance reasoning extends beyond human cognition into AI and robotics [2], [14]. In task-oriented manipulation [15]–[19], agents must process both task requirements (textual input) and environmental perception (visual input) to select and use appropriate objects for specific goals. This process requires reasoning about feature cues and matching them to novel task contexts [20], as illustrated in Figs. 2a and 2b. Developing robust affordance reasoning capabilities could significantly enhance AI systems’ ability to operate flexibly in complex, real-world environments [21].

Despite the significant benefits of affordance reasoning, its computational implementation faces several key challenges, particularly in designing frameworks suitable for local deployment and offline processing. Robotic platforms typically operate under strict computational constraints—whether using NVIDIA’s Jetson Orin development board or even high-end RTX 4090 GPUs with 24GB memory [22]. These limitations prevent the deployment of large-scale pre-trained generative Multimodal Large Language Models (MLLMs) [23] due to computing capabilities, power constraints, usage policies, and information security concerns [24], [25]. While smaller pre-trained generative MLLMs like SPHINX 1.1B [26] can operate locally, their limited knowledge bases and reasoning capabilities prove insufficient for complex affordance reasoning.

Manuscript received February 23, 2025; revised June 16, 2026; accepted August 3, 2026. This work is supported in part by the Brain Science and Brain-like Intelligence Technology—National Science and Technology Major Project (2025ZD0219400), the National Natural Science Foundation of China (62376009), the Beijing Nova program, the NVIDIA Academic Grant Program using Spark and Thor, the State Key Lab of General AI at Peking University, the PKU-BingJi Joint Laboratory for Artificial Intelligence, the Wuhan Major Scientific and Technological Special Program (2025060902020304), the Hubei Embodied Intelligence Foundation Model Research and Development Program, and the National Comprehensive Experimental Base for Governance of Intelligent Society, Wuhan East Lake High-Tech Development Zone. (Corresponding authors: Yixin Zhu; Hao Zhao.)

Xiaomeng Zhu, Yuyang Li, Leiyao Cui, and Yixin Zhu are with Peking University, Beijing 100871, China (email: xiaomeng.zhu@connect.ust.hk, y.li@stu.pku.edu.cn, cuileiyao24@mails.ucas.ac.cn, yixin.zhu@pku.edu.cn).

Xiaomeng Zhu is also with the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong 999077, China.

Leiyao Cui is also with Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China.

Pengfei Li and Hao Zhao are with the Institute for AI Industry Research, Tsinghua University, Beijing 100084, China (li-pf22@mails.tsinghua.edu.cn, zhaohao@air.tsinghua.edu.cn).

Huan-ang Gao is with the Department of Computer Science, Tsinghua University, Beijing 100084, China (gha24@mails.tsinghua.edu.cn).

Yixin Zhu is also with Beijing Key Laboratory of Brain-Computer Interface and Mental Health Modulation, and XS Vision.

Implementable code is available at: <https://zhuxmm.github.io/Afford-X>

^{*} Equal contribution.



Fig. 1: **Affordance reasoning for task-oriented manipulation.** Afford-X provides efficient visual affordance reasoning through: (a) two comprehensive datasets—COCO-Aff (112k images, 1,144 tasks, 80 categories) and LVIS-Aff (119k images, 1,496 tasks, 1,064 categories); (b) real-time processing (2.38 Frames Per Second (FPS)) with a compact 187M-parameter architecture generating bounding boxes and object masks; (c) robust generalization demonstrated through task-specific object selection and multi-object identification at 0.7 confidence threshold; (d) integration with robotic systems for simulated task-oriented object grasping.

In contrast, pre-trained Multimodal Models (MMs) [4], [27]–[31] achieve superior performance with fewer parameters and faster response times, likely because they leverage knowledge directly from the image feature space, capturing fine-grained visual details essential for affordance reasoning [32].

Training these slim MMs presents additional challenges, particularly when pre-training techniques ignore the fundamental nature of affordance. Models trained directly on object detection datasets often develop biased understanding, over-emphasizing familiar nouns. This bias can lead to failures when task descriptions include prominently visible objects, as the model may misinterpret an affordance reasoning task as simple object detection. For example, given the task “*clean bottle with*,” a model might fixate on detecting “*bottle*” while missing the critical action “*clean*,” failing to understand the task’s true intent (see Figs. 2c and 2d). Moreover, some pre-trained MMs, such as Multimodal DETection TRansformer (MDETR) [4], rely on simple concatenation of visual and language features, potentially limiting their comprehension of multimodal data.

Developing slim models with accurate and generalized affordance understanding requires both a large-scale corpus and diverse knowledge representation [33]. This diversity depends on three critical dimensions: the range of tasks, the quantity of images, and the variety of target object categories. Deficiencies in any of these dimensions can limit a robot’s ability to make contextual decisions when encountering dynamic affordances, diverse layouts, and novel objects [2], [32], [34]. However, expanding this knowledge base presents unique challenges due to the complex many-to-many mapping between affordance tasks and target objects—unlike the straightforward one-to-one

relationships in object detection [35], [36]. A single object can serve multiple tasks, and conversely, one task might be accomplished using various objects (see Fig. 3) [37]. Despite previous efforts involving manual annotations [37]–[39] or LLMs, creating truly diverse affordance reasoning datasets remains a significant challenge [17].

To address these challenges, we propose Afford-X, a slim end-to-end multimodal reasoning framework built on a knowledge distillation architecture inspired by TOIST [15]. Our framework consists of paired teacher and student models sharing the same architecture. It processes text-based task descriptions and visual scene inputs to autonomously identify appropriate objects for given tasks, producing both detection bounding boxes and fine-grained segmentation masks (see Fig. 1(b)). The training process occurs in two stages: first, the teacher model learns in an oracle manner using target object category labels in the text input; then, this knowledge transfers to the student model, which operates without such labels. This distillation approach proves particularly effective for scenarios with size constraints, as the student model’s architecture can adapt to practical requirements while maintaining effective supervision from the teacher.

To enhance our framework’s capabilities, we introduce two key modules. The Verb Attention (VA) module emphasizes action words in the input, ensuring accurate task understanding rather than mere object recognition. The Bi-Fusion (BF) module, inspired by BLIP [40], improves upon simple feature concatenation methods to better capture complex interactions between visual and textual information. Our experimental results (see Fig. 11) demonstrate that Afford-X, despite its compact 187M parameters, outperforms even GPT-4V in affordance

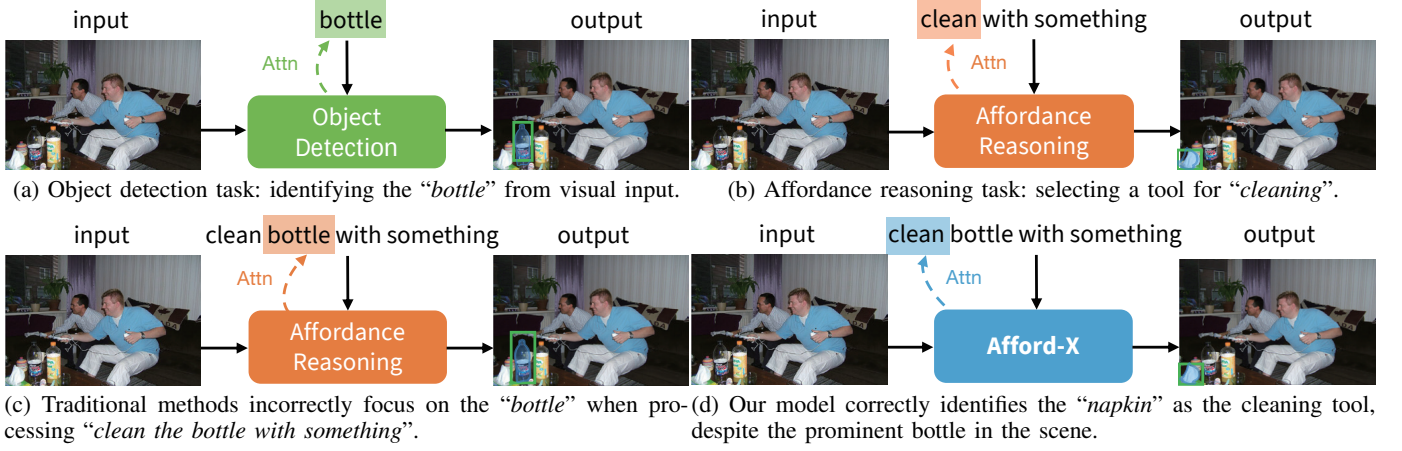


Fig. 2: Comparison between object detection and affordance reasoning. In our task-oriented setting, affordance reasoning aims to identify the object that best supports the intended action, rather than simply localizing the noun mentioned in the instruction. (a) Conventional object detection localizes the object specified by a noun query, *e.g.*, “bottle”. (b) Affordance reasoning selects an object that can serve a functional role for a task, *e.g.*, choosing a tool for “clean with something”. (c) Existing affordance methods may over-attend to the explicit noun in the instruction and incorrectly select the prominent “bottle” for “clean bottle with something”. (d) Afford-X emphasizes the task verb and correctly identifies the “napkin” as the appropriate cleaning tool, despite the visually salient bottle in the scene.

reasoning tasks while achieving real-time inference speeds of 2.38 FPS—making it practical for real-world deployment.

To strengthen Afford-X’s generalization capabilities, we developed an automated pipeline that leverages LLMs to convert object detection datasets into affordance reasoning datasets. This pipeline employs GPT-4 in dual roles: as a producer generating task-object pairs from object categories, and as a quality inspector filtering out errors and inconsistencies. Using this approach, we created two comprehensive datasets: COCO-Aff from COCO2014 [35] and LVIS-Aff from LVIS [36]. COCO-Aff features 1,144 diverse tasks, 112k training images, and 80 object categories, while LVIS-Aff expands to 1,496 tasks, 119k images, and 1,064 object categories, offering broader coverage of both indoor and outdoor scenarios (see Fig. 1(a)). Models trained on these datasets show significant improvements in generalization, with accuracy gains of 22.9% and 24.7% respectively on unseen tasks.

We validate these improvements through extensive testing of Afford-X’s ability to support embodied agents in diverse simulated environments. Beyond evaluations on natural images from datasets like COCO [35], we utilize textured meshes from Objaverse [41] and OmniGibson [42] to create scenes with diverse, randomly placed objects rendered using photorealistic ray-tracing, as shown in Fig. 1(d). This approach enables assessment of our model’s robustness to complex object geometry, appearance variations, and challenging environmental conditions including clustered objects, varying lighting, and visual distractions. We further demonstrate practical applicability through simulated object collection tasks, where Afford-X integrates with standard grasp planners and motion planners on a mobile manipulator to perceive scenes, select appropriate objects, and execute retrieval actions.

This article significantly extends our previous TOIST work [15] through several key contributions:

- Development of knowledge distillation-based Afford-X with innovative VA and BF modules, enhancing action recognition and multimodal interpretation capabilities

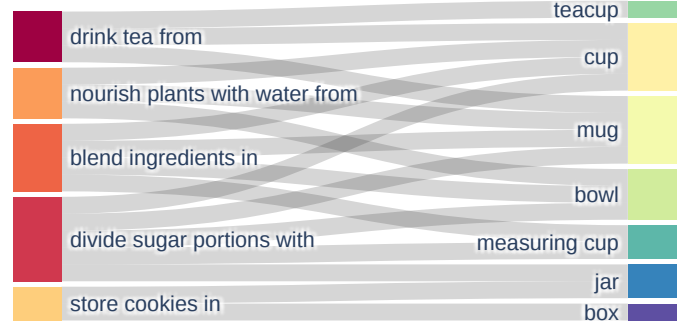


Fig. 3: Task-object mapping in affordance reasoning. Affordance reasoning involves complex many-to-many relationships between tasks and objects. (a) We visualize task descriptions from LVIS-Aff on the left and their compatible object categories on the right. (b) The connections demonstrate how individual tasks can be accomplished with multiple objects. (c) The mapping reveals how single objects can serve multiple different tasks, highlighting the complexity of affordance relationships.

- Creation of comprehensive affordance reasoning datasets: COCO-Aff and its expanded version LVIS-Aff
- Extensive validation demonstrating significant performance improvements over our previous approach through both image datasets and simulated environments

The paper continues with a comprehensive review of related work (Sec. II), followed by detailed descriptions of our dataset construction (Sec. IV), model architecture (Sec. III), and embodied affordance reasoning approach (Sec. V). We present experimental results (Sec. VI), discuss task-oriented manipulation findings (Sec. VI-F), and conclude with key insights (Sec. VII).

II. RELATED WORK

In this section, we review three key aspects of affordance research: learning-based affordance reasoning (Sec. II-A), datasets for training and evaluation (Sec. II-B), and task-oriented manipulation approaches (Sec. II-C).

A. Learning-based Affordance Reasoning

The concept of affordance, introduced by American psychologist James J. Gibson [1], proposes that environmental objects inherently offer action possibilities—for instance, a chair affords sitting. This fundamental notion, which emphasizes the relationship between physical properties and potential functions, has become central to how robots understand and interact with their environment [2].

As affordance research entered the computer vision domain, early approaches focused on establishing direct mappings between visual features and potential actions, bypassing explicit object recognition [43]. These traditional methods combined handcrafted features—including shape, size, texture, color, and material—with Bayesian networks [44] or support vector machines [45] to encode relationships between geometric features and affordances [46]–[48]. However, these approaches struggled to generalize across diverse object appearances and environmental contexts [49].

