Beyond Object Recognition: A New Benchmark towards Object Concept Learning

Yong-Lu Li, Yue Xu, Xinyu Xu, Xiaohan Mao, Yuan Yao, Siqi Liu, Cewu Lu* Shanghai Jiao Tong University

{yonglu_li, silicxuyue, xuxinyu2000, mxh1999, yaoyuan2000, magi-yunan, lucewu}@sjtu.edu.cn



Figure 1: For embodied agents, understanding daily objects requires the ability to perceive not only **category** but also **attribute** and **affordance**. In OCL, we try to reveal object concept learning in both three levels and explore their profound causal relations.

Abstract

Understanding objects is a central building block of AI, especially for embodied AI. Even though object recognition excels with deep learning, current machines struggle to learn higher-level knowledge, e.g., what attributes an object has, and what we can do with it. Here, we propose a challenging **Object Concept Learning** (OCL) task to push the envelope of object understanding. It requires machines to reason out affordances and simultaneously give the reason: what attributes make an object possess these affordances. To support OCL, we build a densely annotated knowledge base including extensive annotations for three levels of object concept (category, attribute, affordance), and the clear causal relations of three levels. By analyzing the causal structure of OCL, we present a baseline, Object Concept Reasoning Network (OCRN). It leverages concept instantiation and causal intervention to infer the three levels. In experiments, OCRN effectively infers the object knowledge while following the causalities well. Our data and code are available at https://mvig-rhos.com/ocl.

1. Introduction

Object understanding is essential for intelligent robots. Recently, benefiting from deep learning and large-scale datasets [1, 2], category recognition [3, 4] has made tremendous progress. But to close the gap between human and machine perception, machines need to pursue deeper understanding, *e.g.*, recognizing higher-level attributes [5] and affordances [6], which may help it establish object concept [7] when interacting with contexts.

Category apple is a symbol indicating its referent (real apples). In line with symbol grounding [8], machines should learn knowledge beyond category to approach concept understanding. According to cognition studies [9, 7], attribute depicting objects from the physical/visual side plays an important role in object understanding. Thus, many works [10, 11, 12] studied to ground objects with attributes, *e.g.*, a hammer consists of a long handle and a heavy head. Moreover, attributes can depict object states [5]. An elegant characteristic of attributes is *cross-category*: objects of the same category can have various states (big or fresh apple), whilst various objects can have the same state (sliced orange or apple). If the category is the **first** level of object concept, the attribute can be seen as the **second** level closer to the physical fact.

However, recognizing attributes is still far away from concept understanding. Given a hammer, we should know it can be held to hit nails, *i.e.*, requiring machines to infer affordance [6] indicating what actions humans can perform with objects. Thus, we refer to affordance as the **third** level, which is closely related to common sense and causal inference [6]. Though affordance has been

^{*}Corresponding author.

studied in robotics [13, 14] and vision [15, 16] communities for decades, it is still challenging. First, previous works [17, 18] often focus on recognizing affordance solely. But we usually infer affordance based on attribute observation. If we need to knock in a nail without a hammer at hand, we may find other hard or heavy objects instead, *e.g.*, a thick book. This profoundly reveals the **causality** between attribute and affordance. Second, previous works are designed for scale/scene-limited tasks, *e.g.*, in [16], 40 objects and 14 affordances are included; Hermans *et al.* [14] collect 375 indoor images of 6 objects, 21 attributes, and 7 affordances; a recent dataset [17] contains 10 indoor objects and 9 affordances. Thus, they cannot afford general affordance reasoning for large-scale applications.

To reshape object learning, we believe it is essential to look at the above three levels in a unified and causal way based on an extensive knowledge base. Hence, we move a step forward to propose the object concept learning (OCL) task: given an object, machines need to infer its category, attributes, and further answer "what can we do upon it and why", as shown in Fig. 1. In a nutshell, machines need to reason affordance based on object appearance, category, and attributes. To this end, we build a largescale and dense dataset consisting of 381 categories, 114 attributes, and 170 affordances. It contains 80,463 images of diverse scenes and 185,941 instances in different states. Different from previous works [19, 14, 16], OCL offers a more subtle angle. It includes: (1) category-level attribute (A) and affordance (B) labels; (2) instance-level attribute (α) and affordance (β) labels. Besides, we annotate the causal relations between three levels to evaluate the reasoning ability of models and keep the follow-up methods from fitting data only. Accordingly, based on the causal structure of OCL, we propose a *neuro-causal* method, Object Concept Reasoning Network (OCRN), as the future baseline. It leverages concept instantiation (from category-level to instance-level) and causal intervention [20] to infer attributes and affordances. OCRN outperforms a host of baselines and shows impressive performance while following the causal relations well.

In summary, our contributions are threefold:

- (1) Introducing the object concept learning task poses challenges and opportunities for object understanding and knowledge-based reasoning.
- (2) Building a benchmark consisting of diverse objects, elaborate attributes, and affordances, together with their clear causal relations.
- (3) An object concept reasoning network is introduced to reason three levels with concept instantiation performing well on OCL.

2. Related Work

Object Attribute depicts the physical properties like color, size, shape, etc. It usually plays the role of intermedia between pixels and higher-level concepts, e.g., prompting object recognition [12], affordance learning [14], zero-shot learning [10], and object detection [21]. Recently, several large-scale datasets [12, 11, 22, 23, 5, 24, 25] are released. For attribute recognition, besides direct attribute classification [10, 26, 11, 23] and leveraging the correlation between attribute-attribute and attribute-object [27, 28, 29], intrinsic properties (compositionality, contextuality [30, 31], symmetry [32, 33]) of attribute-object are also proven useful. [30] uses the model weight space to encode the attributes to model the compositionality and contextuality. [31] uses the attributes as linear operators to transform object embeddings. [32] leverages the symmetry property to model the attribute changes within attribute-object coupling and decoupling.

Object Affordance. is introduced by [6]. Affordance learning has two canonical paradigms: direct mapping [18] or indirect method [16, 34, 35, 36] with intermediates like object category, attribute, and 3D contents. Some works learned affordance from human-object interactions (HOI) to encode the relation between object and action [37, 38, 39]. Visual Genome [24] provides relations between objects, including actions instead of affordances. However, these relations cover limited and sparse affordances. Differently, we use easily accessible object images as the knowledge source and densely annotate all attributes/affordances for all objects. Besides the vision community, the robot community pays much attention to affordance [40, 41, 42] for grasping and manipulation. For instance, [40] utilized the robot to discover the object affordance via self-supervised learning. Recently, several datasets [17, 19, 15] have been proposed. IIT-AFF [17] collected ten daily indoor objects and provided nine common affordances to construct a dataset for robot applications. Zhu et al. [16] built a dataset containing attribute, affordance, human pose, and HOI spatial configuration. But labeling pose [43] and HOI [44, 45] are costly. Chao et al. [19] proposed a semantic category-level affordance dataset including 91 objects [2] and 957 affordances.

Causal Inference. There is increasing literature on exploiting causal inference [20] in machine learning, especially with causal graphical models [46, 20], including feature selection [47] and learning [48], video analysis [49, 50], reinforcement learning [51, 52], *etc.* Recently, Wang *et al.* [53] studied the causal relation between objects in images and used intervention [20] to alleviate the observation bias. Atzmon *et al.* [54] analyze the causal generative model of compositional zero-shot learning and disentangle the representations of attributes and objects. Here, we explore the causal relations between three object levels

| Dataset | # Image | # Instance | # Object | # Attribute | # Affordance |
|---------------------|---------|------------|----------|-------------|--------------|
| APY [12] | 15,339 | 15,339 | 32 | 64 | / |
| SUN [11] | 14,340 | 14,340 | 717 | 102 | / |
| COCO-a [23] | 84,044 | 188,426 | 29 | 196 | / |
| ImageNet150k [22] | 150,000 | 150,000 | 1,000 | 25 | / |
| Chao et al. [19] | / | / | 91 | / | 957 (B) |
| Hermans et.al. [14] | 375 | - | 6 | 21 | 7 |
| Zhu et al. [16] | 4,000 | 4,000 | 40 | 57 | 14 |
| OCL | 80,463 | 185,941 | 381 | 114 | 170 |
| | | | | | |

Table 1: *Dense annotated* datasets. OCL provides categoryand instance- level attributes (A, α) , affordances (B, β) .

and apply backdoor adjustment [20] to alleviate the existing bias.

3. Constructing OCL Benchmark

We construct a benchmark to characterize abundant object knowledge following Fig. 2.

3.1. Fine-Grained Object Knowledge Base

Data Collection. We briefly introduce the collection of affordances, categories, attribute classes, and image sources here.

- (1) **Affordance**: We collect 170 affordances out of 1,006 candidates from widely-used action/affordance datasets [19, 55, 44, 56, 16, 17] given generality and commonness.
- (2) **Category**: Considering the taxonomy (WordNet [57]) and diversity, we collect 381 objects out of 1,742 candidates from object datasets [12, 11, 2, 23, 1, 22].
- (3) Attribute: We manually filter the 500 most frequent attributes from attribute datasets [12, 11, 2, 23, 1, 22, 24] and choose 114 attributes, covering colors, deformations, supercategories, surface, geometrical and physical properties.
- (4) Image: We extract 75,578 images from object datasets [12, 11, 2, 23, 1, 22, 24], together with Ground Truth (GT) boxes. We manually collected 4,885 Internet images of selected categories to ensure diversity. Then, we annotate the missing box and category labels for all instances. Finally, 185,941 instances of 381 categories from 80,463 images are collected: an average of 488 instances per category and 2.31 boxes per image. Details are given in the supplementary. OCL is long-tail distributed, where the head categories have over 5,000 instances each, but the rarest categories have only 9 instances, which challenges the robustness of machines greatly.