The emergence of deep learning transformed affordance reasoning through its powerful feature extraction capabilities. These models leverage multi-layer networks and large-scale datasets to capture rich object features and learn affordance cues, achieving enhanced performance and robustness [50]. Their success in identifying functional regions—such as graspable or supportable areas—has significantly advanced robotic task-oriented manipulation [43], [50]–[52]. However, challenges persist in capturing contextual dependencies and complex semantic relationships, particularly in cluttered scenes with multiple interacting objects.

The integration of semantic information with visual cues marked the next major advance. Modern models leverage category-level knowledge for object function inference, enabling more effective task-oriented object detection [15], [53]. This enhanced contextual understanding provides deeper insights into object-environment interactions [4], [15], [54]–[56]. Transformer-based architectures, especially vision-language pre-training models [4], [27]–[31], have further advanced the field through sophisticated cross-modal alignment techniques. These approaches demonstrate superior flexibility and generalizability compared to traditional Convolutional Neural Network (CNN)-based methods, particularly in complex and dynamic scenarios.

Most recently, the rich commonsense knowledge embedded in LLMs has opened new possibilities for affordance reasoning. CoTDet demonstrates this potential through structured task decomposition [57], while AffordanceLLM enhances open-world inference by combining visual perception with LLM capabilities [32]. Along this line, recent works further exploit foundation models and 3D cues. OOAL learns open affordances from a single example with foundation models [58], and both 3D-AffordanceLLM [59] and GEAL [60] push open-vocabulary affordance reasoning into 3D worlds. However, deploying large-scale LLMs like GPT-4 presents significant challenges for robotic platforms that require local, offline inference. While small-scale LLMs offer an alternative, they struggle with limited knowledge bases and reasoning capabilities—particularly in sparse language spaces

compared to dense image feature spaces [61]. To address these limitations, we propose Afford-X, a MM-based end-to-end framework inspired by TOIST [15]. Our approach achieves efficient inference with compact parameters, enabling broader deployment across manipulation platforms.

B. Dataset for Affordance Reasoning

The proliferation of deep learning in affordance reasoning has highlighted datasets as a critical foundation for model development. A dataset’s effectiveness depends on three key dimensions: the diversity of tasks, the quantity of images, and the range of object categories. Tab. I summarizes the major datasets in this field. The first significant milestone came from Myers *et al.* [62], who introduced both a framework for joint affordance localization and recognition and the field’s first pixel-level annotated dataset. While groundbreaking, this initial dataset focused primarily on surface features, overlooking the crucial role of human-object interactions in affordance reasoning. Chuang *et al.* [39] addressed this limitation with the ADE-Affordance dataset, built upon ADE20K [38], incorporating both physical constraints and social norms to better align with real-world reasoning challenges.

TABLE I: **Comparison of affordance detection datasets.** We analyze the key characteristics of major datasets in affordance reasoning. (a) We evaluate datasets based on their number of images (#Imgs), object categories (#Cats), and affordance/task categories (#Aff). (b) Each dataset builds upon different source datasets, providing varying foundations for affordance learning. (c) Our proposed datasets, COCO-Aff and LVIS-Aff, achieve significant expansion across all three dimensions compared to existing work.

Dataset	#Imgs	#Cats	#Aff	Source
ADE-Aff [39]	1000	150	7	ADE20K [38]
PAD [63]	4002	72	31	\
PADv2 [64]	30000	103	39	
PAD-L [64]	4002	72	31	
COCO-Tasks [37]	39724	49	14	COCO2024 [35]
RIO [17]	40214	69	>100	COCO2024
COCO-Aff	112k	80	1144	COCO2024
LVIS-Aff	119k	1064	1496	LVIS [36]

Recognizing that affordances fundamentally connect to human behavioral goals—reflecting Gibson’s concept of animal-environment complementarity—researchers began developing goal-oriented datasets. Luo *et al.* pioneered this direction with PAD [63] and its successor PADv2 [65], explicitly modeling the relationship between human goals and affordances while encompassing more complex scenarios. Lu *et al.* [64] further advanced this approach through PAD-L, which integrated natural language in affordance detection to enable object segmentation based on phrase-based affordance descriptions. However, their reliance on a limited affordance dictionary for paraphrasing constrained the capture of natural language complexity.

The field then shifted toward task-specific object selection. Sawatzky *et al.* [53] introduced the COCO-Task dataset, derived from COCO [66], marking the first transformation of an object detection dataset into an affordance reasoning dataset. While innovative, its scope remained limited by predefined 14 tasks and rigid phrase representations. Qu *et al.* [17]

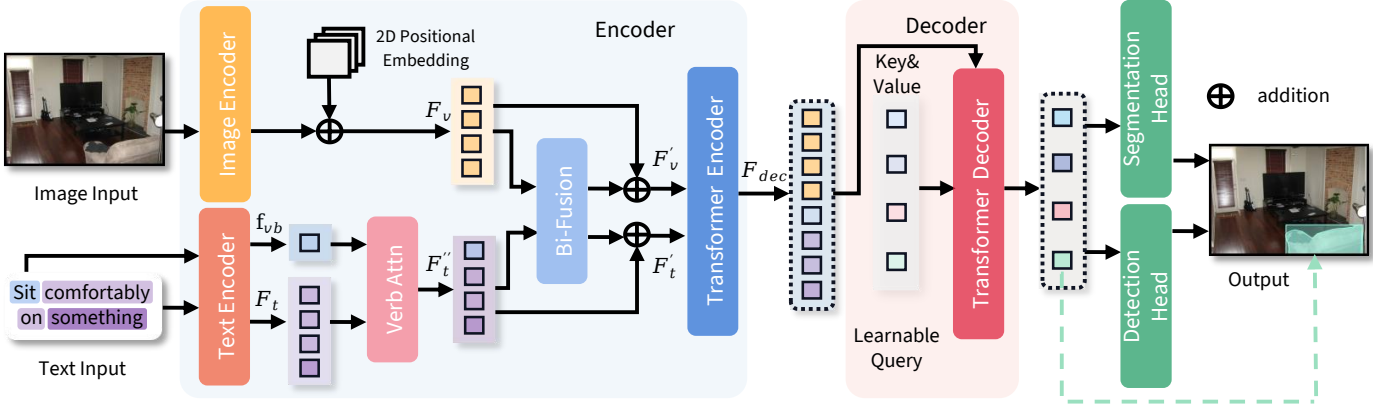


Fig. 4: **Architecture of the Afford-X.** Our framework processes visual and textual inputs through multiple specialized components. (a) The model begins by extracting visual features from images and textual features from prompts containing words like “something”. (b) The **Bi-Fusion** module performs bi-directional attention between visual and textual features to enhance multimodal understanding. (c) The **Verb Attn** module strengthens the model’s focus on action verbs to reduce category interference. (d) A Transformer encoder-decoder processes these enhanced features using learnable query vectors, producing parallel outputs for object detection and instance segmentation.

addressed these constraints with the RIO dataset, also built upon COCO2014, offering richer tasks and more diverse descriptions for broader scenario coverage. However, RIO’s limited object categories prove insufficient for dynamic open-world environments, and despite leveraging LLM for task construction, it still requires extensive human annotation for task-object pair filtering.

To overcome these limitations, we propose an automated pipeline for converting object detection datasets into affordance reasoning datasets. Our approach employs LLM in dual roles—as both producer and inspector—significantly reducing the need for human annotation. Through this pipeline, we have created COCO-Aff and LVIS-Aff, establishing new benchmarks in task coverage, image quantity, and object category diversity. These datasets provide models with a substantially richer knowledge base while maintaining data quality through automated verification, enabling more robust and generalizable affordance reasoning.

C. Task-oriented Manipulation

Ikeuchi and Hebert’s seminal work [67] established that vision systems should adapt their architectures to specific tasks rather than pursuing a general-purpose approach. This task-oriented vision paradigm has become fundamental to robotic manipulation, enabling systems to handle diverse tasks posed by algorithms or human users. These tasks span a broad spectrum—from object manipulation with varying grasp types [67]–[69] and purposes [2], [13], [19], [70] to complex environmental interactions such as door opening and water pouring [71]–[73]. At its core, task-oriented manipulation requires optimal system configuration through the selection of appropriate sensor signals [74], task representations [13], [68], [69], processing modules, and manipulation policies [18], [19], [75], [76], conditioning on specific goals and target objects.

Current task-oriented robotic vision and manipulation systems, however, typically operate under a significant constraint: they assume the suitable object for a task has already been specified and represented with proper visual representations (e.g., 3D mesh or point cloud) [55], [77]–[79]. This limitation

reduces system autonomy by requiring human input for object selection. Our approach addresses this fundamental gap by integrating affordance reasoning capabilities within scene understanding [80], enabling robots to identify appropriate objects for the given task within open-ended scenes with multiple candidate objects and obtain their 2.5D representations, *i.e.* segmented colored point clouds, which are vital for many downstream manipulation tasks [55], [77]–[79]. This integration represents a crucial step toward autonomous task-oriented manipulation systems that can adapt to diverse environments and task requirements.

III. THE AFFORD-X

This section presents our Afford-X framework for affordance reasoning. We begin with a formal problem formulation (Sec. III-A), followed by the model architecture (Sec. III-B), noun-pronoun distillation strategy (Sec. III-C), and key architectural components (Sec. III-D).

A. Problem Formulation

Given an RGB image $X_v \in \mathbb{R}^{3 \times H_0 \times W_0}$ and a task description X_l (e.g., “sit comfortably on”), our goal is to detect and segment objects most suitable for the specified task. The model predicts bounding boxes $B_{\text{pred}} = \{b_1, \dots, b_{n_{\text{pred}}}\}$, instance segmentation masks $M_{\text{pred}} = \{m_1, \dots, m_{n_{\text{pred}}}\}$, and preference scores $S_{\text{pred}} = \{\hat{s}_1, \dots, \hat{s}_{n_{\text{pred}}}\} \in [0, 1]^{n_{\text{pred}}}$. Each bounding box $b_i \in [0, 1]^4$ contains normalized center coordinates and dimensions, while preference scores indicate object suitability for the task. We denote the complete set of predictions as $O_{\text{pred}} = \langle B_{\text{pred}}, M_{\text{pred}}, S_{\text{pred}} \rangle$. Formally, we seek a function f such that:

$$f(X_v, X_l) = \langle B_{\text{pred}}, M_{\text{pred}}, S_{\text{pred}} \rangle. \quad (1)$$

B. Affordance Reasoning Model Architecture

To identify suitable objects without explicit object category labels, we design the Afford-X with parallel visual and textual pathways, enhanced by specialized modules for multimodal understanding (illustrated in Fig. 4). A pre-trained visual

encoder processes the input image X_v to extract visual features F_v , while a text encoder processes the task description X_l to generate textual features F_t . Two key components enhance these representations: the **VA module** processes F_t to produce enhanced text features F'_t that emphasize task-specific actions, while the **BF module** integrates F'_t with F_v to generate fused features F'_v and F''_t that capture fine-grained associations between vision and language.

The fused features pass through a transformer encoder-decoder architecture, where the encoder captures global relationships through self-attention mechanisms, and the decoder employs learnable query vectors to generate refined outputs. These outputs are projected through parallel heads to produce bounding boxes B_{pred} and segmentation masks M_{pred} . The decoder also outputs logits $\mathbf{G}_{\text{pred}} = [\hat{\mathbf{g}}_1, \dots, \hat{\mathbf{g}}_{n_{\text{pred}}}] \in \mathbb{R}^{n_{\text{pred}} \times n_{\text{max}}}$ for computing preference scores S_{pred} . Here, n_{pred} is the number of decoder object queries, and n_{max} is the fixed padded text length that includes the special “no-object” token. For each predicted object i , the preference score $\hat{s}_i \in S_{\text{pred}}$ is computed as:

$$\hat{s}_i = 1 - \frac{\exp(\hat{g}_i^{n_{\text{max}}})}{\sum_{j=1}^{n_{\text{max}}} \exp(\hat{g}_j^i)}, \quad (2)$$

where $\hat{g}_j^i$ is the logit of object query i at token position j and $\hat{g}_{n_{\text{max}}}^i$ is the “no-object” logit. Thus $\hat{s}_i = 1 - P_i(\text{no-object})$ measures how likely query i is a valid task-relevant target.

During training, we compute a bipartite matching between predicted and ground truth objects using the Hungarian algorithm [81]. The matched predictions are supervised with localization losses (L1 loss and Generalized Intersection over Union loss [82]) and segmentation losses (Dice/F-1 loss [83] and focal cross-entropy loss [84]). We incorporate the soft-token prediction loss and contrastive alignment loss from MDETR [4], adapting them to focus on the entire verb-pronoun description rather than individual tokens. The total loss for the Afford-X is formulated as:

$$\mathcal{L}_{\text{Afford-X}} = \lambda_1 \mathcal{L}_{\text{L1}} + \lambda_2 \mathcal{L}_{\text{GIoU}} + \lambda_3 \mathcal{L}_{\text{Dice}} + \lambda_4 \mathcal{L}_{\text{Cross}} + \lambda_5 \mathcal{L}_{\text{Token}} + \lambda_6 \mathcal{L}_{\text{Align}}, \quad (3)$$

where λ_1 to λ_6 are weights for the respective loss components.

C. Noun-Pronoun Distillation

To enable object inference without explicit category labels, we introduce a noun-pronoun distillation framework (illustrated in Fig. 5). Our approach uses a teacher model trained on task descriptions containing object labels (e.g., “sit comfortably with a couch”) and a student model that processes category-free descriptions (e.g., “sit comfortably with something”). Through clustering and preference distillation, the teacher transfers its object-centric knowledge to the student.