Annotating Attribute in two levels of granularity: (1) **Category-level** attribute (*A*) contains common sense. For each category, we annotate its *most common* attributes. In concept learning, the usage of the category-level labels as common knowledge can date back to [58]. Following [58], to avoid bias, annotators are given *category-attribute pairs* (category names instead of images) and multiple annotators vote to build the binary A matrix M_A in size of [381, 114]. (2) **Instance-level** attribute (α) is the individual attributes of *each instance*. The annotation unit is an *attribute-instance pair* and each pair is labeled by multiple annotators.

Annotating Affordance in two levels of granularity: (1) Category-level affordance B, similar to A, is annotated in category-affordance pairs, indicating the common affordances of each category. Following [19], the annotators label B matrix M_B in size of [381, 170]. (2) Instance-level affordance β is annotated for *each instance* with the help of object state. As B is defined by common states, objects in specific states may have different affordances from B: if a service robot finds a broken cup, it may infer that the cup can still hold water as it is trained with B labels. Thus, we need detailed β beyond B. β exhibits evident similarities for objects in similar status forming "state" aligning with commonsense, thus we use them to streamline annotation and reduce the annotator discrepancy. A state is defined as an [category, description (e.g., a set of attributes)] pair, and instances in a state usually possess similar affordances, e.g., fresh, juicy, clean oranges are eatable. First, six experts conclude the states by scanning all instances of a category and listing all states according to affordance. Then these states were merged manually. In total, 1,376 states are defined, and each category has 3.6 states on average. Next, β is annotated for each state, and the instances are first assigned with the state-level β . Bext, the instance-level β is **detailed** based on the state-level β according to the visual content of each instance. Note that the state is category-dependent and can not be transferred among object categories, which is different from attribute and affordance. Besides, the composition of attributes makes the state space huge and there can be many unseen states. Thus, we only use them in annotation but not in our method.

Fig. 3 shows some examples of OCL. We compare OCL with previous dense datasets in Tab. 1. More details, figures, and tables are given in the supplementary.

3.2. Causal Graph Definition

We use a causal graph to shed light on the subtle causalities of our knowledge base in Fig. 4. Causal graph [20] indicates the underlying causalities based on components:

- O: object category
- I: object instance in an image
- A: category-level attribute



Figure 2: OCL construction. a) Data collection. b) Annotating category-level attributes and affordances. c) Annotating instance-level attributes and affordances. d) Finding direct and clear instance-level causal relations.



Figure 3: OCL samples including category, α (red), β (blue), and their causal relations in various contexts.

- α : instance-level attribute
- *B*: category-level affordance
- β : instance-level affordance

According to the prior knowledge about the causalities between three levels, a hierarchical structure is depicted: (a) the **inner** triangle with dotted lines is the **category**-level: object category O, category-level attributes A, and affordances B; (b) the **outer** triangle is the **instance**-level: instance visual appearance I, instance-level attributes α , and affordances β . Each directed *possible* arc in the graph indicates the *possible* causality between two nodes.

Here, besides the red arcs indicating the common causal relations (*e.g.*, $I \rightarrow \alpha$, $I \rightarrow \beta$ as attribute/affordance recognition from images), we define some special arcs given our

category-level attribute and affordance settings: (1) $O \rightarrow A$, $O \rightarrow B$ (dotted arcs): Given O, A, B are strictly determined within labels. (2) $O \rightarrow I$, $A \rightarrow \alpha$, $B \rightarrow \beta$ (blue arcs): The category-level O, A, and B are direct causes of instance-level I, α , and β during the concept *instantiation*. Note that, according to the previous analysis, we focus on the $A \rightarrow B$ and $\alpha \rightarrow \beta$ but sometimes the opposite can also happen: $A \leftarrow B$ and $\alpha \leftarrow \beta$ ("or" in Fig. 4). In annotation and experiments, we observe that $\alpha \rightarrow \beta$ is stronger and more common and natural to human perception, so we focus more on $\alpha \rightarrow \beta$ in our causal benchmark (Sec. 3.4).

In this work, we focus on α, β perception $(I \rightarrow \alpha, I \rightarrow \beta)$ and visual reasoning (with *I*, inferring β given α) for embodied AI. Thus, Fig. 4 is simplified. Our knowledge base can support more tasks such as attribute/affordance



(a) OCL: Object Concept Learning

Figure 4: Causal graph of our OCL task. "or" indicates that either $A \leftarrow B$ or $A \rightarrow B$ ($\alpha \rightarrow \beta$ or $\alpha \leftarrow \beta$) exists.

conditioned image generation $(\alpha \rightarrow I, \beta \rightarrow I)$ [59]. However, they are beyond the scope of this paper (Suppl. Sec. 3).

3.3. Causal Inference Benchmark on $\alpha \rightarrow \beta$

We annotate instance-level (considering the context of each instance) causality of $\alpha \rightarrow \beta$ to answer "which attribute(s) are the critical and direct causes of a certain affordance?" in two phrases:

Filtering: Initially, we need to make binary decisions on all *instance*- α - β triplets, which is far beyond handleable. Fortunately, we find that **most** α - β classes (*e.g.*, shiny and kick) are meaningless and always of no causality. Thus, we exclude the most impossible pairs and only annotate existing rules without ambiguity, meanwhile, guaranteeing the completeness of causality. For each of the 114×170 α - β pairs, we attach 10 samples for reference and 3 experts vote yes/no/not-sure. We take the majority vote and the not-sure and controversial pairs are rechecked. The not-sure and no pairs are removed, and so do the ambiguous pairs. Finally, we obtain about 10% α - β classes as candidates. The left 90% pairs may hold value, we plan to use LLMs to mine new rules in future work, especially from ambiguous pairs.

Instance-level Causality: we adopt object states as a reference. Multiple annotators have been involved for each *state*- α - β triplet and are asked whether the specific attribute is the *clear* and *direct* cause of this affordance in this state. The answers are combined and checked for all instances of a state. Finally, we obtain about 2 M *instance*- α - β triplets of causal relations. As we have labeled all α and β for all instances, the causal relations would be in four situations: [0,0], [1,1]; [0,1], [1,0]. The former two are "positive", e.g., $fresh(1/0) \rightarrow eat(1/0)$ for an apple. While the last two are "negative", e.g., broken $(1/0) \rightarrow drive(0/1)$ for a car.

Fig. 3 shows some causal examples. These causalities are not thoroughly studied in previous datasets [16, 17, 18, 14]. For more details, please refer to the supplementary.

3.4. Task Overview

Here, we formulate the OCL task formally. Given an instance I (content in box b_o representing an object instance), OCL aims to infer attribute α and affordance β while following the causalities. Formally, OCL can be described as:

$$\langle P_{\alpha}, P_{\beta} \rangle = \mathcal{F}(I, P(O|I)),$$
 (1)

where P_{α}, P_{β} are the probabilities of $\alpha, \beta, P(O|I)$ is the predicted category probability from an object detector [4].

We aim at benchmarking the reasoning ability of machines, causal relations in Fig. 4 can all be candidates. However, annotating causal relations is usually ambiguous and it is impractical to cover all relations. In a user study, experts met significant divergence when annotating different arcs. For embodied AI, affordance β is more important in robot-world interactions. Moreover, both the causal relation annotation and the ablations support that the causal effect of $\alpha \rightarrow \beta$ is more significant than the other alternatives. Thus, we only annotate the unambiguous $\alpha \rightarrow \beta$ (Sec. 3.3) and mainly measure the learning of $\alpha \rightarrow \beta$ here. Formally, the evaluation of $\alpha \rightarrow \beta$ learning follows

$$\Delta P_{\beta} = ITE[\mathcal{F}(I, P(O|I))], \qquad (2)$$

where ΔP_{β} is the Individual Treatment Effect [60] of **af**fordance prediction change after we operate $ITE[\cdot]$ on a model $\mathcal{F}(\cdot)$. ΔP_{β} is expected to follow the GT causal relation between α, β from humans. For example, when the attributes of an object change, then the causal-related affordances should also change accordingly. We will detail the ITE evaluation in Sec. 5. Note that A, B are decided by O. Given O, we can get A, B via querying the prior M_A, M_B (Sec. 3). Thus, we do not evaluate $A \rightarrow B$ here.

We split images into the train, validation, and test sets with 56K:14K:9K images. The validation and test sets cover 221 of the 381 categories, and the train set covers all categories. OCL is a long-tailed recognition task [61, 62] and requires generalization to cover the whole object category-attribute-affordance space with imbalanced information. Thus, it is challenging for current machines without the reasoning ability to understand the causalities.

4. Object Concept Reasoning Network

Before proposing the OCRN, we first simplify the causal graph in Fig. 4 to facilitate the implementation. We focus on $\alpha \to \beta$ and omit $\beta \to \alpha$. Similarly, we omit $B \to A$. Besides, I, α, β are the *instantiations* of O, A, B respectively and we use a O' node to represent O, A, B. The adapted causal graph is shown in Fig. 5. OCRN implements the instantiation of attribute and affordance, corresponding to $A \to \alpha, B \to \beta$. Thus the model can propose a coarse estimation of attribute and affordance at category-level, then

tune the results with the image patterns as a condition for a more accurate prediction. Besides, we exploit **intervention** to remove the causal relation between I and O to construct a category-agnostic model. It suffers less from category bias and is more capable of learning uncommon cases.

Object Category Bias. OCL can be depicted as $P(\alpha|I)$ and $P(\beta|I, \alpha)$. As the samples of different categories are usually imbalanced, conventional methods may suffer from severe *category bias* [53], *e.g.*, animal accounts for 22% instances in OCL, and home appliance only accounts for 3%. In $P(\alpha|I)$, category bias is imported following

$$P(\alpha|I) = \sum_{i}^{m} P(\alpha|I, O_i) P(O_i|I),$$
(3)

where $P(O_i|I)$ is the predicted category probability. That is, O is a confounder [20] and pollutes attribute inference, especially for the *rare* categories.

Causal Intervention. To tackle this, we propose OCRN using intervention [20] to deconfound the confounder O for α (Fig. 5). In α estimation, we use $do(\cdot)$ operation [20] to eliminate the arc from O to $I: P(\alpha|do(I))$ is

$$\sum_{i}^{m} P(\alpha|I, O_{i})P(O_{i})$$

$$= \sum_{i}^{m} P(O_{i}) \sum_{j}^{m} P(\alpha|I, A_{j})P(A_{j}|O_{i}) \qquad (4)$$

$$= \sum_{i}^{m} P(\alpha|I, A_{i})P(O_{i}),$$

where m = 381. A_j is the category-attribute vector of j^{th} category. As A is decided by O, $P(A_j|O_i) = 1$ if i = j and $P(A_j|O_i) = 0$ if $i \neq j$, where O_i is the i^{th} category and A_j is the category-attribute of j^{th} category. $P(O_i)$ is the **prior** probability of the *i*-th category (frequency in our train set). We apply the intervention to reduce the bias from O recognition for an **category-agnostic** model.

Similar to α , in β estimation, category bias also exists:

$$P(\beta|I,\alpha) = \sum_{i}^{m} P(\beta|I,\alpha,O_i) P(O_i|I,\alpha).$$
(5)

With Eq. 4, α is beforehand estimated and thus can be seen as "enforced" and deconfounded. For *I*, we again use the intervention [20]:

$$P(\beta|do(I,\alpha)) = \sum_{i}^{m} P(\beta|I,\alpha,B_i)P(O_i).$$
 (6)

Similar to Eq. 4, $P(B_j|O_i) = 1$ if i = j, $P(B_j|O_i) = 0$ if $i \neq j$, we omit the process for clarity.

4.1. Model Implementation

We represent nodes $\{I, A, B, \alpha, \beta\}$ as $\{f_I, f_A, f_B, f_\alpha, f_\beta\}$ respectively in latent space. f_I is the RoI pooling feature of an instance extracted by a COCO pre-trained ResNet-50 [63]. Following Eq. 4, we represent category-level attribute A based on the *mean* object category feature \bar{f}_{O_i} , which is the mean of f_I of all **training** samples in category O_i . We map \bar{f}_{O_i} to the attribute latent space f_{A_i} with fully-connected layers (FC) (Fig. 5). f_{A_i} stands for the category-attribute representation for ith category.

Attribute Instantiation. Next, we obtain α representation following Eq. 4:

$$f_{\alpha_i} = \mathcal{F}_{\alpha}(f_I, f_{A_i}), \quad f_{\alpha} = \sum_i^m f_{\alpha_i} \cdot P_{O_i}, \tag{7}$$

where P_{O_i} is the *prior* category probability $(P(O_i))$ in Eq. 4). Eq. 7 indicates the attribute *instantiation* from A to α with I as the *condition*. Hence, we can equally translate the α estimation problem into a **conditioned instantiation problem**. $\mathcal{F}_{\alpha}(\cdot)$ is implemented with multi-head attention [64] with two entries (Fig. 5). The attention output is compressed by a linear layer to the instantiated representation f_{α_i} . The debiased representation f_{α} is the expectation of f_{α_i} w.r.t P_{O_i} according to back-door adjustment in Eq. 4.

We also get the feature for specific attributes for ITE operation (Sec. 5). f_{α} is first separated to f_{α_p} for each attribute $p \ (p \in [1, 114])$ by multiple independent FCs, then we can manipulate specific attributes by masking some certain f_{α_p} . Next, the features are aggregated via concatenatingcompressing by an FC to f'_{α} as shown in Fig. 5.

Affordance Instantiation. Similarly, FCs are used to obtain f_B from \bar{f}_{O_i} and f_{A_i} and Eq. 6 is implemented as:

$$f_{\beta_i} = \mathcal{F}_{\beta}(f_I, f'_{\alpha}, f_{B_i}), \quad f_{\beta} = \sum_i^m f_{\beta_i} \cdot P_{O_i}.$$
(8)

 $\mathcal{F}_{\beta}(\cdot)$ operates instantiation with conditions $\{f_I, f'_{\alpha}, f_{B_i}\}$.

4.2. Learning Objectives.

To drive the learning, we devise several objectives:

Category-level loss L_C . We input category-level f_A , f_B to two linear-Sigmoid classifiers to classify A, B. The binary cross-entropy losses are L_A and L_B . The total category-level loss is $L_C = L_A + L_B$.

Instance-level loss L_I . We input instance-level f_{α} , f_{β} , together with f_{α_i} , f_{β_i} to linear-Sigmoid classifiers. The separated f_{α_p} are also sent to independent binary classifiers. The binary cross-entropy losses are represented as L_{α} , L_{β} . The total instance-level loss is $L_I = L_{\alpha} + L_{\beta}$.

The total loss is $L = \lambda_C L_C + L_I$. We adopt a two-stage policy: first inferring attributes, then reasoning affordances.



Figure 5: OCRN overview. The arc from O to I is deconfounded. Thus, we can eliminate the bias from the O imbalance. Equations below the graphs are the original or deconfounded estimations of α, β . Attribute and affordance modules are the **instantiations** of category-level features: categorical features f_{A_i} or f_{B_i} are obtained following the left-bottom-most causal graph and then instantiated via \mathcal{F}_{α} or \mathcal{F}_{β} conditioned by the instance representations. f_{α} and f_{β} after intervention are the expectations of instantiated f_{α_i} and f_{β_i} w.r.t **prior** P_{O_i} . At last, linear-Sigmoid classifiers give the final predictions.



Figure 6: Example of ITE reasoning benchmark.

5. Experiment

5.1. Metrics

 α, β **Recognition**: we measure the correctness of model prediction $\hat{\alpha}$ and $\hat{\beta}$. For multi-label classification tasks, we use the mean Average Precision (mAP) metric.

Reasoning: we use **Individual Treatment Effect** (ITE) [60]. $ITE_i = Y_{i,T=1} - Y_{i,T=0}$ measures the causal effect $T \to Y$ of ith individual with the difference between outcomes (Y) with or without receiving the treatment (T). In OCL, we discuss the causal relation between pth attribute and qth affordance: $\alpha_p \to \beta_q$. So we interpret the treatment T as the **existence of** α_q and the outcome Y as the β_q output. We measure the difference of β_q output when the whole α_q feature is wiped out or not, which should be non-zero when the causal relation $\alpha_p \to \beta_q$ exists.

In detail, given a model, for an instance with causal relation $\alpha_p \rightarrow \beta_q$ ($p \in [1, 114], q \in [1, 170]$), we first formulate ITE as the affordance probability change following Eq. 2:

$$ITE = \Delta \hat{\beta}_q = \hat{\beta}_q |_{do(\alpha_p)} - \hat{\beta}_q |_{do(\mathbf{x}_q)}.$$
(9)

 $\hat{\beta}_q|_{do(\alpha_p)}$ is the factual output of the affordance probability. $\hat{\beta}_q|_{do(\alpha_p)}$ is the counterfactual output when the α_p is wiped out, which can be got by assign zero-mask [65] to the feature of α_p (e.g., f_{α_p} in OCRN) and keep the other features.

Then, based on ITE, we benchmark instances following: **ITE**: If the causality $\alpha_p \rightarrow \beta_q$ exists on the instance, ITE should be non-zero when eliminating the effect of α_p . And the direction of ITE depends on the affordance ground-truth β_q : if $\beta_q = 0$, the predicted $\hat{\beta}_q$ tend to be 1 after wiping out α_p so ITE should be a negative value; contrarily, ITE should be positive if $\beta_q = 1$. Hence we compute the ITE score as:

$$S_{\text{ITE}} = \begin{cases} \max(\Delta \hat{\beta}_q, 0), & \beta_q = 1, \\ \max(-\Delta \hat{\beta}_q, 0), & \beta_q = 0, \end{cases}$$
(10)

so that larger S_{ITE} indicates the model infers more accurate ITE directions and has better reasoning performance. An example is given in Fig. 6.

 α - β -ITE: we combine recognition and reasoning performances. We multiply S_{ITE} with $P(\hat{\alpha}_p = \alpha_p)$ and $P(\hat{\beta}_q = \beta_q)$ as a unified metric S_{α - β -ITE.

For all metrics, we compute AP for each $[\alpha_p, \beta_q]$ and average them to mAP. Non-existing pairs are not considered.

5.2. Baselines

Different methods exploit different causal paths including the sub-graphs with $\alpha \rightarrow \beta$ or $\alpha \leftarrow \beta$ based on Fig. 4. We implement a series of baselines following different subgraphs to fully exert the potential of OCL and divide them into 3 folds w.r.t. $\alpha - \beta$ causal structure. We briefly list them here and detail them in the supplementary:

Fold I. No arc connecting α and β :

(1) Direct Mapping from f_I to P_{α} , P_{β} via an MLP (DM-V): feeding f_I into MLP-Sigmoids to predict P_{α} , P_{β} .

(2) DM Linguistic feature (DM-L): replacing the f_I of DM-V with linguistic feature f_L , which is the expectation of Bert [66] embeddings of category names w.r.t $P(O_i|I)$.