Clustering Distillation: We maintain a text feature memory bank that stores noun features from the teacher model, enabling the student to select appropriate noun prototypes for pronoun replacement. The process operates on enhanced text features F'_{noun} and F'_{pron} , corresponding to noun and pronoun tokens after VA module processing.

The memory bank is structured as a $n_{\text{task}} \times n_{\text{mem}} \times d$ tensor, where n_{task} represents the number of tasks, n_{mem} is the per-task

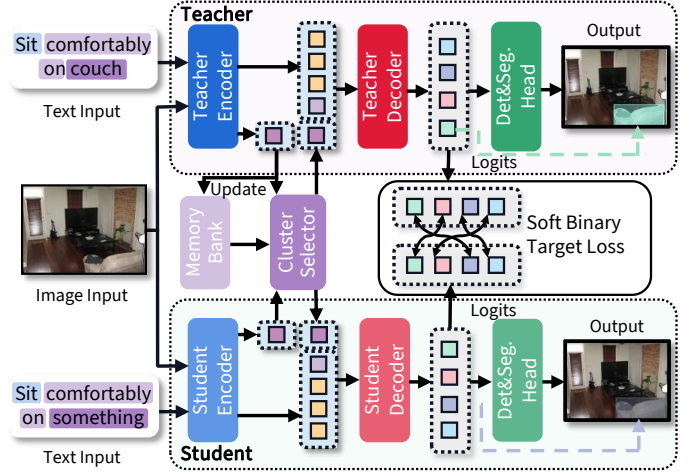


Fig. 5: **Noun-pronoun distillation framework.** Our framework employs parallel teacher-student encoder-decoder architectures for affordance learning. The teacher processes specific noun-based descriptions (e.g., “sit comfortably on couch”), while the student handles generalized pronoun-based inputs (e.g., “sit comfortably on something”). Knowledge transfer occurs through two mechanisms: (i) a memory bank storing noun features that guides the student’s cluster selector, and (ii) a soft binary target loss that aligns teacher-student logits. This design enables category-agnostic inference while maintaining category-informed understanding.

memory size, and d is the feature dimension. For each task j , we maintain a queue $\mathbf{F}_{\text{mem}}^j = [F_1^j, F_2^j, \dots, F_{n_{\text{mem}}}^j]$ of noun features. During training, we add teacher noun features F'_{noun} and evict their nearest neighbors, a redundancy-aware update that fixes the memory size while preserving feature diversity. We then apply K-means clustering to $\mathbf{F}_{\text{mem}}^j$ to obtain K cluster centers $\mathbf{F}_c^j = \{F_{c_1}^j, F_{c_2}^j, \dots, F_{c_K}^j\}$.

Knowledge transfer occurs through a cluster selector in the student model, which uses nearest neighbor classification to select a prototype $F_{c_s}^j$ from $\mathbf{F}_c^j$ based on the pronoun feature F'_{pron} . This prototype replaces F'_{pron} in the student’s feature sequence. To ensure proper alignment, we define the cluster loss $\mathcal{L}_{\text{cluster}} = \|F'_{\text{pron}} - F_{c_s}^j\|_2$, which minimizes the Euclidean distance between the pronoun feature and selected cluster center.

Preference Distillation: We align teacher and student predictions through a soft binary target loss based on the Kullback-Leibler (KL) divergence. For each object query, we compute binary probabilities indicating positive (ground truth object) or negative matches: $\mathbf{p} = [p^{\text{pos}}, p^{\text{neg}}] \in \mathbb{R}^{1 \times 2}$. These probabilities are defined using the softmax function:

$$p^{\text{pos}} = \frac{\sum_{j=1}^{n_{\text{max}}-1} \exp(\hat{g}_j)}{\sum_{j=1}^{n_{\text{max}}} \exp(\hat{g}_j)}, \quad p^{\text{neg}} = \frac{\exp(\hat{g}_{n_{\text{max}}})}{\sum_{j=1}^{n_{\text{max}}} \exp(\hat{g}_j)}, \quad (4)$$

The probability sequences for teacher and student models, denoted as $\mathbf{P}_t = [\mathbf{p}_{t_1}, \dots, \mathbf{p}_{t_{n_{\text{pred}}}}]$ and $\mathbf{P}_s = [\mathbf{p}_{s_1}, \dots, \mathbf{p}_{s_{n_{\text{pred}}}}]$, are aligned through bipartite matching. Using the Hungarian algorithm [81], we find an optimal permutation $\sigma \in \mathfrak{S}_{n_{\text{pred}}}$ that minimizes the matching cost:

$$\hat{\sigma} = \arg \min_{\sigma \in \mathfrak{S}_{n_{\text{pred}}}} \sum_{i=1}^{n_{\text{pred}}} \mathcal{L}_{\text{match}}(y_{t_i}, y_{s_{\sigma(i)}}), \quad (5)$$

where $y_{t_i} = (\hat{b}_{t_i}, \mathbf{p}_{t_i})$ combines the teacher’s bounding box

prediction $\hat{b}_{t_i}$ and probabilities, and $\mathcal{L}_{\text{match}}$ incorporates both box prediction losses and KL divergence.

The soft binary target loss is then defined using the optimal assignment $\hat{\sigma}$:

$$\mathcal{L}_{\text{binary}} = \sum_{i=1}^{n_{\text{pred}}} \mathcal{L}_{\text{KL}}(\mathbf{p}_{t_i}, \mathbf{p}_{s_{\hat{\sigma}(i)}}), \quad (6)$$

where the KL divergence between teacher and student probabilities is:

$$\mathcal{L}_{\text{KL}}(\mathbf{p}_{t_i}, \mathbf{p}_{s_{\hat{\sigma}(i)}}) = p_{t_i}^{\text{pos}} \log \left(\frac{p_{t_i}^{\text{pos}}}{p_{s_{\hat{\sigma}(i)}}^{\text{pos}}} \right) + p_{t_i}^{\text{neg}} \log \left(\frac{p_{t_i}^{\text{neg}}}{p_{s_{\hat{\sigma}(i)}}^{\text{neg}}} \right). \quad (7)$$

Minimizing $\mathcal{L}_{\text{binary}}$ aligns the student’s binary query probabilities with the teacher’s. Since the preference score $\hat{s}_i$ (Eq. (2)) follows a similar formulation to p^{pos} , this effectively transfers preference knowledge from teacher to student.

Overall Training Loss: We combine all components into a final training objective for the Afford-X with noun-pronoun distillation:

$$\mathcal{L}_{\text{Afford-X-NP}} = \mathcal{L}_{\text{Afford-X}}^t + \mathcal{L}_{\text{Afford-X}}^s + \lambda_7 \mathcal{L}_{\text{cluster}}^s + \lambda_8 \mathcal{L}_{\text{binary}}^s, \quad (8)$$

where $\mathcal{L}_{\text{Afford-X}}^t$ and $\mathcal{L}_{\text{Afford-X}}^s$ represent the teacher and student model losses, and λ_7, λ_8 weight the distillation components. The distillation losses $\mathcal{L}_{\text{cluster}}^s$ and $\mathcal{L}_{\text{binary}}^s$ apply only to the student model. During inference, we employ only the student model with the fixed memory bank, maintaining category-agnostic object identification.

We provide additional methodological details, including *noun feature* representations in Sec. A and *loss function* derivations in Sec. B. The effectiveness of our proposed modules is validated through extensive experiments detailed in Secs. H and VI-C. Furthermore, in Sec. VI-B, we demonstrate that the integration of these modules achieves state-of-the-art performance across multiple benchmarks.

D. Verb Attention module and Bi-Fusion module

To enhance multimodal understanding, we introduce two specialized components: the BF module for cross-modal feature integration and the VA module for action-focused reasoning.

BF module: We design this module to overcome limitations of MDETR’s feature concatenation approach by implementing bi-directional cross-modal attention between visual and textual features. This direct interaction captures fine-grained vision-language associations through parallel bi-directional attention:

$$\begin{aligned} F'_v &= F_v + \gamma_v \cdot \text{Attn}_{v \rightarrow t}(\text{LN}(F_v), \text{LN}(F_t)), \\ F'_t &= F_t + \gamma_t \cdot \text{Attn}_{t \rightarrow v}(\text{LN}(F_t), \text{LN}(F_v)), \end{aligned} \quad (9)$$

where $\text{Attn}_{v \rightarrow t}$ and $\text{Attn}_{t \rightarrow v}$ implement multi-head attention between modalities, LN performs layer normalization, and learnable parameters γ_v, γ_t control cross-modal influence.

VA module: This module enhances action-related information processing while reducing interference from dominant object categories. Leveraging the standardized verb-object-preposition format of task descriptions, it applies cross-attention between verb features and the complete textual prompt. Given the first verb’s feature $F_{vb} \in \mathbb{R}^{C_t}$ and full text features $F_t \in \mathbb{R}^{L \times C_t}$ from the text encoder (where L

is sequence length and C_t is feature dimension), the module computes:

$$F''_t = F_t + \text{CrossAttn}(\text{LN}(F_t), \text{LN}(F_{vb})), \quad (10)$$

where CrossAttn performs cross-attention to produce enhanced text features F''_t with dimensions matching F_t , effectively amplifying action-related information in the final representation.

E. Summary

The combination of noun-pronoun distillation, bi-directional feature fusion, and verb-focused attention enables Afford-X to effectively reason about object affordances without relying on explicit category labels. The teacher-student framework transfers object-centric knowledge while maintaining category-agnostic inference, the BF module ensures comprehensive multimodal understanding, and the VA module emphasizes action-specific features critical for affordance reasoning. Together, these components form a robust architecture that bridges the gap between category-specific training and category-agnostic deployment while maintaining high performance in affordance detection tasks.

IV. DATASET CONSTRUCTION

We present a scalable approach for creating large-scale affordance knowledge bases through automated conversion of object detection datasets. Our pipeline transforms standard detection annotations into rich affordance-task pairs, yielding two comprehensive datasets: COCO-Aff from COCO and LVIS-Aff from LVIS. We detail our conversion methodology in Sec. IV-A and analyze dataset characteristics in Sec. IV-B.

A. Dataset Collection

A primary challenge in developing Afford-X is establishing comprehensive affordance knowledge without depending on LLM inference. While manual dataset construction is possible, it becomes impractical due to the complex many-to-many relationships between tasks and objects: single objects can serve multiple purposes, tasks can utilize various objects, and objects have different levels of suitability for each task. This intricate mapping makes manual knowledge base construction both time-intensive and potentially inconsistent.

We address this challenge through an automated pipeline that leverages LLM capabilities for dataset construction. Our approach prioritizes three critical factors: task diversity (enabling broad affordance reasoning), image quantity (supporting physical property learning and scene generalization), and object category variety (covering diverse usage scenarios). While existing detection datasets provide rich visual and categorical resources, we needed a systematic method to generate diverse tasks and establish meaningful task-object relationships.

Our pipeline (illustrated in Fig. 6) employs GPT-4 for both task-object pair generation and quality inspection. After filtering uncommon object categories (e.g., animals, musical instruments, food items) that rarely serve as tools, we proceed through four systematic steps:

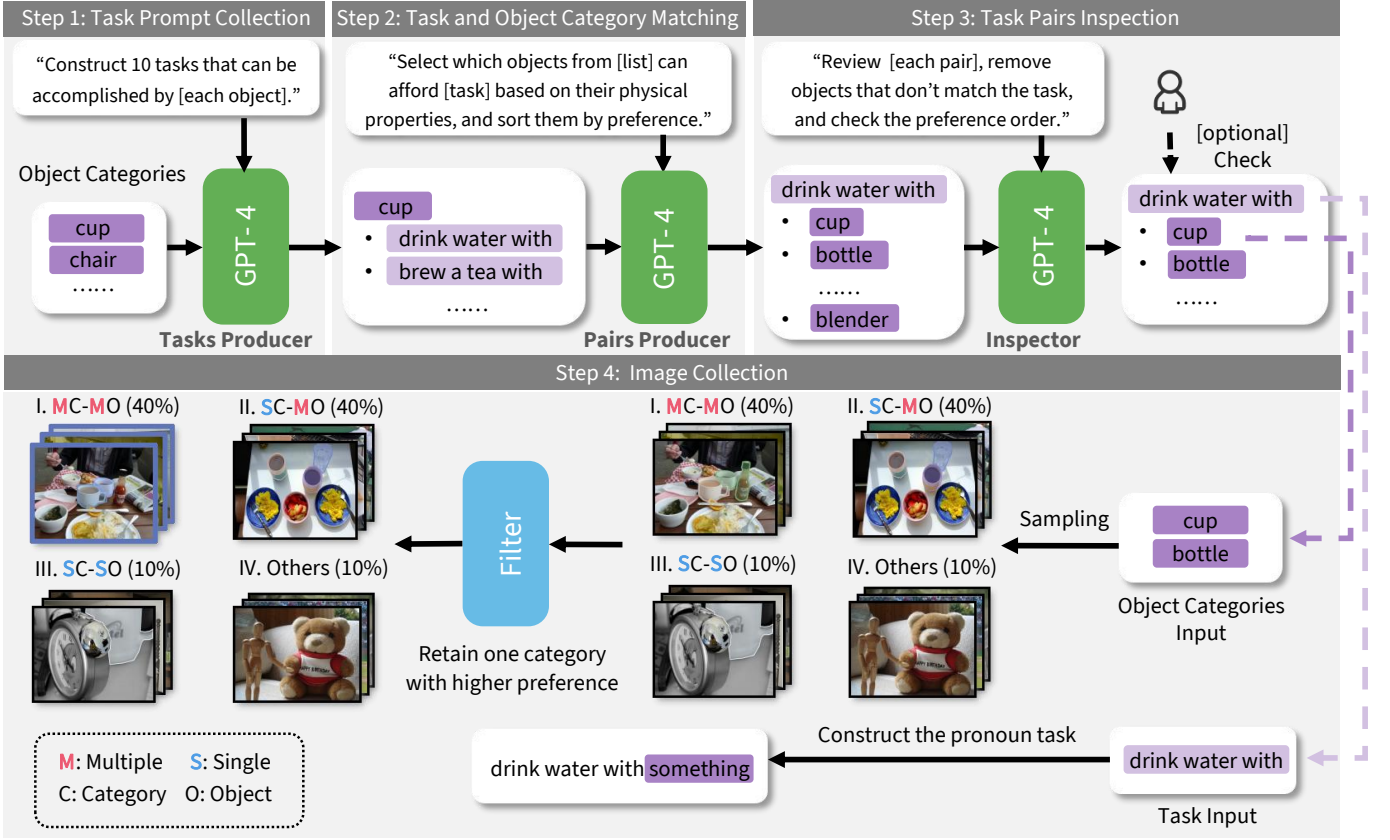


Fig. 6: **Pipeline for affordance dataset construction.** Our automated pipeline transforms detection datasets into affordance knowledge bases through four steps: (i) task generation—creating diverse tasks per object category, (ii) pair matching—associating tasks with suitable objects and establishing affordance rankings, (iii) quality inspection—validating task-object pairs and ranking accuracy, and (iv) image sampling—following composition rules (40% MC-MO, 40% SC-MO, 10% SC-SO, 10% no targets) while retaining highest-priority objects per task. This systematic approach ensures comprehensive coverage of affordance relationships while maintaining data quality.