(3) Visual-Linguistic alignment, *i.e.*, Multi-Modality (MM): mapping f_I to a latent space and minimizing the distance to f_L , feeding it to an MLP-Sigmoids to get α , β .

(4) Linguistic Correlation of O- α , O- β (LingCorr): measuring the correlation between object and α or β classes via their Bert [66] embedding cosine similarities. P_{α} , P_{β} are given by multiplying P(O|I) to correlation matrices.

(5) Kernelized Probabilistic Matrix Factorization (KPMF) [67]: calculating feature similarity to all training samples as weights. Taking the weighted sum of GT α or β of training samples as predictions.

(6) **A**&**B** Lookup: getting P_A, P_B from M_A, M_B .

(7) Hierarchical Mapping (HMa): mapping f_I to category-level attribute or affordance space by an MLP, then feeding it to an MLP-Sigmoids to predict P_{α} or P_{β} .

Fold II. $\beta \rightarrow \alpha$:

(8) DM from β to α (DM- $\beta \rightarrow \alpha$): same as DM-V but using f_{β} to infer α .

(9) DM from β and I to α (DM- $\beta I \rightarrow \alpha$): same as DM-V but using both f_I and f_β to infer α .

Fold III. $\alpha \rightarrow \beta$:

(10) DM from α to β (DM- $\alpha \rightarrow \beta$): same as DM-V but using both f_I and f_{α} to infer β .

(11) DM from α and I to β (DM- $\alpha I \rightarrow \beta$): same as DM-V but using both f_I and f_{α} to infer β .

(12) Retrieving α - β relation by Ngram [68] (Ngram): adopting Ngram to retrieve the relevance of $\alpha \& \beta$. Then we use DM predicted α and the relevance to estimate β .

(13) Markov Logic Network [69] (MLN-GT): using GT α to infer β with MLN.

(14) Instantiation with attention (Attention): feeding $[f_{\alpha}, f_I]$ to an MLP-Sigmoid to generate attentions and predicting P_{β} by multiplying the attentions with P_B .

(15) DM with multi-head attention (DM-att): the α and β features are sent to multi-head attention to learn their interaction, then use MLP-Sigmoids to get predictions.

(16) Vanilla CLIP: CLIP [70] trained from scratch.

5.3. ITE loss

Though machines are expected to learn the causalities given α , β labels only. We wonder how it would perform given *causal supervision*. We adopt an extra Hinge loss to maximize the ITE score of all $[\alpha_p, \beta_q]$. In detail, we intend the ITE of causal relations larger than a margin τ (= 0.1 in experiments), so the loss term is:

$$\begin{cases} \max\{0, \tau - \Delta \hat{\beta}_q\}, & \beta_q = 1, \\ \max\{0, \tau + \Delta \hat{\beta}_q\}, & \beta_q = 0. \end{cases}$$
(11)

We enumerate all *annotated* $[\alpha_p, \beta_q]$ of an instance to obtain L_{ITE} . Different from the default, the total loss here is $L = \lambda_C L_C + L_I + \lambda_{ITE} L_{ITE}$.

5.4. Implementation Details

For a fair comparison, all methods adopt a shared COCO [2] pre-trained ResNet-50 [63] (frozen) to extract f_I and use the same object boxes in training and inference. In OCRN, the dimension of f_I and all f_{A_i} , f_{B_i} , f_{α} , f_{β} is 1024. The individual features of each attribute category are 512d and aggregated to 1024d by an FC. We train the attribute module with a learning rate of 0.3 and batch size of 1024 for 470 epochs. Then the attribute module is frozen, and the affordance module is trained with a learning rate of 3.0e-3 and batch size of 768 for 20 epochs. In training, $\lambda_C = 0.03$, $\lambda_{ITE} = 3$.

5.5. Results

Tab. 2 presents the results. We can find that the causal structure of the models matters in OCL. Comparing DM methods implementing different causal graphs (including $\alpha \rightarrow \beta, \alpha \leftarrow \beta$), α as intermediate knowledge (DM- $\alpha \rightarrow \beta$ and DM- $\alpha I \rightarrow \beta$) could advance β perception (DM-V). But when β serves as intermediate (DM- $\beta \rightarrow \alpha$ and DM- $\beta I \rightarrow \alpha$), β perception is comparable or even worse than DM-V. So the causal relation $\alpha \rightarrow \beta$ is more evident than $\beta \rightarrow \alpha$ in the realistic dataset, which supports our choice in Sec. 3.4 that we focus more on the $\alpha \rightarrow \beta$ arc and implement our model with only $\alpha \rightarrow \beta$.

OCRN outperforms the baselines and achieves decent improvements on all tracks. In terms of α recognition, with or without L_{ITE} , OCRN outperforms the second-best method with 1.7 and 2.5 mAP respectively. As for β recognition, the improvements are 0.7 and 1.1 mAP with or without L_{ITE} . Comparatively, HMa utilizes the supervision of A, B, but it performs much worse. A&B Lookup directly uses GT A, B to infer α , β , but its poor performance verifies the significant difference between A, B and α , β . Moreover, we find that all methods perform better on β than α , and the improvement of OCRN on α is larger too. This may be because α are more diverse than β , e.g., we can eat lots of foods, but foods usually have various attributes (fruit vs. pizza). And OCL also has fewer attribute classes than affordance classes (114 vs. 170). Another reason is that the positive samples in β labels (23.2%) are much more than the positives in α labels (9.4%). The different pos-neg ratio affects learning a lot and results in the above gap.

| Fold | Method | α | β | $\mathcal{S}_{	ext{ITE}}$ | $\mathcal{S}_{\alpha-\beta-\mathrm{ITE}}$ |
|--------------------------------|--|-------------|-------------|---------------------------|---|
| | DM-V | 29.9 | 51.8 | - | - |
| | DM-L | 21.2 | 47.5 | - | - |
| | MM | 23.8 | 48.9 | - | - |
| i N/A | LingCorr | 7.9 | 25.9 | - | - |
| | KPMF | 25.4 | 49.1 | - | - |
| | A&B-Lookup | 18.9 | 30.9 | - | - |
| | HMa | 28.6 | 51.7 | - | - |
| | DM-att | 21.9 | 49.2 | - | - |
| | Vanilla CLIP | 23.6 | 49.6 | - | - |
| | $DM-\beta \rightarrow \alpha$ | 30.0 | 52.0 | - | - |
| II: $\rho \rightarrow \alpha$ | $DM-\beta I \rightarrow \alpha$ | 29.5 | 51.8 | - | - |
| | $DM-\alpha \rightarrow \beta$ | 28.7 | 52.6 | 7.6 | 6.7 |
| | $DM-\alpha I \rightarrow \beta$ | 29.0 | 52.6 | 8.1 | 7.0 |
| iiii a NB | Ngram | 22.6 | 50.8 | <u>8.3</u> | 7.6 |
| III. $\alpha \rightarrow \rho$ | MLN-GT | - | 33.4 | 9.5 | <u>9.1</u> |
| | Attention | 24.1 | 48.9 | 8.1 | 7.1 |
| | OCRN | 31.6 | 53.3 | 9.5 | 9.2 |
| | $\text{DM-}\alpha \rightarrow \beta \text{ w/ } L_{ITE}$ | 28.8 | 52.4 | 15.5 | 14.0 |
| $\alpha \rightarrow \beta$ | $\text{DM-}\alpha I \rightarrow \beta \text{ w/ } L_{ITE}$ | <u>29.0</u> | <u>52.5</u> | 15.4 | 13.6 |
| | Ngram w/ L _{ITE} | 22.2 | 49.9 | 14.1 | 12.9 |
| | MLN-GT w/ L _{ITE} | - | 33.7 | 12.3 | 11.8 |
| | Attention w/ L_{ITE} | 23.9 | 49.0 | <u>17.8</u> | <u>15.5</u> |
| | OCRN w/ L _{ITE} | 31.5 | 53.6 | 20.3 | 16.9 |

Table 2: OCL results. w/ L_{ITE} means that training with ITE loss. The baselines in the upper block cannot operate ITE due to the model structure. Different α - β relations are exploited for causal graph comparison.

In ITE evaluation, without the guidance of L_{ITE} , all methods achieve unsatisfactory performances. However, OCRN still has an advantage. Only MLN-GT adopting the first-order logic and $GT \alpha$ labels is comparable with OCRN. If trained with L_{ITE} and direct causality labels, all methods perform much better to learn the causalities, *e.g.*, on OCRN, the ITE loss brings 10.8 and 7.7 mAP improvements on the two ITE tracks. Particularly, the typical deep learning model Attention performs best in baselines, but MLN-GT no longer holds the advantage. Relatively, OCRN shows more improvements and outperforms Attention with 2.5 and 1.4 mAP on the two ITE tracks.

We provide more visualizations and discussions in the supplementary. In particular, we also apply OCRN to **Human-Object Interaction Detection** [44], where OCRN boosts the performances of multiple HOI models and verifies the generalization and application potential of OCL.

5.6. Ablation Study

We verify the components of OCRN on the validation set in Tab. 3.