Step 1: Task Prompt Collection. A GPT-4-based task producer generates 10 diverse tasks per object category, building an initial task pool. This step captures various potential uses for each object, ensuring comprehensive affordance coverage while maintaining natural and practical tasks.

Step 2: Task and Object Category Matching. A GPT-4-based pair producer matches tasks with relevant object categories, incorporating commonsense preference rankings. For instance, in “drink water with,” cups receive higher rankings than bottles, reflecting intuitive usage preferences. This ranking system captures nuanced distinctions in object suitability for specific tasks.

Step 3: Task Pairs Inspection. A GPT-4-based inspector performs multi-level quality control: filtering tasks against predefined criteria, verifying object-task match rankings, and removing inappropriate pairs (e.g., excluding blenders from “drink water with”). An optional manual review reduces task redundancy to optimize training efficiency, though this primarily serves computational rather than quality purposes.

Step 4: Image Collection. Following COCO-Tasks [37], we organize images into four configurations: MC-MO (multiple categories, multiple objects), SC-MO (single category, multiple objects), SC-SO (single category, single object), and Others (random images without target categories). We retain highest-priority objects per task and append “something” to prompts, balancing task specificity with visual diversity.

This automated pipeline enables efficient construction of large-scale affordance datasets while maintaining data quality. By leveraging LLM capabilities for generation and inspection, we create comprehensive knowledge bases that capture complex task-object relationships. Human validation and preference-aware target-selection examples are provided in Secs. P and Q.

B. Dataset Statistics

Our construction pipeline produced two comprehensive affordance datasets: COCO-Aff and LVIS-Aff. COCO-Aff, derived from COCO2014 [35], contains 112k images spanning 1130 task prompts and 80 object categories. We partitioned the dataset into training (600 images/task) and test (150 images/task) sets, sampling from COCO2014’s training and validation sets respectively for fair evaluation.

LVIS-Aff represents a significant expansion using LVIS [36], encompassing 119k images, 1494 task prompts, and 1064 object categories. We followed the same sampling strategy, drawing training (600 images/task) and test (150 images/task) sets from COCO2017’s training and validation sets respectively, maintaining strict separation for reliable evaluation.

Both datasets exhibit long-tail distributions in task-object relationships, reflecting natural variations in object functionality. Objects with limited functional diversity generated fewer

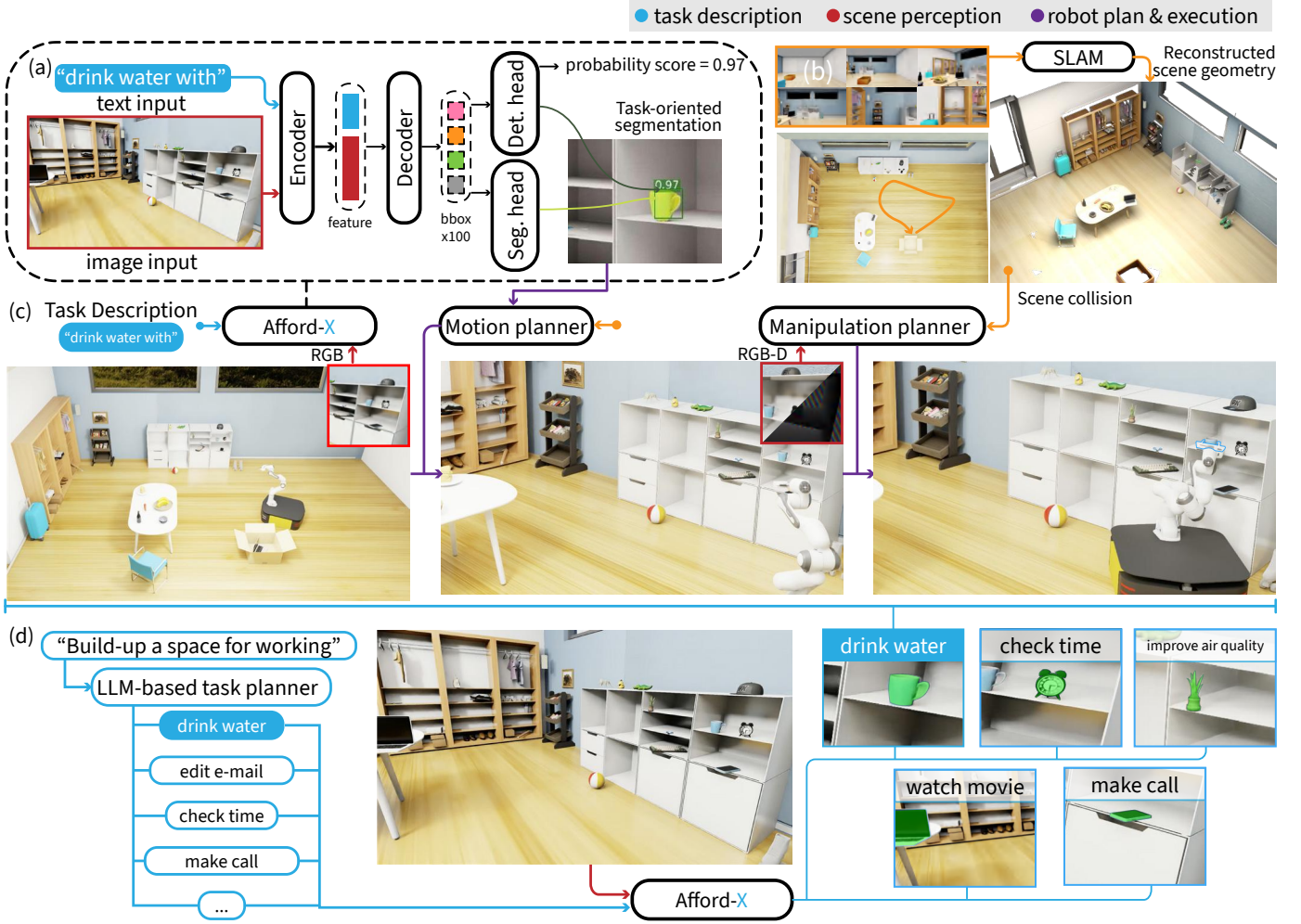


Fig. 7: **System infrastructure for object selection and grasping as a showcase of task-oriented manipulation.** Our robot executes the task “drink water with” through a multi-stage process: (a) Afford-X performs initial affordance reasoning on RGB input, generating instance segmentation masks for suitable objects. (b) Optionally, the robot navigates the scene to reconstruct a point cloud with SLAM for collision-free motion planning in completely unseen scenes. (c) The robot executes task-oriented object grasping by locating an object via Afford-X, approaching for detailed perception, and performing grasping. (d) For complex tasks (e.g., “build up a space for working”), the system decomposes the high-level goal into sequential sub-tasks, applying procedure (c) to each. See more examples in Sec. K.

unique tasks during Step 1 due to task overlap, leading to subsequent filtering. Despite this natural skew, the datasets provide comprehensive coverage of daily-life affordance scenarios, establishing robust knowledge bases for affordance reasoning. More details for the specifications and characteristics of both constructed datasets is presented in Sec. N

V. EMBODIED AFFORDANCE REASONING

As discussed in Sec. II-C, traditional task-oriented robotics often relies on human supervision or predefined rules for object selection prior to specific manipulation, limiting autonomy in open-world scenarios. Afford-X focuses on selecting the suitable task-relevant object in an open-ended scene and obtaining its visual representation for downstream manipulation modules. This step is especially important when diverse objects and contexts are present, where the target object cannot be assumed to be specified in advance.

To demonstrate this, we evaluate Afford-X’s practical applications by integrating it into embodied agents within simulated

environments. Our evaluation examines two key capabilities: (i) scene-level affordance reasoning, targeted on analyzing object affordance in complex 3D environments, and (ii) task-oriented object grasping, as a minimal yet representative showcase of manipulation based on the outputs of affordance reasoning. We leverage Isaac Sim’s photorealistic rendering to provide both high-fidelity rendering and precise ground-truth annotations, enabling systematic assessment of how our model bridges perception and action in embodied contexts.

A. Affordance Reasoning in the Scene

We developed a comprehensive evaluation within NVIDIA Isaac Sim to rigorously evaluate Afford-X’s affordance reasoning in 3D environments. This evaluation assesses the model’s understanding of object-task relationships across diverse environmental configurations and viewing conditions.

Our test environments employ a multi-source scene composition approach. We begin with base scenes from the Ev-ermotion dataset and OmniGibson [42], providing realistic

room layouts and furniture arrangements. We augment these with randomly placed textured meshes from Objaverse [41], following two principles: (i) including both task-suitable and unsuitable objects to test discrimination capability, and (ii) randomized object placement for complexity.

We evaluate Afford-X’s ability in selecting appropriate object for the desired task in the simulated 3D scenes, and compare it with LLM-integrated pipelines. The evaluation features an RGB-D camera for environmental perception. Isaac Sim’s physics engine ensures authentic environmental interactions, while its ray-tracing renderer generates photo-realistic images with lighting and materials. The simulator provides ground-truth annotations including object bounding boxes and segmentation masks for quantitative evaluation. During testing, Afford-X processes RGB images alongside task instructions, generating segmentation masks and for identifying task-relevant objects. This enables the robot to identify suitable objects for task execution. These procedures enables systematic evaluation across varying scenes, viewpoints, and task contexts while maintaining experimental reproducibility. The results are reported and analyzed in Sec. VI-E. Further towards the open-ended world, Sec. K reports a more in-depth evaluation on Afford-X’s capabilities in diverse 3D scenes, and Sec. L reports analysis on object size and occlusion.

B. Task-oriented Object Grasping

We use object grasping as a minimal yet representative task to evaluate whether the predicted object-level affordance mask can support downstream manipulation. Our system architecture (Fig. 7) combines Afford-X with established manipulation modules: GraspNet [87] as the grasp planner and cuRobo [88] as the collision-aware motion planner. The manipulation follows the four-phase procedure: (i) **Affordance reasoning**: Afford-X processes the scene’s RGB frame with a task description, identifying suitable target objects through segmentation masks. (ii) **Viewpoint optimization**: The robot positions itself at a prescribed viewing distance d_{view} from the identified target for detailed perception, where the target camera location is computed using the depth image additionally acquired in step (i); An affordance mask is inferred from each perspective and used to segment the RGB-D image into an object point cloud from the masked region and a scene point cloud from the remaining region. (iii) **Grasp planning**: The grasp planner [87] generates object-level grasp proposals (each consists of a gripper pose, a width, and a quality score) using the segmented object point cloud as the grasping target. (iv) **Grasp execution**: A motion planner generates trajectories for the grasp proposals using the scene point cloud as collider and executes the planned grasp motion with the highest grasp score on the robot. We operate in simulated scenes and objects from OmniGibson [42] (see Fig. A7). During grasp execution, we use a fixed joint to attach the object to the gripper when the gripper is closing if the object is close enough to the gripper, which is a standardized protocol to isolate affordance reasoning from unstable contact simulation.

For complex tasks requiring multiple steps (e.g., “build up a space for working”, Fig. 7(d)), we employ an LLM to

decompose the high-level goal into atomic sub-tasks. Each sub-task follows the same three-phase protocol sequentially, enabling structured completion of complex manipulations through affordance-guided action sequences. This integration of affordance reasoning with manipulation planning advances autonomous task-oriented robotics, enabling robots to independently identify and utilize appropriate objects based on task requirements. Sec. M provides more details and results.

VI. EXPERIMENT

To validate the effectiveness of Afford-X, we conduct comprehensive experiments across three key dimensions: (i) performance comparison and ablation studies (Secs. VI-B and VI-C), (ii) dataset analysis (Sec. VI-D), and (iii) real-world applicability (Secs. VI-E and VI-F). The latter includes comparative analysis against LLM-based approaches and validation in task-oriented manipulation scenarios.

A. Implementation Details

Model Architecture: We implement Afford-X using RoBERTa-base [89] for text encoding and ResNet-101 [90] as the CNN backbone. To leverage existing vision-language understanding capabilities, we initialize our model with pre-trained weights from MDETR [4]. We evaluate this architecture across three progressively more challenging datasets: COCO-Tasks, COCO-Aff, and LVIS-Aff.

Training Protocol: Our training follows a carefully designed multi-stage approach. In the initial verb-pronoun and verb-noun stages, we employ the Adam optimizer with a batch size of 36, applying a uniform learning rate of 10^{-5} across the text encoder, backbone network, and BF module. The subsequent distillation stage requires more precise parameter updates, leading to the reduced batch size 18 and text encoder learning rate 5×10^{-6} while maintaining other learning rates.

Data Augmentation: To enhance model robustness to real-world variations, we implement a comprehensive data augmentation pipeline. This includes dynamic image resizing that randomly scales the shortest side between 480-800 pixels while capping the longest side at 1333 pixels. We further augment the training data through random cropping (probability 0.5), generating diverse viewports between 384-1333 pixels.

Evaluation Metrics: We evaluate model performance using the AP@0.5 metric, which assesses localization accuracy and ranking effectiveness through predicted preference scores S_{pred} . For comprehensive evaluation across different tasks, we report the mean Average Precision (mAP@0.5), calculated by averaging AP@0.5 scores across all experimental tasks.

B. Performance Analysis

Our comprehensive experimental results are presented in Tab. II. We evaluate our approach against existing methods based on pre-trained MMs, excluding LLM-augmented baselines for fair comparison. Prior approaches fall into two categories: two-stage methods combining object detection with Gated Graph Neural Networks (GGNN) [53] (rows (a)-(g)), and one-stage methods built on MDETR (rows (h)-(k)).

TABLE II: **Comparison of Afford-X with state-of-the-art methods.** Comprehensive evaluation across COCO-Tasks, COCO-Aff, and LVIS-Aff datasets demonstrates the effectiveness of our approach, with Afford-X consistently achieving superior performance in both affordance understanding and instance segmentation tasks. The performance gains stem from our proposed VA and BF modules, which enhance the baseline architecture. Results marked with $\dagger$ are from original papers, with $\ddagger$ are from our previous conference work, and with **bold** and underlined values indicating best and second-best performance, respectively.

Index	Method	COCO-Tasks		COCO-Aff		LVIS-Aff	
		mAP^{box}	mAP^{mask}	mAP^{box}	mAP^{mask}	mAP^{box}	mAP^{mask}
(a)	Fast R-CNN [85] + GGNN $\dagger$	32.6	-	-	-	-	-
(b)	YOLO + GGNN [53] $\dagger$	33.2	-	-	-	-	-
(c)	MDETR (w/o pretraining) + GGNN $\ddagger$	9.6	8.6	-	-	-	-
(d)	MDETR + GGNN $\ddagger$	36.8	30.3	-	-	-	-
(e)	ViTDet (ViT-B) [86] + GGNN	22.5	29.5	27.4	22.8	6.6	5.9
(f)	ViTDet (ViT-L) + GGNN	32.1	24.6	29.7	24.7	8.0	7.1
(g)	ViTDet (ViT-H) + GGNN	33.8	25.9	31.5	26.1	8.5	7.4
(h)	MDETR [4]	41.3 $\ddagger$	35.2 $\ddagger$	44.7	41.0	25.1	22.7
(i)	MDETR (w/ VA & BF)	43.2	36.9	<u>45.2</u>	<u>41.4</u>	<u>26.8</u>	<u>24.2</u>
(j)	TOIST [15]	<u>44.1</u> $\ddagger$	<u>39.0</u> $\ddagger$	44.9	41.3	26.2	23.4
(k)	Afford-X (w/ VA & BF)	45.3	39.2	45.8	42.5	27.7	24.8

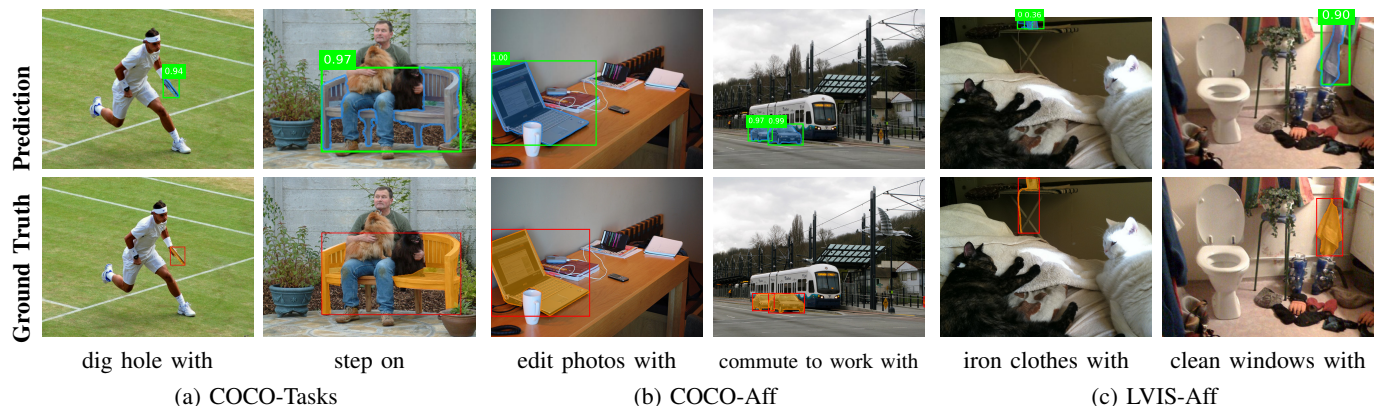


Fig. 8: **Qualitative results of Afford-X on three datasets.** The first row of each section shows model predictions, with corresponding ground truth in the second row. Results demonstrate performance across (a) COCO-Tasks, showing sports and office scenes; (b) COCO-Aff, depicting indoor and outdoor environments; and (c) LVIS-Aff, illustrating diverse interaction scenarios. Green boxes indicate model predictions while orange boxes show ground truth annotations.

In the two-stage category, we evaluated three object detection frameworks combined with GGNN: CNN-based Fast R-CNN [85], YOLO (version 11), and transformer-based approaches (MDETR, ViTDet [86]). Despite using COCO-pretrained weights, these methods showed limited performance. Even on the relatively simple COCO-Tasks dataset, Fast R-CNN and YOLO implementations achieved only 32-33% mAP^{box} . While ViTDet offered modest improvements, scaling from ViT-B to ViT-H backbone failed to overcome fundamental performance limitations.

The one-stage category, represented by MDETR-based approaches, demonstrated notably stronger results. The baseline MDETR implementation (row (h)) achieved 41.3% mAP^{box} and 35.2% mAP^{mask} on COCO-Tasks, suggesting that vision-language aligned features better capture object semantics across diverse scenarios. Importantly, this one-stage approach significantly outperformed its two-stage counterpart (row (d)), despite sharing the same MDETR detection backbone. This performance gap highlights the advantages of end-to-end training for affordance reasoning tasks.

Building on these insights, we enhanced the baseline

MDETR with our proposed VA and BF modules. This enhancement yielded substantial improvements across all datasets: COCO-Tasks performance increased to 43.2% mAP^{box} and 36.9% mAP^{mask} , with similar gains on COCO-Aff and LVIS-Aff. Our full Afford-X framework, which integrates these modules with TOIST, further pushes performance boundaries, achieving 45.3% mAP^{box} and 39.2% mAP^{mask} on COCO-Tasks. Notably, Afford-X maintains its leading position on the more challenging LVIS-Aff dataset, which features long-tail distributions and diverse affordance relationships, reaching 27.7% mAP^{box} and 24.8% mAP^{mask} .

Qualitative results in Fig. 8 reveal both strengths and limitations across all three datasets. While Afford-X successfully identifies major affordance regions, it sometimes struggles with fine-grained functional elements crucial for complete affordance understanding. For instance, in the “iron clothes with” example, the model misses the iron’s power cord, while the “clean windows with” case shows imprecise bounding box localization extending beyond the cleaning implement. A more detailed failure analysis is provided in Sec. G.

C. Ablation Study

We conduct comprehensive ablation studies to evaluate three critical components of Afford-X: (i) distillation architecture, (ii) VA and BF modules, and (iii) cluster number K. Using the TOIST framework as our baseline, we systematically analyze each component’s contribution through extensive experiments.

Knowledge Transfer through Distillation: To address categorical bias in the COCO-Tasks dataset, we introduce a novel two-stage Noun-Pronoun Distillation framework. Rather than directly minimizing the distance between pronoun-based ($l_{\text{pron}}^{\text{tr}}$) and noun-based ($l_{\text{noun}}^{\text{tr}}$) representations, our framework strategically transfers knowledge through an intermediate feature space ($l_{c_s}^j$). This is achieved by first training a teacher model with explicit noun categories and then distilling this knowledge into a baseline model operating on verb-pronoun instructions. As shown in Tab. III, this approach improves the verb-pronoun mAP^{box} from 41.3% to 44.1% compared to MDETR re-trained on COCO-Tasks with verb-pronoun inputs and without noun-pronoun distillation. Notably, this sophisticated distillation strategy outperforms direct distillation between $l_{\text{pron}}^{\text{tr}}$ and $l_{\text{noun}}^{\text{tr}}$, confirming the effectiveness of our architectural design.

TABLE III: **Different distillation methods on COCO-Tasks.** Values in parentheses show improvements over the baseline.

Method	mAP^{box}	mAP^{mask}
MDETR	41.3	35.2
distill from $l_{c_s}^j$ to $l_{\text{pron}}^{\text{tr}}$	44.1 (+2.8)	39.0 (+3.8)
distill from $l_{\text{noun}}^{\text{tr}}$ to $l_{\text{pron}}^{\text{tr}}$	41.9 (+0.6)	36.0 (+0.8)

To further refine feature grouping and preference scoring, we propose a Clustering Distillation (CD) approach comprising three complementary components: Cluster Center Replacement (CCR) for anchoring pronoun tokens, Cluster Loss (CL) to promote refined feature grouping, and Soft Binary Target Loss (SBTL) to improve preference modeling. As demonstrated in Tab. IV, while individual components provide modest improvements, their combination yields substantial gains of +2.8% mAP^{box} and +3.8% mAP^{mask} over the baseline. When integrated into the complete Afford-X framework, these enhancements achieve state-of-the-art performance of 45.3% mAP^{box} on verb-pronoun tasks.

TABLE IV: **Ablation study of clustering distillation components on COCO-Tasks.** Each row shows performance impact when specific components are enabled (✓) or disabled (×). CCR: Cluster Center Replacement, CL: Cluster Loss, SBTL: Soft Binary Target Loss.

Index	Method Components			Performance	
	CCR	CL	SBTL	mAP^{box}	mAP^{mask}
(a)	×	×	×	41.3	35.2
(b)	×	×	✓	43.4 (+2.1)	38.0 (+2.8)
(c)	×	✓	×	42.0 (+0.7)	37.1 (+1.9)
(d)	×	✓	✓	43.8 (+2.5)	38.6 (+3.4)
(e)	✓	×	×	42.0 (+0.7)	37.0 (+1.8)
(f)	✓	×	✓	42.3 (+1.0)	37.3 (+2.1)
(g)	✓	✓	×	42.3 (+1.0)	37.5 (+2.3)
(h)	✓	✓	✓	44.1 (+2.8)	39.0 (+3.8)

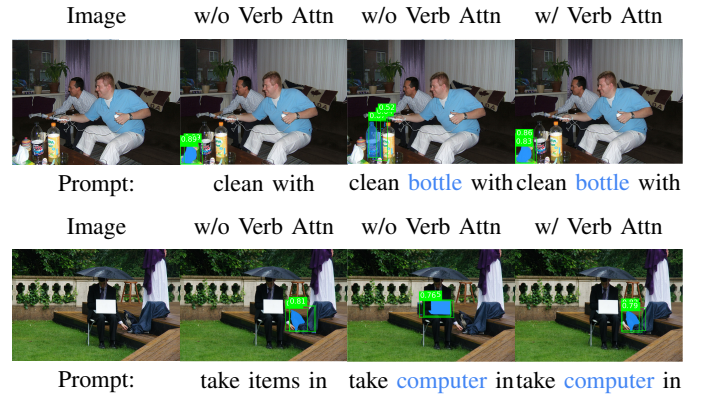


Fig. 9: **Impact of VA module on affordance reasoning performance.** The VA module reduces failures in affordance reasoning when text prompts include prominent objects. Column 1 shows the original images. Columns 2 and 3 present results without the module for prompts without and with object labels, respectively. Column 4 displays results with the module for prompts with object labels.

Analysis of VA and BF Modules: Tab. V presents ablation results for the VA and BF modules across three input settings: verb-pronoun, verb-noun, and distill. We evaluate their effectiveness by sequentially integrating them into our baseline.

TABLE V: **Ablation study of Afford-X components.** Analysis showing performance impact of VA and BF modules across verb-pronoun, verb-noun, and distillation settings. Values in parentheses indicate improvements over baseline configuration.