(1) **Deconfounding.** OCRN w/o deconfounding is implemented following Eq. 3 and 5, where P(O|I) and $P(O|I, \alpha)$ are the category predictions of pre-trained detectors [71]. All the α , β , and ITE performances drop due to the object bias. For more bias analyses please refer to the

| Method | α | β | $\mathcal{S}_{\mathrm{ITE}}$ | $\mathcal{S}_{\alpha-\beta-\mathrm{ITE}}$ |
|-------------------------|------|------|------------------------------|---|
| OCRN | 32.4 | 52.2 | 20.5 | 17.0 |
| w/o deconfounding | 32.1 | 51.8 | 18.2 | 16.1 |
| w/o L_{A_i}, L_{B_i} | 32.1 | 51.8 | 19.8 | 16.7 |
| w/o L_{lpha}, L_{eta} | 10.0 | 27.0 | 16.6 | 16.4 |
| 128 Dims | 31.7 | 51.5 | 18.0 | 16.0 |
| 512 Dims | 32.3 | 52.1 | 19.9 | 16.7 |
| 2048 Dims | 32.2 | 51.5 | 19.1 | 16.3 |
| Mean aggregation | 32.2 | 51.3 | 18.9 | 16.7 |
| Max-pooling aggregation | 32.1 | 49.1 | 19.0 | 16.8 |
| Random counterfactual | 32.4 | 51.8 | 5.1 | 5.1 |

Table 3: Ablation study results (validation set).

supplementary.

(2) Losses. The performances slightly drop after removing category-level L_{A_i}, L_{B_i} , but significantly drop without instance-level L_{α}, L_{β} by over 20 mAP.

(3) Feature dimension. We compare different dimentionality for feature $f_{A_i}, f_{B_i}, f_{\alpha}, f_{\beta}$. Smaller and larger feature sizes than 1024 all have degrading effects.

(4) ITE-related implementations. We probe some different methods: (a) Mean aggregation: $f'_{\alpha} = \sum_{i} f_{\alpha_{p}}$; (b) Max-pooling aggregation: f'_{α} is the max value of $f_{\alpha_{p}}$ as each component; (c) Random counterfactual feature: assigned random vector as the counterfactual attribute feature (instead of zero vector) during ITE. These methods perform worse than the chosen setting on ITE performance but are comparable on α and β performance.

5.7. Discussion

Overall, OCL poses extreme challenges to current AI systems. It expects representative learning to accurately recognize attributes and affordances from raw data meanwhile causal inference to capture the causalities within diverse instances and contexts, *i.e.*, both the *intuitive System 1 and logical System 2* [72]. From the experiments, we find that models struggle to achieve satisfying results on all tracks **simultaneously**. Notably, it is difficult to achieve a satisfying ITE score via data fitting. There is much room for improvement. For future studies, a harmonious performance on α , β , and causality learning are encouraged to better capture object knowledge. Potential directions may include causal representation learning [73], neuralsymbolic reasoning [74], and Foundation Models [75]. etc.

6. Conclusion

In this work, we introduce object concept learning (OCL) expecting machines to infer affordances and explain what attributes enable an object to possess them. Accordingly, we build an extensive dataset and present OCRN based on casual intervention and instantiation. OCRN achieves decent performance and follows the causalities well. However, OCL remains challenging and would inspire a line of studies on reasoning-based object understanding.

Acknowledgment: Supported by the National Key R&D Program of China (No.2021ZD0110704), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Shanghai Qi Zhi Institute, Shanghai Science and Technology Commission (21511101200).

References

- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1, 3, 18
- [2] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 2, 3, 8, 18, 23
- [3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1
- [4] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1, 5
- [5] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *CVPR*, 2015. 1, 2, 13
- [6] James J Gibson. The ecological approach to the visual perception of pictures. *Leonardo*, 11(3):227–235, 1978. 1, 2
- [7] Alex Martin. The representation of object concepts in the brain. Annu. Rev. Psychol., 58:25–45, 2007. 1
- [8] Stevan Harnad. The symbol grounding problem. *Physica* D: Nonlinear Phenomena, 42(1-3):335–346, 1990.
- [9] B Ross. Category learning: Learning to access and use relevant knowledge. *Memory and mind: A Festschrift for Gordon H. Bower*, pages 229–246, 2008. 1
- [10] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 1, 2
- [11] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 1, 2, 3, 13, 18
- [12] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth.
 Describing objects by their attributes. In *CVPR*, 2009. 1, 2, 3, 13, 18
- [13] Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *ICRA*, 2018. 2
- [14] Tucker Hermans, James M Rehg, and Aaron Bobick. Affordance prediction via learned object attributes. In *ICRA Workshop*, 2011. 2, 3, 5
- [15] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. In CVPR, 2018. 2
- [16] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. Reasoning about object affordances in a knowledge base representation. In ECCV, 2014. 2, 3, 5, 13, 22, 29

- [17] Anh Nguyen, Dimitrios Kanoulas, Darwin G Caldwell, and Nikos G Tsagarakis. Object-based affordances detection with convolutional neural networks and dense conditional random fields. In *IROS*, 2017. 2, 3, 5, 13
- [18] David F Fouhey, Xiaolong Wang, and Abhinav Gupta. In defense of the direct perception of affordances. arXiv preprint arXiv:1505.01085, 2015. 2, 5
- [19] Yu-Wei Chao, Zhan Wang, Rada Mihalcea, and Jia Deng. Mining semantic affordances of visual object categories. In *CVPR*, 2015. 2, 3, 13
- [20] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. Causal inference in statistics: A primer. John Wiley & Sons, 2016. 2, 3, 6, 15, 28, 29
- [21] Krishna Kumar Singh, Santosh Divvala, Ali Farhadi, and Yong Jae Lee. Dock: Detecting objects by transferring common-sense knowledge. In ECCV, 2018. 2
- [22] H. Liu, R. Wang, S. Shan, and X. Chen. Learning multifunctional binary codes for both category and attribute oriented retrieval tasks. In *CVPR*, 2017. 2, 3, 13
- [23] Genevieve Patterson and James Hays. Coco attributes: Attributes for people, animals, and objects. In ECCV, 2016. 2, 3, 13
- [24] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2016. 2, 3, 13
- [25] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In CVPR, 2019. 2
- [26] Devi Parikh and Kristen Grauman. Relative attributes. In ICCV, 2011. 2
- [27] Sung Ju Hwang, Fei Sha, and Kristen Grauman. Sharing features between objects and their attributes. In CVPR, 2011. 2
- [28] Chao-Yeh Chen and Kristen Grauman. Inferring analogous attributes. In CVPR, 2014. 2
- [29] Dhruv Mahajan, Sundararajan Sellamanickam, and Vinod Nair. A joint learning framework for attribute models and object descriptions. In *ICCV*, 2011. 2
- [30] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In CVPR, 2017. 2
- [31] Tushar Nagarajan and Kristen Grauman. Attributes as operators: factorizing unseen attribute-object compositions. In ECCV, 2018. 2
- [32] Yong-Lu Li, Yue Xu, Xiaohan Mao, and Cewu Lu. Symmetry and group in attribute-object compositions. *CVPR*, 2020. 2
- [33] Yong-Lu Li, Yue Xu, Xinyu Xu, Xiaohan Mao, and Cewu Lu. Learning single/multi-attribute of object with symmetry and group. *TPAMI*, 2021. 2
- [34] Yibiao Zhao and Song-Chun Zhu. Scene parsing by integrating function, geometry and appearance models. In *CVPR*, 2013. 2

- [35] Xiaolong Wang, Rohit Girdhar, and Abhinav Gupta. Binge watching: Scaling affordance learning from sitcoms. In *CVPR*, 2017. 2
- [36] Anirban Roy and Sinisa Todorovic. A multi-scale cnn for affordance segmentation in rgb images. In ECCV, 2016. 2
- [37] Abhinav Gupta and Larry S Davis. Objects in action: An approach for combining action understanding and object perception. In CVPR, 2007. 2
- [38] Bangpeng Yao, Jiayuan Ma, and Li Fei-Fei. Discovering object functionality. In *ICCV*, 2013. 2
- [39] Keizo Kato, Yin Li, and Abhinav Gupta. Compositional learning for human object interaction. In ECCV, 2018. 2
- [40] Lerrel Pinto and Abhinav Gupta. Supersizing selfsupervision: Learning to grasp from 50k tries and 700 robot hours. In *ICRA*, 2016. 2
- [41] Spyridon Thermos, Georgios Th Papadopoulos, Petros Daras, and Gerasimos Potamianos. Deep affordance-grounded sensorimotor object recognition. In *CVPR*, 2017.
 2
- [42] Lerrel Pinto, Dhiraj Gandhi, Yuanfeng Han, Yong-Lae Park, and Abhinav Gupta. The curious robot: Learning visual representations via physical interactions. In *ECCV*, 2016. 2
- [43] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *TPAMI*, 2023. 2
- [44] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In WACV, 2018. 2, 3, 9, 13, 25, 26, 27
- [45] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Shiyi Wang, Hao-Shu Fang, Ze Ma, Mingyang Chen, and Cewu Lu. Pastanet: Toward human activity knowledge engine. In *CVPR*, 2020. 2
- [46] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search.* MIT press, 2000. 2
- [47] Isabelle Guyon, Constantin Aliferis, and André Elisseeff. Causal feature selection. *Computational methods of feature selection*, pages 63–82, 2007. 2
- [48] Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Visual causal feature learning. arXiv preprint arXiv:1412.2309, 2014. 2
- [49] Lyndsey C Pickup, Zheng Pan, Donglai Wei, YiChang Shih, Changshui Zhang, Andrew Zisserman, Bernhard Scholkopf, and William T Freeman. Seeing the arrow of time. In CVPR, 2014. 2
- [50] Karel Lebeda, Simon Hadfield, and Richard Bowden. Exploring causal relationships in visual object tracking. In *ICCV*, 2015. 2
- [51] Suraj Nair, Yuke Zhu, Silvio Savarese, and Li Fei-Fei. Causal induction from visual observations for goal directed tasks. arXiv preprint arXiv:1910.01751, 2019. 2
- [52] Ishita Dasgupta, Jane Wang, Silvia Chiappa, Jovana Mitrovic, Pedro Ortega, David Raposo, Edward Hughes, Peter Battaglia, Matthew Botvinick, and Zeb Kurth-Nelson. Causal reasoning from meta-reinforcement learning. arXiv preprint arXiv:1901.08162, 2019. 2