Index	Method		Performance		
	VA	BF	verb-pronoun	verb-noun	distill
(a)	×	×	41.3	53.2	44.1
(b)	✓	×	43.2 (+1.9)	53.8 (+0.6)	44.8 (+0.7)
(c)	×	✓	43.0 (+1.7)	53.9 (+0.7)	44.6 (+0.5)
(d)	✓	✓	43.4 (+2.1)	54.8 (+1.6)	45.3 (+1.2)

Individual module integration shows consistent improvements across all input configurations. The VA module (row (b)) yields gains of +1.9%, +0.6%, and +0.7% across the three input types, demonstrating its effectiveness in emphasizing action-related cues while reducing noun-based interference. Similarly, the BF module (row (c)) improves performance by +1.7% and +0.7% for pretrained student and teacher models respectively, with an additional +0.5% gain during distillation, confirming that bidirectional cross-modal attention enhances visual-language alignment for affordance reasoning. The combination of both modules (row (d)) demonstrates synergistic benefits, achieving +2.1% improvement for verb-pronoun input and +1.6% for verb-noun input, surpassing individual module gains. While the joint implementation shows a more modest +1.2% improvement under the distill setting, likely due to architectural constraints, the results validate the complementary nature of both modules in enhancing affordance reasoning.

Qualitative analysis in Fig. 9 illustrates the VA module’s impact on attention mechanisms. The baseline model correctly identifies target objects for tasks like “clean with” (towel) and “take item in” (backpack) but becomes susceptible to interference from distracting nouns (bottle and computer). With the VA module, the model maintains accurate affordance reasoning.

dance reasoning while effectively filtering such distractions, providing visual confirmation of our quantitative findings and demonstrating enhanced functional reasoning capabilities. Additional qualitative visualizations and free-form task-description evaluations are provided in Secs. C and E.

Analysis of Cluster Number K : We systematically investigate optimal cluster numbers (K) in the distillation framework; see also Fig. 10. We evaluate K values from 1 to 10, noting that higher values introduce excessive clustering complexity. Results show that all tested configurations improve upon the baseline model’s performance (41.3% mAP^{box} , 35.2% mAP^{mask}). $K = 3$ emerges as the optimal choice, suggesting that a moderate number of clusters effectively balances feature aggregation and information preservation during knowledge distillation from noun features to the student model. Specifically, a moderate K balances prototype diversity and stability, whereas a larger K over-fragments noun features and destabilizes prototypes for pronoun-based inference.

D. Dataset Analysis

To examine how large-scale visual and linguistic information improves affordance reasoning, we analyze two key factors: *task-scale* and *category-scale*.

Task-scale Analysis: We evaluate models trained on three datasets of increasing scale: COCO-Tasks (14 tasks), COCO-Aff (1,144 tasks), and LVIS-Aff (1,496 tasks). As shown in Tab. VI, while the COCO-Tasks-trained model achieves strong performance on seen tasks ($\text{mAP}^{\text{box}} = 45.3\%$), it fails to generalize to unseen tasks ($\text{mAP}^{\text{box}} = 1.6\%$). In contrast, training on COCO-Aff substantially improves generalization to unseen tasks ($\text{mAP}^{\text{box}} = 24.5\%$), with further gains achieved by LVIS-Aff ($\text{mAP}^{\text{box}} = 26.3\%$), demonstrating the benefits of expanded task and object coverage.

TABLE VI: **Task-scale analysis.** Comparison of model performance on 14 *seen tasks* from COCO-Tasks and 80 *unseen tasks* from LVIS.

Training Data.	Seen-Tasks (14)		Unseen-Tasks (80)	
	mAP^{box}	mAP^{mask}	mAP^{box}	mAP^{mask}
COCO-Tasks	45.3	39.2	1.6	1.3
COCO-Aff	43.9	38.3	24.5	24.1
LVIS-Aff	—	—	26.3	25.6

Category-scale Analysis: We further evaluate generalization using 40 unseen tasks: 20 with common categories (present in both COCO and LVIS) and 20 with novel categories (LVIS-exclusive). As shown in Tab. VII, the COCO-Aff-trained model performs well on common category tasks

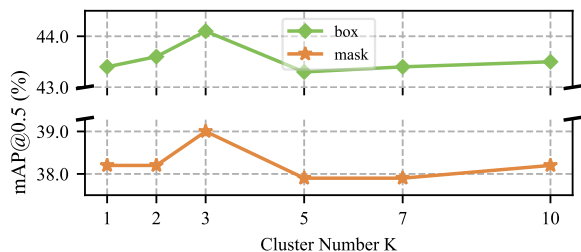


Fig. 10: **Effects of clustering number.** A moderate number of clusters balances prototype diversity and stability.

(35.0% mAP^{box}) but struggles with novel categories (4.9% mAP^{box}). Models trained on LVIS-Aff show improved performance across both task types (37.2% and 8.3% mAP^{box} respectively), demonstrating that expanded category coverage enhances both affordance reasoning and generalization.

TABLE VII: **Category-scale analysis.** Performance comparison between COCO-Aff and LVIS-Aff trained models on unseen tasks involving common categories (shared between COCO/LVIS) and novel categories (LVIS-exclusive).

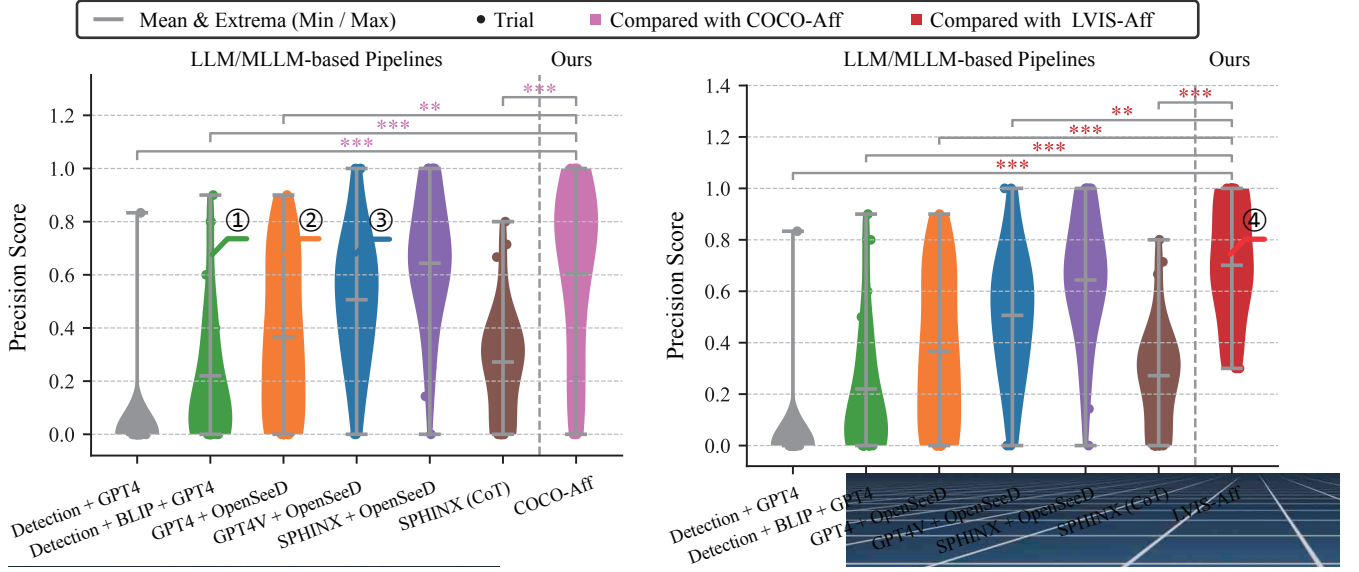
Training Data.	Common categories		Novel categories	
	mAP^{box}	mAP^{mask}	mAP^{box}	mAP^{mask}
COCO-Aff	35.0	33.9	4.9	4.8
LVIS-Aff	37.2	35.3	8.3	8.2

E. Affordance Reasoning in Simulated 3D Scenes

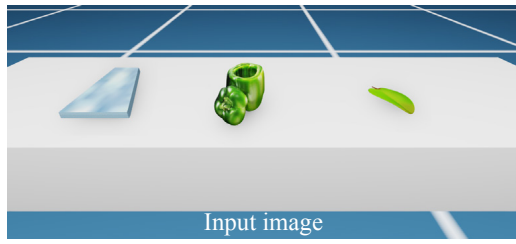
We evaluate Afford-X’s potential as an alternative to LLM/MLLM-based pipelines in robotic deployments through extensive testing in simulated 3D environments with textured object meshes, as described in Sec. V-A. Our evaluation compares baseline methods trained on COCO-Aff and LVIS-Aff against three categories of LLM/MLLM-based pipelines: (i) Detect before Reasoning (rows (a) and (b)), where object detection precedes LLM processing; (ii) Reason before Detection (rows (c)-(e)), where an LLM or MLLM identifies task-relevant candidate nouns or object categories before localization with OpenSeeD; and (iii) Simultaneous Perception and Reasoning (row (f)), where an MLLM jointly processes the task and image for direct bounding box output. Additional comparisons with MLLM-based pipelines are provided in Sec. D.

Our experimental framework encompasses 50 tasks (30 seen, 20 unseen) in simulated environments, with each task evaluated across 10 images containing three candidate objects—one capable of affording the task and two non-affording objects randomly selected from the candidate pool. Success criteria require bounding box predictions to achieve mIoU above 0.5, as shown in Fig. 11, where the “Detection” methods utilize RAM++ [91] and Grounding Dino [8] for object detection based on given labels. To ensure practical relevance, all evaluations were conducted on a single 24GB RTX 3090 GPU, comparing API-based services with memory-constrained ($\leq 24\text{GB}$) SPHINX tiny models while accounting for API communication latency in GPT-4 implementations. The FPS and parameter sizes for each baseline are detailed in Tab. VIII, where $>$ indicates API usage, excluding these model parameters from statistics. Notably, our model requires only 3.2 GB locally, far below SPHINX (10–11.5 GB), leaving ample headroom for concurrent robotic modules.

The experimental results revealed significant performance variations across pipeline architectures. LLM-based label reasoning pipelines demonstrated notably poor performance (2.33% recall, 3.44% mAP), primarily due to semantic ambiguity in object labeling—exemplified by cases where distinct tools like hammers and chisels share generic labels, compromising reasoning accuracy. While image captioning



(a) Precision comparison of different methods



Task: drink water with

- ① Textual ambiguity
- "glass, green pepper, leaf"
 - "glass"
- ② Missing Perception
- "water bottle, glass, mug, straw, water dispenser, sports sipper"
 - Detection failed
- ③ "cup"
- ④

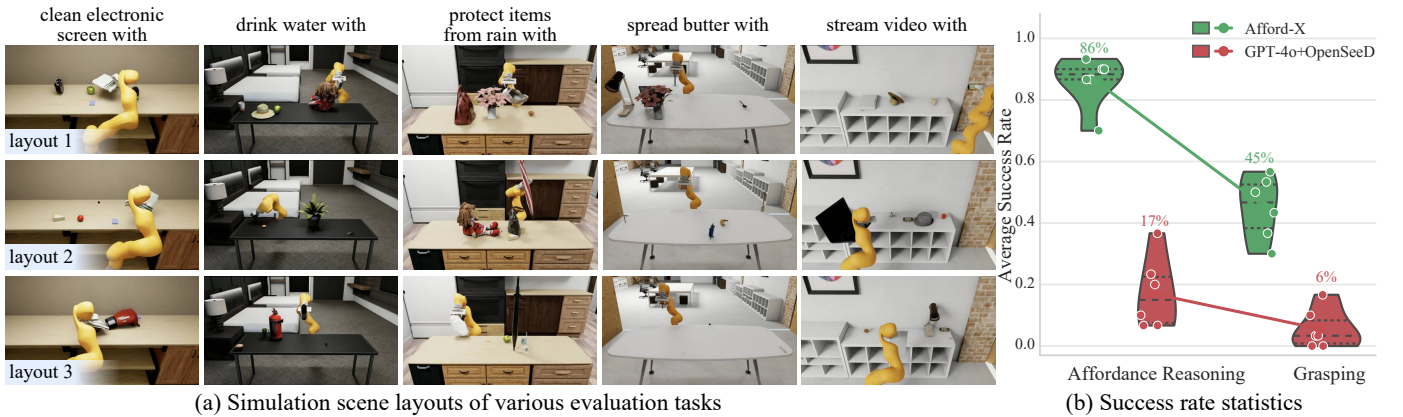
(b) Analysis of method performance in affordance reasoning

Fig. 11: Comparison between Afford-X and LLM/MLLM-based methods. Afford-X demonstrates superior performance in affordance reasoning through two key analyses. First, quantitative precision comparisons reveal significant advantages over LLM/MLLM-based methods, as shown in (a) the violin plot where statistical significance is denoted by stars. Second, (b) qualitative analysis highlights Afford-X's ability to overcome two fundamental limitations of these pipelines: textual ambiguity and limited visual grounding. While LLMs can suffer from semantic uncertainty and MLLMs may still struggle with precise task-relevant grounding, Afford-X directly identifies and selects objects based on their functional affordances.

integration showed improvements, limited caption semantic density continued to impede accurate object identification.

Reason-before-detection pipelines showed improved results, with GPT-4V + OpenSeeD achieving 49.61% recall and

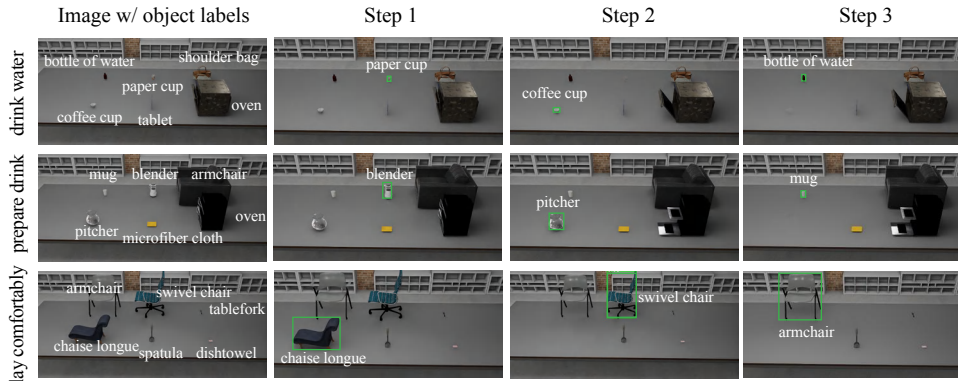
50.63% mAP. However, this performance remained slightly below the SPHINX model, potentially due to GPT-4V's inherent value-based constraints on safety and societal norms, particularly evident in scenarios like "hold water" tasks where



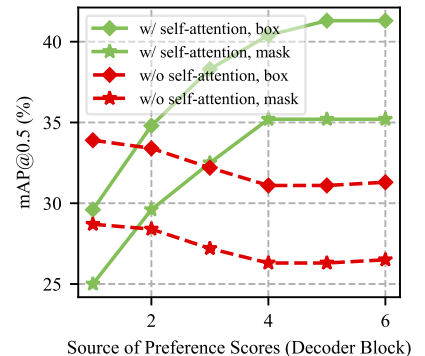
(a) Simulation scene layouts of various evaluation tasks

(b) Success rate statistics

Fig. 12: Task-oriented object grasping across different layouts in simulated environments. (a) As a showcase for task-oriented manipulation, the robotic system integrated with Afford-X successfully identifies the appropriate object across various spatial and object configurations in simulated environments, and executes grasping as a showcase of task-oriented manipulation. (b) Success rates of each step throughout the manipulation trial demonstrate Afford-X's capability in supporting task-oriented manipulation in simulated environments.



(a) Afford-X's object selection preferences.



(b) Impact of self-attention.

Fig. 13: **Sequential object preference and self-attention analysis.** (a) We demonstrate Afford-X's affordance-based selection process across three sequential steps. The model exhibits hierarchical preferences as objects are progressively removed from the environment. Each experimental row presents a different task scenario where available objects are labeled to show possible choices. (b) The performance curves reveal the significant impact of self-attention layers on preference modeling. Our analysis shows that models with self-attention consistently achieve higher accuracy across decoder blocks compared to variants without this mechanism.

TABLE VIII: **Computational efficiency comparison.** Analysis of FPS, parameters, and GPU memory across LLM/MLLM-based pipelines and our proposed approaches, measured on a standard NVIDIA 3090 GPU (24GB). SPHINX parameters include the visual encoder, and GPU Mem. denotes local deployment memory.

Index	Method	FPS	Parameters	GPU Mem.
(a)	Detection + GPT-4	1.18	>369M	—
(b)	Detection + BLIP [40] + GPT-4	0.27	>498M	—
(c)	GPT-4 + OpenSeeD [92]	0.11	>116M	> 1.3 GB
(d)	GPT-4V + OpenSeeD	0.04	>116M	> 1.3 GB
(e)	SPHINX [26] + OpenSeeD	0.32	3.2B	11.5 GB
(f)	SPHINX (CoT)	0.34	3.1B	10.0 GB
(g)	COCO-Aff	2.38	187M	3.2 GB
(h)	LVIS-Aff	2.38	187M	3.2 GB

unconventional but viable solutions like bamboo tubes might be overlooked. End-to-end MLLM systems, while promising, faced practical limitations—the memory-constrained SPHINX tiny model required supplementary detection algorithms or chain-of-thought prompting for bounding box output, achieving 39.76% mAP at 0.49 FPS, indicating limited practical utility.

In contrast, Afford-X demonstrated superior performance through direct bounding box output, achieving 60.67% mAP on COCO-Aff and 67.26% mAP on LVIS-Aff while maintaining 2.38 FPS. These results emphasize the importance of specialized task-oriented datasets for training efficient end-to-end models. Qualitative analysis through four representative case studies revealed the limitations of existing approaches—detection models struggled with semantic ambiguity (particularly in “glass” interpretation between structural glass and drinking vessels), while LLMs without visual perception showed a tendency to infer non-existent objects. Both MLLM-based pipelines and our Afford-X exhibited advanced affordance understanding, successfully identifying unconventional affordances, such as recognizing modified green peppers as viable containers.

This comprehensive evaluation demonstrates that domain-specific training can outperform generic pre-trained MLLMs for affordance reasoning tasks, suggesting a promising direc-

tion for practical robotic applications.

F. Task-oriented Object Grasping

We have conducted comprehensive evaluations through a series of systematic experiments in simulated environments, designed to assess Afford-X's adaptability in diverse indoor contexts, utilizing controlled virtual environments to ensure precise manipulation of experimental variables, including object placement, scene complexity, and task parameters.

Robot Evaluation in Indoor Scenarios: Our primary evaluation focuses on the robot's performance across distinct indoor settings—dining room, living room, and office—each rendered in multiple stylistic variants with varying object configurations. We tasked the robot with performing affordance-based grasping (e.g., “drink water with”) that align with typical user interactions in residential and workplace settings. Figs. A7 and 12 illustrates the results in a table-top manipulation setting using a fixed KUKA IIWA-Panda robot arm. The manipulation is performed within various scenes featuring substantial variation in object placement, room layouts, and environmental complexity, providing a comprehensive test of our model's adaptability.

Fig. 12(b) reports the cumulative success rates of the manipulation pipeline over 150 trials (6 tasks $\times$ 5 scenes $\times$ 5 random object layouts), where we compare Afford-X against the powerful baseline “GPT-4o + OpenSeeD”. The success rates are evaluated for two key steps: affordance reasoning (step i in Sec. V-B), and grasping (steps ii-iv in Sec. V-B). For Afford-X, affordance reasoning succeeds in 86% of the trials, showing that Afford-X can reliably select and segment task-relevant objects in 3D simulated environments. The grasp execution succeeds in 45% of the trials (52.3% of trials with successful affordance reasoning), where most failures occur during grasp planning and execution (e.g., grasping thin and flat objects like spoon and fork) that are challenging for a generic grasp pose planner. The baseline shows 17% and 6% success rates in the two steps. These results demonstrate that Afford-X effectively provides object-level grounding for

task-oriented object grasping, and reveals the valuable future study of integrating finer part-level affordance reasoning and task-specific manipulation strategies for more advanced task-oriented manipulation. Further, Sec. M extends this manipulation to identifying and collecting appropriate objects for a complex task in a larger interior scene with more objects in sight. After decomposing the complex task into multiple steps with an LLM-based task planner, a mobile base manipulator grasps and collects the objects identified by Afford-X.

Preference on Multiple Options: In real-world scenarios where multiple options are available for the same task, Afford-X demonstrates sophisticated preference modeling aligned with human utility patterns. For instance, when prompted with “*prepare drinks with*,” the model exhibits hierarchical understanding of utility, prioritizing a blender, followed by a pitcher, and finally a mug. To systematically evaluate this capability, we designed an experimental setup featuring six candidate objects placed on a large table in an interior scene, including three positive object categories for specific tasks (e.g., “*prepare drinks with*” or “*lay comfortably on*”). Through an iterative elimination procedure, the robot selects the object with the highest score, which is subsequently removed from the environment for the next query. This process, illustrated in Fig. 13a, thoroughly tests the model’s ability to dynamically adapt its preferences. Similarly, for the task “*lay comfortably on*,” the model demonstrates nuanced understanding by favoring a lounge over a chair, indicating awareness of comfort-oriented furniture hierarchies.

The model’s sophisticated preference modeling capability is fundamentally enabled by self-attention (SA) layers in the decoder. To validate this mechanism, we conducted a comparative analysis between two model variants—one incorporating SA layers and one without—while maintaining identical parameter budgets to isolate the SA component’s contribution. Using bounding boxes and masks generated by the final decoder block, we computed mean Average Precision (mAP) values based on preference scores from each intermediate decoder layer. The SA-equipped model demonstrates progressive improvement through deeper decoder layers, achieving final scores of 41.3% (mAP^{box}) and 35.2% (mAP^{mask}) from initial values of 29.6% and 25.0%, respectively. In contrast, the non-SA version shows limited improvement, reaching maximum values of only 33.9% (mAP^{box}) and 28.7% (mAP^{mask}). These results directly implicate SA layers as crucial components in capturing and refining pairwise preferences—critical for strategic decision-making during progressive selection.

For comprehensive results and detailed analyses beyond the scope of this section, readers are directed to Sec. K.

VII. CONCLUSION

We present Afford-X, a novel framework addressing a fundamental challenge in robotics: achieving sophisticated affordance reasoning while maintaining computational efficiency for local deployment. Our approach integrates a noun-pronoun distillation pipeline with specialized VA and BF modules, complemented by an automated pipeline for constructing comprehensive affordance reasoning datasets—COCO-Aff and

LVIS-Aff. Through extensive experimentation, we demonstrate our framework’s enhanced generalization capabilities while maintaining real-time processing speeds, validated through robotic manipulation experiments in simulated environments.

Our current framework faces two primary limitations. First, geometric features alone prove insufficient for certain affordance reasoning scenarios—for instance, distinguishing between drinking cups and toothbrush holders despite identical geometric features, where usage context and hygiene considerations necessitate different affordances. Second, the lack of 3D information constrains system performance in complex spatial environments, particularly in assessing object relationships and occlusion scenarios. Future work could address them through integration of contextual knowledge bases and incorporation of depth information to enhance spatial understanding.

REFERENCES

- [1] J. J. Gibson, “The theory of affordances. the ecological approach to visual perception,” *The people, place and, space reader*, pp. 56–60, 1979.
- [2] Y. Zhu, T. Gao, L. Fan, S. Huang, M. Edmonds, H. Liu, F. Gao, C. Zhang, S. Qi, Y. N. Wu, *et al.*, “Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense,” *Engineering*, vol. 6, no. 3, pp. 310–345, 2020.
- [3] D. Marr, *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010.
- [4] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, “Mdetr-modulated detection for end-to-end multi-modal understanding,” in *ICCV*, 2021.
- [5] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, “Scaling up visual and vision-language representation learning with noisy text supervision,” in *ICML*, 2021.
- [6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, “Learning transferable visual models from natural language supervision,” in *ICML*, 2021.
- [7] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, *et al.*, “Grounded language-image pre-training,” in *CVPR*, 2022.
- [8] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, *et al.*, “Grounding dino: Marrying dino with grounded pre-training for open-set object detection,” *arXiv preprint arXiv:2303.05499*, 2023.
- [9] M. Qin, J. Brawer, and B. Scassellati, “Robot tool use: A survey,” *Frontiers in Robotics and AI*, vol. 9, p. 1009488, 2023.
- [10] R. Wu, K. Cheng, Y. Zhao, C. Ning, G. Zhan, and H. Dong, “Learning environment-aware affordance for 3d articulated object manipulation under occlusions,” in *NeurIPS*, 2024.
- [11] K. Vaesen, “The cognitive bases of human tool use,” *Behavioral and Brain Sciences*, vol. 35, no. 4, pp. 203–218, 2012.
- [12] T. McCormack, C. Hoerl, and S. Butterfill, *Tool use and causal cognition*. Oxford University Press, 2011.
- [13] Y. Zhu, Y. Zhao, and S. Chun Zhu, “Understanding tools: Task-oriented object modeling, learning and recognition,” in *CVPR*, 2015.
- [14] M. Hassani, S. Khan, and M. Tahtali, “Visual affordance and function understanding: A survey,” *ACM Computing Surveys (CSUR)*, vol. 54, no. 3, pp. 1–35, 2021.
- [15] P. Li, B. Tian, Y. Shi, X. Chen, H. Zhao, G. Zhou, and Y.-Q. Zhang, “Toist: Task oriented instance segmentation transformer with noun-pronoun distillation,” in *NeurIPS*, 2022.
- [16] J. Tang, G. Zheng, J. Yu, and S. Yang, “Cotdet: Affordance knowledge prompting for task driven object detection,” in *ICCV*, 2023.
- [17] M. Qu, Y. Wu, W. Liu, X. Liang, J. Song, Y. Zhao, and Y. Wei, “Rio: A benchmark for reasoning intention-oriented objects in open environments,” in *NeurIPS*, 2024.
- [18] K. Fang, Y. Zhu, A. Garg, A. Kurenkov, V. Mehta, L. Fei-Fei, and S. Savarese, “Learning task-oriented grasping for tool manipulation from simulated self-supervision,” *IJRR*, vol. 39, no. 2-3, pp. 202–216, 2020.
- [19] Z. Wang and G. Tian, “Task-oriented robot cognitive manipulation planning using affordance segmentation and logic reasoning,” *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
- [20] Z. Zhang, Z. Jiao, W. Wang, Y. Zhu, S.-C. Zhu, and H. Liu, “Understanding physical effects for effective tool-use,” *RA-L*, vol. 7, no. 4, pp. 9469–9476, 2022.