- [53] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *CVPR*, 2020. 2, 6, 20, 28, 29
- [54] Yuval Atzmon, Felix Kreuk, Uri Shalit, and Gal Chechik. A causal view of compositional zero-shot recognition. arXiv preprint arXiv:2006.14610, 2020. 2
- [55] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, 2018. 3, 13
- [56] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. arXiv preprint arXiv:1505.04474, 2015. 3, 13, 25
- [57] Christiane Fellbaum. Wordnet. *The encyclopedia of applied linguistics*, 2012. **3**, **13**
- [58] Daniel N Osherson, Joshua Stern, Ormond Wilkie, Michael Stob, and Edward E Smith. Default probability. *Cognitive Science*, 15(2):251–269, 1991. 3, 13
- [59] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 5
- [60] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005. 5, 7, 19
- [61] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 5
- [62] Yue Xu, Yong-Lu Li, Jiefeng Li, and Cewu Lu. Constructing balance from imbalance for long-tailed image recognition. In ECCV, 2022. 5
- [63] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016. 6, 8
- [64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017. 6
- [65] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *CVPR*, 2020. 7, 20, 22, 28, 29
- [66] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018. 8, 20, 21
- [67] Tinghui Zhou, Hanhuai Shan, Arindam Banerjee, and Guillermo Sapiro. Kernelized probabilistic matrix factorization: Exploiting graphs and side information. In SDM, 2012. 8, 21
- [68] Y. Lin, J.-B. Michel, E. L. Aiden, J. Orwant, W. Brockman, and S. Petrov. Syntactic annotations for the google books ngram corpus. In ACL, 2012. 8, 21
- [69] Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine learning*, 62(1-2):107–136, 2006. 8, 22

- [70] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 8
- [71] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030, 2021. 9, 23, 27
- [72] Yoshua Bengio. From system 1 deep learning to system 2 deep learning. In *Posner lecture at NeurIPS*'2019, 2019. 9
- [73] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. 9
- [74] Tarek R Besold, Artur d'Avila Garcez, Sebastian Bader, Howard Bowman, Pedro Domingos, Pascal Hitzler, Kai-Uwe Kühnberger, Luis C Lamb, Daniel Lowd, Priscila Machado Vieira Lima, et al. Neural-symbolic learning and reasoning: A survey and interpretation. arXiv preprint arXiv:1711.03902, 2017. 9
- [75] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155, 2022. 9
- [76] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *TIP*, 2019. 16
- [77] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. JMLR, 2008. 17
- [78] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. arXiv preprint arXiv:2009.12991, 2020. 20, 28, 29
- [79] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. 20
- [80] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013. 21
- [81] Yong-Lu Li, Xinpeng Liu, Han Lu, Shiyi Wang, Junqi Liu, Jiefeng Li, and Cewu Lu. Detailed 2d-3d joint representation for human-object interaction. In CVPR, 2020. 24
- [82] Xinpeng Liu, Yong-Lu Li, and Cewu Lu. Highlighting object category immunity for the generalization of humanobject interaction detection. In AAAI 2022, 2022. 24
- [83] Yong-Lu Li, Xiaoqian Wu, Xinpeng Liu, Yiming Dou, Yikun Ji, Junyi Zhang, Yixing Li, Jingru Tan, Xudong Lu, and Cewu Lu. From isolated islands to pangea: Unifying semantic space for human action understanding. arXiv preprint arXiv:2304.00553, 2023. 24
- [84] Xinpeng Liu, Yong-Lu Li, Xiaoqian Wu, Yu-Wing Tai, Cewu Lu, and Chi-Keung Tang. Interactiveness field in human-object interactions. In CVPR, 2022. 25
- [85] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Xijie Huang, Liang Xu, and Cewu Lu. Transferable interactiveness

knowledge for human-object interaction detection. In *TPAMI*, 2022. 25

- [86] Xiaoqian Wu, Yong-Lu Li, Xinpeng Liu, Junyi Zhang, Yuzhe Wu, and Cewu Lu. Mining cross-person cues for body-part interactiveness learning in hoi detection. In ECCV, 2022. 25
- [87] Yong-Lu Li, Hongwei Fan, Zuoyu Qiu, Yiming Dou, Liang Xu, Hao-Shu Fang, Peiyang Guo, Haisheng Su, Dongliang Wang, Wei Wu, and Cewu Lu. Discovering a variety of objects in spatio-temporal human-object interactions. arXiv preprint arXiv:2211.07501, 2022. 25
- [88] Xinyu Xu, Yong-Lu Li, and Cewu Lu. Dynamic context removal: A general training strategy for robust models on video action predictive tasks. *IJCV*, 2023. 25
- [89] Xinyu Xu, Yong-Lu Li, and Cewu Lu. Learning to anticipate future with dynamic context removal. In *CVPR*, 2022. 25
- [90] Hao-Shu Fang, Yichen Xie, Dian Shao, Yong-Lu Li, and Cewu Lu. Decaug: Augmenting hoi detection via decomposition. In AAAI, 2021. 25
- [91] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Mingyang Chen, Ze Ma, Shiyi Wang, Hao-Shu Fang, and Cewu Lu. Hake: Human activity knowledge engine. arXiv preprint arXiv:1904.06539, 2019. 25
- [92] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, Zuoyu Qiu, Liang Xu, Yue Xu, Hao-Shu Fang, and Cewu Lu. Hake: A knowledge engine foundation for human activity understanding. arXiv preprint arXiv:2202.06851, 2022. 25
- [93] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In *CVPR*, 2019. 25, 27
- [94] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instancecentric attention network for human-object interaction detection. In *BMVC*, 2018. 25, 27
- [95] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, and Cewu Lu. Hoi analysis: Integrating and decomposing human-object interaction. In *NeurIPS*, 2020. 25, 27
- [96] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In CVPR, 2020. 28
- [97] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. arXiv preprint arXiv:1707.09457, 2017. 28
- [98] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018. 28, 29
- [99] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision, pages 2980–2988, 2017. 28, 29
- [100] Patricia W Cheng and Laura R Novick. Causes versus enabling conditions. *Cognition*, 40(1-2):83–120, 1991. 29

We report more details and analyses here:

- Sec. A: Category/Attribute/Affordance Selection
- Sec. B: Annotation Details
- Sec. C: Causal Graph
- Sec. D: OCL Characteristics
- Sec. E: ITE Metric Details
- Sec. F: Baseline Details
- Sec. G: Detailed Result Analysis
- Sec. H: Application on HOI Detection
- Sec. I: Comparison on Debiasing
- Sec. J: Discussion about States
- Sec. K: Discussion about Causality and Causal Graph
- Sec. L: Detailed Lists

A. Category/Attribute/Affordance Selection

We choose affordances, categories, and attributes, considering their causal relations. Their word clouds are shown in Fig. 7. The complete lists can be found in Suppl. Sec. L.

(1) Affordance: To build a general and applicable knowledge base, we collect 1,006 affordance candidates from several widely-used action/affordance datasets: 957 from [19], 160 from [55], 146 from [44], 97 from [56], 41 from [16], 21 from [17] (with *overlaps*). We find that not all affordances are in common use and some of them are difficult for visual recognition, *e.g.*, accept (consider right and proper). So each candidate is scored by 5 human experts from 0.0 to 5.0 according to generality and commonness. We keep **170** top-scored affordances in our base (134 from [19], 78 from [55], 127 from [44], 53 from [56], 13 from [16], 11 from [17], with *overlaps*).

(2) **Category**: Considering the taxonomy (Word-Net [57]), we collect a pool with over 1,742 object categories from previous datasets: 32 from [12], 28 from [23], 717 from [11], 1,000 from [22] (with *overlaps*). Then we merge the similar categories according to WordNet [57] and filter out the categories which are not common daily objects (man, planet), unrelated to the above 170 affordances (skyscraper) or too uncommon (malleefowl). Finally, our database has **381** common object categories. These object categories are divided into **12** super categories, shown in Fig. **8**.

(3) Attribute: We extract the attributes from several large-scale attribute datasets: 64 from [12], 203 from [23], 66 from [11], 25 from [22], top 500 from [24]), and manually filter the 500 most frequent attributes. Five experts give 0 to 5 scores based on their relevance to human actions and the selected 170 affordances to better explore the causal relations between attributes and affordances. Some attributes (cloudy, competitive) that are not useful for affordance reasoning are discarded. Finally, **114** attributes are kept, covering colors, deformations, supercategories, surface, geometrical, and physical properties.

B. Annotation Details

B.1. Attribute Annotation

(1) **Category-level attribute** (*A*). Following [58], to avoid bias, annotators are given *category-attribute pairs* (category *names*, not images). They propose a 0-3 score according to the category concept in their minds (0: No, 1: Normally No, 2: Normally Yes, 3: Yes). Each pair is annotated by three annotators and takes the plurality as the *A* label. If the range of 3 proposals exceeds 1, another three annotators will re-annotate this pair until achieving consensus. We binarize the annotations (0: No, 1: Yes) with a threshold of 2 and get a category-level attribute matrix M_A ([381, 114]).

(2) **Instance-level attribute** (α). Two annotators label each pair with 0 (No) and 1 (Yes). If they give different labels, this pair will be handed over to another two annotators until meeting consensus.

B.2. Affordance Annotation

(1) **Category-level affordance** *B*. Following [19], the annotators are given category-affordance pairs. The pairs are annotated in four bins (0-3) and normalized (same as *A*) to describe the possibility of an affordance in a category. Each pair is annotated by three annotators and makes consensus the same as *A*. The 0-3 scores are binarized (1: Yes, 0: No) with a threshold of 2. The final category-level affordance matrix M_B is [381, 170].

(2) **Instance-level affordance** β is annotated for every instance with the help of *object states* [5]. As B is determined by common states, objects in specific states may have different affordances from B, e.g., we cannot board a flying plane. As the instances in the same state should have similar β (all rotten apples cannot be eaten), six experts first conclude the states. The experts scan all instances of each category and use their knowledge of affordance to define all the existing states. Then all 186 K instances are dispatched to the concluded states via crowdsourcing. If some instances do not belong to any predefined states, they will be returned to the experts to add more states. In total, 1,376 states are defined, and each category has 3.6 states on average. Next, β is annotated for each state. Given a state-affordance pair and example images, two annotators mark it with 0 (No) and 1 (Yes). The results are combined in the same way as α . Thus, each instance would have a state and the corresponding β . An annotator would recheck each instance together with its state and β labels to ensure the quality. If its state is inaccurate or the state β labels are unsuitable, this annotator would correct them.



(a) Category (b) Attribute (c) Affordance Figure 7: Word clouds of object categories, attributes, and affordance (by positive frequencies in OCL).



Figure 8: Super-categories of objects in OCL.

B.3. Causal Relation Annotation

(1) Filtering. As exhaustive annotation is arduous, we only annotated existing rules without ambiguity. Starting from the [114,170] matrix of α - β classes, we ask three experts to vote on the causal relation of each class. They scan all instances to answer whether the relationship exists in any case. That is, we just annotate the *least* pairs with the largest possibility to be casually related. Some causal pairs may be excluded. In detail, for each of the $114 \times 170 \ \alpha \beta$ pairs, we attach 10 samples for reference and 3 experts vote yes/no/not sure. We take the majority vote and the not sure and controversial pairs are rechecked. The not sure and no pairs are removed, and so do the ambiguous pairs. The pairs we selected are checked carefully to ensure the causalities and we only evaluate models on them. Thus, the missed causal pairs or non-causal pairs would not affect the results. Finally, we obtain about 10% α - β classes as candidates. The left 90% pairs may hold value and we will mine new rules with LLMs in future work, especially from ambiguous pairs.

(2) Instance-level causality: we also adopt object states as a reference. For each *state*- α - β triplet, two annotators are asked whether the specific attribute is the *direct* and *unam*-

biguous cause of this affordance in this state and gives their binary answer. We use the same method in annotating β to combine results and assign *state-level* labels to instances. Next, for all instances of a state, an expert decides whether the state-level relations are reasonable for each *instance* in specific contexts and correct the inaccurate ones. Finally, we obtain about 2 M *instance-* α - β triplets of causal relations.

B.4. A Running Example of Dataset Construction.

A running example is shown in Fig. 9 to show the process of annotations clearly.

C. Causal Graph

In this section, we first briefly introduce the causal graph model and causal intervention. Then we introduce the details of the causal graph our knowledge base can support. Then, we detail the implementation of the causal graphs used by different methods.

C.1. Basics of Causal Inference and Causal Graph

A causal graph is a DAG that describes the causal relations between multiple factors. Each directed edge points





Figure 10: A more complex causal graph of our knowledge base. A, B, O are the object category and category-level attribute and affordance. I is the object appearance, α, β are the instance-level attribute and affordance. Note that "or" indicates that the arcs between $A, B, \alpha, \beta, I, \alpha$, and I, β indicate that either $A \leftarrow B$ or $A \rightarrow B$ (the others are similar) is considering in the setting.



Figure 11: An example of the causal graph and causal intervention. We study the causal relation $X \to Y$ while confounder Z exists and brings bias. After the intervention on variable X, the poisonous relation $Z \to X$ is eliminated.

from the "cause" to its "effect", *e.g.* in Fig. 11, node X is the cause of node Y. Under the scenario that causal variables and causal graphs are known, **causal inference** studies how to infer the strength of causal edges given observations, or infer the outcomes given some of the causal variable values.

However, the causal relation in the real world is sophisticated. The causal relation that we observed may have been polluted by spurious variables. For example, let X in Fig. 11 be ice cream sales and Y be drownings, one may observe that more ice cream sales lead to more drownings and infer that they are causally related. Actually, the observed relation is due to another factor Z: weather temperature. These variables are called **confounders**, which is the common cause of two causal variables that we are studying, *e.g.* in the left graph in Fig. 11, Z is a confounder when we focus on the causal edge $X \rightarrow Y$.

In causal inference, confounders should be eliminated to avoid biases on causal learning, by applying **intervention** on the cause variables (*e.g.* X in our example) to "control" its distribution to block the effect of confounder. Traditional scientific research on causality adopts Randomized Controlled Trial (RCT) to completely remove the confounder, but it is not applicable when we only have observational data. Pearl. [20] *et al.* propose *do-calculus* to systematically analyze the causal graph and alleviate the confounder bias in a probabilistic view. In the simple case in Fig. 11, the confounder Z can be eliminated with **Back-door Adjustment**:

$$P(Y|do(X)) = \sum_{z} P(Y|X, Z = z)P(Z = z), \quad (12)$$

where z is the specific value of the random variable Z. The causal graph of our OCRN also meets the back-door criterion so we apply the back-door adjustment to alleviate bias from the confounder O.

C.2. Causal Graph of Our Knowledge Base

A more complicated causal graph considering more arcs between nodes is shown in Fig. 10. The causal relations between nodes or arcs in Fig. 10 are determined as follows:

Firstly, we introduce two kinds of special arcs. $Q \rightarrow A Q \rightarrow B$ (dotted arcs): in OCL A and

 $O \rightarrow A, O \rightarrow B$ (dotted arcs): in OCL, A and B are defined as the category-level annotations. Given O, A, and

B are strictly determined. In Fig. 10, we use two dotted arrows from O to A, B respectively to indicate this deterministic relation to distinguish them from the other causal relations.

 $O \rightarrow I, A \rightarrow \alpha, B \rightarrow \beta$ (blue arcs): we see the category-level O, A, and B are direct causes of instancelevel I, α , and β during the concept *instantiation* according to OCL definition. Because the visual representation I and properties α, β of an instance are derived from the conceptlevel categorical ones. The reversed arcs $O \leftarrow I, A \leftarrow \alpha, B \leftarrow \beta$ mean that O, A, B are the *aggregations* of instances and would be marginally affected by one specific instance, thus we do not include these arcs here for clarity.

Next, we illustrate the regular causal arcs as follows.

 $I \rightarrow \alpha$, $I \rightarrow \beta$: the recognition process of α and β . As *I* indicates the *physical noumenon*, it is the source of semantic and functional properties and decides/causes α, β .

 $\alpha \rightarrow I, \beta \rightarrow I$: the generation of visual pattern from attribute or affordance descriptions and can be utilized in image generation/manipulation tasks [76].

 $A \leftarrow B \text{ or } A \rightarrow B, \alpha \leftarrow \beta \text{ or } \alpha \rightarrow \beta$: the causal direction between attribute and affordance can be reversed sometimes. The arc from α to β is evident, *e.g.*, a broken cup is not useable. Sometimes, the reverse arc causal effect from β to α also exists, *e.g.*, an eatable banana would not be unripe.

C.3. Causal Graph Implementation

In this work, we mainly study the recognition and reasoning of attribute and affordance for robotics and embodied AI, hence we remove the two arcs corresponding to image generation $\alpha \rightarrow I$, $\beta \rightarrow I$. Due to the deterministic relation between O, A, and B, we can simplify the three nodes to a single node O' (Fig. 12).

Different *methods* can exploit different causal paths. We propose diverse baselines to implement different causal subgraphs, including the subgraphs with $\alpha \rightarrow \beta$, and $\alpha \leftarrow \beta$. The causal graphs of some baselines are shown in Fig. 13.

The ablation experiment with arc $\alpha \rightarrow \beta$ and $\alpha \leftarrow \beta$ shows that the causal effect of $\alpha \rightarrow \beta$ is stronger than the alternative in our datasets. Besides, from the aspect of embodied AI and robotics, affordance is more important in practical applications like object manipulation, so we focus more on affordance recognition and regard β inference as our main goal. Therefore, in OCRN and some other baselines, we keep the arc $\alpha \rightarrow \beta$. And in causal reasoning, we focus on the evaluation of $\alpha \rightarrow \beta$ too. The causal graph of OCRN is shown in Fig. 14.

D. OCL Characteristics

D.1. Object Box Size

We visualize the distribution of normalized object box size in Fig. 15, where the box width and height are normalized by the width and height of the whole image. It shows that most objects in our knowledge base are *small objects*, providing abundant regional information.

D.2. Annotator Information

Annotators' age, major, and education degree are presented in Fig. 16, 17, and 18.

D.3. Matrix Samples

The category-level attribute and affordance (A, B) matrices are detailed in Fig. 19, 20 as heatmaps, and the cells with dark color indicate positive samples. For example, ice cream is cold while clock is not natural, cake can be eaten while eraser can not be cooked. These are in line with our common sense.

D.4. State Distribution

Before annotating the affordances, we first define the object states for all object categories and annotate the state affordances. In total, we define 1,376 states for 381 object categories. And Fig. 21 shows the state distribution per object category.