- [21] K. R. Allen, K. A. Smith, and J. B. Tenenbaum, "Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning," *PNAS*, vol. 117, no. 47, pp. 29302–29310, 2020.
- [22] J. Suder, K. Podbucki, and T. Marciniak, "Power requirements evaluation of embedded devices for real-time video line detection," *Energies*, vol. 16, no. 18, p. 6677, 2023.
- [23] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
- [24] F. J. R. Lera, C. F. Llamas, Á. M. Guerrero, and V. M. Olivera, "Cybersecurity of robotics and autonomous systems: Privacy and safety," in *Robotics-legal, ethical and socioeconomic impacts*, InTech London, UK, 2017.
- [25] I. N. Mseer and S. M. Ahmed, "Artificial intelligence and security challenges," in *From Industry 4.0 to Industry 5.0: Mapping the Transitions*, pp. 49–55, Springer, 2023.
- [26] Z. Lin, D. Liu, R. Zhang, P. Gao, L. Qiu, H. Xiao, H. Qiu, W. Shao, K. Chen, J. Han, S. Huang, Y. Zhang, X. He, Y. Qiao, and H. Li, "Sphinx: A mixer of weights, visual embeddings and image scales for multi-modal large language models," in *ECCV*, 2024.
- [27] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "Vi-bert: Pre-training of generic visual-linguistic representations," in *ICLR*, 2019.
- [28] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *NeurIPS*, 2019.
- [29] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, *et al.*, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *ECCV*, 2020.
- [30] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *ECCV*, 2020.
- [31] J. Lu, V. Goswami, M. Rohrbach, D. Parikh, and S. Lee, "12-in-1: Multi-task vision and language representation learning," in *CVPR*, 2020.
- [32] S. Qian, W. Chen, M. Bai, X. Zhou, Z. Tu, and L. E. Li, "Affordancellm: Grounding affordance from vision language models," in *CVPR*, 2024.
- [33] X. Wang, G. Chen, G. Qian, P. Gao, X.-Y. Wei, Y. Wang, Y. Tian, and W. Gao, "Large-scale multi-modal pre-trained models: A comprehensive survey," *Machine Intelligence Research*, vol. 20, no. 4, pp. 447–482, 2023.
- [34] K. Fang, F. Liu, P. Abbeel, and S. Levine, "Moka: Open-world robotic manipulation through mark-based visual prompting," *RSS*, 2024.
- [35] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.
- [36] A. Gupta, P. Dollar, R. Girshick, K. He, P. Dollar, and R. Girshick, "Lvis: A dataset for large vocabulary instance segmentation," in *CVPR*, 2019.
- [37] J. Xin, L. Wang, S. Wang, D. Kong, J. Li, and B. Yin, "A visual affordance reasoning network based on graph attention," in *International Conference on Digital Home (ICDH)*, 2022.
- [38] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *CVPR*, 2017.
- [39] C.-Y. Chuang, J. Li, A. Torralba, and S. Fidler, "Learning to act properly: Predicting and explaining affordances from images," in *CVPR*, 2018.
- [40] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*, 2022.
- [41] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, E. VanderBilt, L. Schmidt, K. Ehsani, A. Kembhavi, and A. Farhadi, "Objaverse: A universe of annotated 3d objects," in *CVPR*, 2023.
- [42] C. Li, R. Zhang, J. Wong, C. Gokmen, S. Srivastava, R. Martín-Martín, C. Wang, G. Levine, M. Lingelbach, J. Sun, *et al.*, "Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation," in *CoRL*, 2023.
- [43] A. Nguyen, D. Kanoulas, D. G. Caldwell, and N. G. Tsagarakis, "Detecting object affordances with convolutional neural networks," in *IROS*, 2016.
- [44] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine Learning*, vol. 29, pp. 131–163, 1997.
- [45] W. S. Noble, "What is a support vector machine?," *Nature Biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.
- [46] M. Lopes, F. S. Melo, and L. Montesano, "Affordance-based imitation learning in robots," in *IROS*, 2007.
- [47] L. Montesano, M. Lopes, A. Bernardino, and J. Santos-Victor, "Learning object affordances: from sensory-motor coordination to imitation," *T-RO*, vol. 24, no. 1, pp. 15–26, 2008.
- [48] E. Uğur and E. Şahin, "Traversability: A case study for learning and perceiving affordances in robots," *Adaptive Behavior*, vol. 18, no. 3-4, pp. 258–284, 2010.
- [49] D. Chen, D. Kong, J. Li, S. Wang, and B. Yin, "A survey of visual affordance recognition based on deep learning," *IEEE Transactions on Big Data*, 2023.
- [50] T.-T. Do, A. Nguyen, and I. Reid, "Affordancenet: An end-to-end deep learning approach for object affordance detection," in *ICRA*, 2018.
- [51] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *JMLR*, vol. 17, no. 39, pp. 1–40, 2016.
- [52] S. Gupta, J. Davidson, S. Levine, R. Sukthankar, and J. Malik, "Cognitive mapping and planning for visual navigation," in *CVPR*, 2017.
- [53] J. Sawatzky, Y. Souri, C. Grund, and J. Gall, "What object should i use?-task driven object detection," in *CVPR*, 2019.
- [54] Y. Zhu, A. Fathi, and L. Fei-Fei, "Reasoning about object affordances in a knowledge base representation," in *ECCV*, 2014.
- [55] T. Nguyen, M. N. Vu, A. Vuong, D. Nguyen, T. Vo, N. Le, and A. Nguyen, "Open-vocabulary affordance detection in 3d point clouds," in *IROS*, 2023.
- [56] T. Van Vo, M. N. Vu, B. Huang, T. Nguyen, N. Le, T. Vo, and A. Nguyen, "Open-vocabulary affordance detection using knowledge distillation and text-point correlation," in *ICRA*, 2024.
- [57] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," in *NeurIPS*, 2022.
- [58] G. Li, D. Sun, L. Sevilla-Lara, and V. Jampani, "One-shot open affordance learning with foundation models," in *CVPR*, 2024.
- [59] H. Chu, X. Deng, Q. Lv, X. Chen, Y. Li, J. Hao, and L. Nie, "3d-affordancellm: Harnessing large language models for open-vocabulary affordance detection in 3d worlds," in *ICLR*, 2025.
- [60] D. Lu, L. Kong, T. Huang, and G. H. Lee, "Geal: Generalizable 3d affordance learning with cross-modal consistency," in *CVPR*, 2025.
- [61] C. Xu, Y. Chen, H. Wang, S.-C. Zhu, Y. Zhu, and S. Huang, "Partafford: Part-level affordance discovery from 3d objects," in *ECCV VOLI Workshop*, 2022.
- [62] A. Myers, C. L. Teo, C. Fermüller, and Y. Aloimonos, "Affordance detection of tool parts from geometric features," in *ICRA*, 2015.
- [63] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao, "One-shot affordance detection," in *IJCAI*, 2021.
- [64] L. Lu, W. Zhai, H. Luo, Y. Kang, and Y. Cao, "Phrase-based affordance detection via cyclic bilateral interaction," *IEEE Transactions on Artificial Intelligence*, vol. 4, no. 5, pp. 1186–1198, 2022.
- [65] W. Zhai, H. Luo, J. Zhang, Y. Cao, and D. Tao, "One-shot object affordance detection in the wild," *IJCV*, vol. 130, no. 10, pp. 2472–2500, 2022.
- [66] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.
- [67] K. Ikeuchi and M. Hebert, "Task-oriented vision," in *Exploratory vision: the active eye*, pp. 257–277, Springer, 1996.
- [68] A. T. Miller, S. Knoop, H. I. Christensen, and P. K. Allen, "Automatic grasp planning using shape primitives," in *ICRA*, 2003.
- [69] G. Vezzani, U. Pattacini, and L. Natale, "A grasping approach based on superquadric models," in *ICRA*, 2017.
- [70] D. Song, C. H. Ek, K. Huebner, and D. Kragic, "Task-based robot grasp planning using probabilistic inference," *T-RO*, vol. 31, no. 3, pp. 546–561, 2015.
- [71] V. Narayanan and M. Likhachev, "Task-oriented planning for manipulating articulated mechanisms under model uncertainty," in *ICRA*, 2015.
- [72] R. Gong, J. Huang, Y. Zhao, H. Geng, X. Gao, Q. Wu, W. Ai, Z. Zhou, D. Terzopoulos, S.-C. Zhu, *et al.*, "Arnold: A benchmark for language-grounded task learning with continuous states in realistic 3d scenes," in *ICCV*, 2023.
- [73] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, "Rlbench: The robot learning benchmark & learning environment," *RA-L*, vol. 5, no. 2, pp. 3019–3026, 2020.
- [74] Z. Xia, Z. Deng, B. Fang, Y. Yang, and F. Sun, "A review on sensory perception for dexterous robotic manipulation," *IJRR*, vol. 19, no. 2, p. 17298806221095974, 2022.
- [75] A. Agarwal, S. Uppal, K. Shaw, and D. Pathak, "Dexterous functional grasping," in *CoRL*, 2023.
- [76] T. Zhu, R. Wu, J. Hang, X. Lin, and Y. Sun, "Toward human-like grasp: Functional grasp by dexterous robotic hand via object-hand semantic representation," *TPAMI*, vol. 45, no. 10, pp. 12521–12534, 2023.
- [77] G. Li, V. Jampani, D. Sun, and L. Sevilla-Lara, "Locate: Localize and transfer object parts for weakly supervised affordance grounding," in *CVPR*, 2023.

- [78] C. Zhang and G. H. Lee, “Iaao: Interactive affordance learning for articulated objects in 3d environments,” in *CVPR*, 2025.
- [79] L. Wu, W. Wei, P. Yu, and J. Lan, “Open-vocabulary 3d affordance understanding via functional text enhancement and multilevel representation alignment,” in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025.
- [80] C.-Y. Chuang, J. Li, A. Torralba, and S. Fidler, “Learning to act properly: Predicting and explaining affordances from images,” in *CVPR*, 2018.
- [81] H. W. Kuhn, “The hungarian method for the assignment problem,” *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.
- [82] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, “Generalized intersection over union: A metric and a loss for bounding box regression,” in *CVPR*, 2019.
- [83] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” in *3DV*, 2016.
- [84] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *ICCV*, 2017.
- [85] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *TPAMI*, vol. 39, no. 6, pp. 1137–1149, 2016.
- [86] Y. Li, H. Mao, R. Girshick, and K. He, “Exploring plain vision transformer backbones for object detection,” in *ECCV*, 2022.
- [87] H.-S. Fang, C. Wang, M. Gou, and C. Lu, “Graspnet-1billion: A large-scale benchmark for general object grasping,” in *CVPR*, 2020.
- [88] B. Sundaralingam, S. K. S. Hari, A. Fishman, C. Garrett, K. Van Wyk, V. Blukis, A. Millane, H. Oleynikova, A. Handa, F. Ramos, *et al.*, “Curobo: Parallelized collision-free robot motion generation,” in *ICRA*, IEEE, 2023.
- [89] Y. Liu, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, vol. 364, 2019.
- [90] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *CVPR*, 2016.
- [91] Y. Zhang, X. Huang, J. Ma, Z. Li, Z. Luo, Y. Xie, Y. Qin, T. Luo, Y. Li, S. Liu, *et al.*, “Recognize anything: A strong image tagging model,” in *CVPR*, 2024.
- [92] H. Zhang, F. Li, X. Zou, S. Liu, C. Li, J. Yang, and L. Zhang, “A simple framework for open-vocabulary segmentation and detection,” in *ICCV*, 2023.



Xiaomeng Zhu is a Ph.D. student in the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology. She received her B.Eng. degree from the University of Electronic Science and Technology of China and her M.S. degree from the Institute of Automation, Chinese Academy of Sciences. Her research interests include robotic perception, human intent understanding, and human-robot collaboration.



Yuyang Li is a Ph.D. candidate at the Institute for AI, Peking University. He received his bachelor's degree in engineering from the Department of Automation at Tsinghua University in 2024. He aspires to advance the development of embodied intelligence with versatile dexterity based on multi-modal perception.



Leiyao Cui received his B.Eng. degree from the school of Information and Electronics at the Beijing Institute of Technology in 2024. Currently, he is a Ph.D. student at the Shenyang Institute of Automation, Chinese Academy of Sciences. He is also a research intern at the Institute for AI, Peking University. His research interest currently focuses on robot perception.



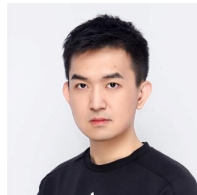
Pengfei Li is a fourth-year Ph.D. student at the Institute for AI Industry Research, Tsinghua University. He received the B.E. degree in Computer Science and Technology from University of Chinese Academy of Sciences, Beijing, China, in 2022. His research interests include autonomous driving and robotics.



Huan-ang Gao is a second-year PhD student at the Department of Computer Science and Technology and the Institute for AI Industry Research at Tsinghua University. He earned his bachelor's degree in engineering from the same department at Tsinghua University in 2024. His research focuses on developing generative simulation methods for evaluating and training embodied AI policies.



Yixin Zhu received the Ph.D. degree in statistics from the University of California, Los Angeles (UCLA), in 2018. He is currently an Assistant Professor with Peking University, Beijing, China, jointly appointed in the School of Psychological and Cognitive Sciences and the Institute for Artificial Intelligence. His research interests include cognitive reasoning, embodied intelligence, computer vision, human-centric AI, and BCI, with a particular emphasis on integrating commonsense knowledge with multimodal perception to build interactive intelligent systems. He has published in *Science*, *Nature Machine Intelligence*, *Nature Human Behaviour*, *Science Robotics*, and *Science Advances*, among others.



Hao Zhao received the B.E. degree and the Ph.D. degree both from the EE department of Tsinghua University, Beijing, China. He is currently an Assistant Professor with the Institute for AI Industry Research (AIR), Tsinghua University. He was a research scientist at Intel Labs China and a joint postdoc affiliated to Peking University. His research interests cover various computer vision topics related to robotics, especially 3D scene understanding. His work has been recognized with the Best Student Paper Award at CVPR 2026 and the Best Paper Award at CVPR 2026 and the Best Paper Award at 3DV 2024, finalist selections at ICRA 2026 and RSS 2026, and a runner-up award at CICA 2023.