D.5. Attribute-Affordance Relation

We analyze the instance-level attribute-affordance relations in our knowledge base under three criteria. (1) Attribute Conditioned Affordance Probability. It is computed as $P(\beta|\alpha)$ to estimate affordance probability given an attribute. The range is [0,1]. (2) Attribute-Affordance **Correlation.** For all instances in our dataset, we evaluate the label correlation of each attribute-affordance pair, whose scale is in [-1,1]. (3) Attribute-Affordance Causality. Starting with the annotated cause-effect $(\alpha - \beta)$ labels, we count for how many times each attribute-affordance pair appear in our dataset and normalize the value by the maximum occurrences, leading to a value in the range [0,1]. It should be mentioned that we only annotate whether an attribute-affordance pair has explicit and key causality, but the detailed effect (positive or negative) should be referred to instance labels.

We visualize the samples of attribute-affordance relation matrices in Fig. 22, 23, 24 and observe some interesting properties of them. They reveal some common relations, such as what is between *tasty* and *eat*. However, some of the criteria suffer from data bias. For the condition matrix in Fig. 22, it only cares about cases with *positive* attribute labels, which is not good in highlighting the negative relations, *e.g.*, the relation between *natural* and *produce*. For



Figure 12: Simplified causal graph for OCL task. Note that "or" indicates that the arcs between A, B and α , β are either $A \leftarrow B$ or $A \rightarrow B$ ($\alpha \leftarrow \beta$ or $\alpha \rightarrow \beta$), instead of concurrence.



Figure 13: Causal graphs of the baselines.

the former two matrices in Fig. 22, 23, they all point out the relation between *tasty* and *pick*, since most *tasty* objects are *pickable food*. This finding is simply misled by the data bias but violates the causal graph (inference from attribute to object category, then affordance). Last, the matrix obtained from our causal annotation in Fig. 24 is more sparse and clear of causality.

D.6. Unified Object Representation

To compare the difference between attribute-only and attribute-affordance representations, we cluster the object instances of two similar animals (zebra and horse) with their attribute labels and attribute-affordance labels, respectively. The results are shown in Fig. 25 via t-SNE [77]. With both attribute and affordance labels, zebra and horse can be better separated than attribute only. And attribute and affordance together can differentiate specific **states** well, such as riding, pulling car, etc.

D.7. Difference between Category- and Instance-Level Labels

We analyze the differences between category-level A, B labels and instance-level α, β labels. For each object category, we compute the *average ratio* of changed at-



Figure 14: Causal graph of OCRN.



Figure 15: Distribution of normalized object box width (left) and height (right).

tribute/affordance classes during each instantiation from A to α or from B to β . The top 50 categories with the most significant differences between A and α as well as B and β are reported respectively in Fig. 26. We find that affordance labels change more dramatically than attribute labels during instantiations. This is because **each** attribute change may affect **several** affordances, *e.g.*, when a common book becomes burning, we can neither open nor read it.

D.8. Attribute-Affordance Causal Relations

We annotate all object instances' causal relations of filtered $[\alpha_p, \beta_q]$ pairs. In total, 1,085 $[\alpha_p, \beta_q]$ pairs are chosen for the causality annotation, and over 2 M *instance*- α - β triplets are annotated. In the ITE evaluation (main text Sec. 5), we report the mean AP of top-300 [α_p , β_q] pairs to avoid the biased influence of very rare [α_p , β_q] pairs that include less than 35 object instances.

D.9. Data Partitioning

For the OCL task, our knowledge base is split into the train, val, and test sets. The statistical details of the split are listed in Tab. 4. The image number ratio of the three sets is nearly 4:1:0.6, and the instance ratio is around 5:1:1.

| Set | Image | Object Instance | Object category |
|----------|--------|-----------------|-----------------|
| Train | 56,916 | 135,148 | 381 |
| Val | 14,446 | 25,176 | 221 |
| Test | 9,101 | 25,617 | 221 |
| Val+Test | 23,547 | 50,793 | 221 |
| All | 80,463 | 185,941 | 381 |

Table 4: Detailed data split of our knowledge base.

D.10. Images and Instances

Some additional data samples of our knowledge base are shown in Fig. 27, 28a, 28b, 29, 30, and 31, including samples of diverse object categories with various bounding box distributions, different attributes and affordances, and human-labeled object states and obvious causal relations. We also show the counts of object categories, attributes, and affordances in instance/image in Fig. 32, 33, and 34.

D.11. More Statistics of Annotation

We divide A, B, α, β , causality annotation into multiple finer-grained small sets in our pipeline. Generally, we have 13, 19, 124, 140, and 85 annotator sets (381 total) for A, B, α, β , and causality annotation respectively. We assign each small set to 2 annotators. However, considering the controversial situations introduced, part of the annotation are confused cases based on their results. In the whole process, 9.6% of A, 7.7% of B, 5.2% of $\alpha, 7.9\%$ of β , and 13.7% of causality are confusing and re-assigned to additional annotators. These indeterminable ones will be sent to two extra annotators until agreement. The quality of the dataset is guaranteed by a low confusion ratio and multiple refining stages.

D.12. Potential Bias

We have considered the bias issue in the construction of our dataset. (1) In our dataset, the existing datasets (ImageNet [1], COCO [2], aPY [12], SUN [11]) are opensourced datasets and the images collected from the Internet are publicly accessible too. The dataset is constructed for only non-commercial purposes. We will only provide the





Figure 16: Age information of annotators.



Figure 17: Major information of annotators.

Figure 18: Degree information of annotators.



Figure 19: Category-level attribute (A) matrix.

Figure 20: Category-level affordance (B) matrix.

URLs of these images to avoid copyright infringement. (2) During image collection, we choose images with general objects and are particularly careful with the image selection to avoid unsuitable content, private images, or implicit biases. (3) During annotation, the annotators cover different genders, ages, and fields of expertise to avoid potential annotation biases. And they are all informed on how we will use the annotations in our research.

E. ITE Metric Details

ITE (**Individual Treatment Effect (ITE)** [60]) is to measure whether a model infers affordance with proper attention to the causality-related attribute. That said, when removing the attribute, the model is expected to have *large prediction difference further away from the ground truth*.

We detail some settings in our ITE metric. For the ITE score:

$$S_{\text{ITE}} = \begin{cases} \max(\Delta \hat{\beta}_q, 0), & \beta_q = 1, \\ \max(-\Delta \hat{\beta}_q, 0), & \beta_q = 0, \end{cases}$$
(13)

where

$$\Delta \hat{\beta}_q = \hat{\beta}_q |_{do(\alpha_p)} - \hat{\beta}_q |_{do(\mathbf{x}_q)} == \hat{\beta}_q - \hat{\beta}_q |_{do(\mathbf{x}_q)}, \quad (14)$$

we want the moving direction of affordance prediction after the intervention to be correct according to the GT affordance labels (β_q). Concretely, for an instance with the labeled causal relation between [α_p , β_q], if the label $\beta_q = 1$, we expect the prediction change $\Delta \hat{\beta}_q$ to be larger, indicating the elimination of α_p leads to a drop of predicted probability. Because without the effect of α_p , the probability of β_q should be **contrary** to the fact ($\beta_q = 1$). Similarly, if $\beta_q = 0$, we expect $\Delta \hat{\beta}_q$ to be smaller, i.e. the elimination of α_p leads to an increase of predicted probability. The design of the ITE loss also follows the setting of this ITE score.

In α - β -ITE, the ITE score is multiplied by two factors of recognition performance:



Figure 21: State distributions of different object categories.



Figure 22: Attribute **conditioned** affordance matrix.





Figure 23: Attribute-affordance **correlation**.

Figure 24: Attribute-affordance causality.

$$P(\hat{\alpha_p} = \alpha_p) = \begin{cases} \hat{\alpha_p}, & \alpha_p = 1, \\ 1 - \hat{\alpha_p}, & \alpha_p = 0, \end{cases}$$

$$P(\hat{\beta_q} = \beta_q) = \begin{cases} \hat{\beta_q}, & \beta_q = 1, \\ 1 - \hat{\beta_q}, & \beta_q = 0. \end{cases}$$
(15)

And the overall metric is:

$$S_{\alpha-\beta-\text{ITE}} = S_{\text{ITE}} P(\hat{\alpha_p} = \alpha_p) P(\hat{\beta_q} = \beta_q)$$
(16)

The factors measure the correctness of attributes and affordances. Hence a model achieves a high $S_{\alpha-\beta-\text{ITE}}$ only if it correctly predicts attribute and affordance and learns the causal relation between them.

In our experiments, for attribute/affordance recognition only, all methods adopt labels to learn knowledge from the data. In the evaluation of causal relation, only the "w/ L_{ITE} " models adopt the causal relation labels. We hope the models can automatically learn to mine and learn the intrinsic causalities. Thus, we design the ITE to evaluate this ability. Similar to our OCRN, some works [53, 78, 65] also try to marry supervised deep learning and causal inference.

F. Baseline Details

We introduce the details of all baselines here:

Fold I. No arc between α and β .

(1) Direct Mapping from Visual Feature (DM-V): feeding f_I into MLP-Sigmoids to predict P_{α}, P_{β} . Each α and β class owns customized MLP followed by Layer-Norm [79] to generate class-specific features and share the same MLP-Sigmoid in classification.

(2) DM from Linguistic Representation (DM-L): replacing the input representation f_I of DM-V with linguistic feature f_L , which is the expectation of Bert [66] of category names w.r.t $P(O_i|I)$.

(3) Multi-Modality (MM): mapping f_I to the semantic space via minimizing the distance to its f_L . The multimodal aligned f_I is fed to an MLP-Sigmoids to predict P_{α}, P_{β} .



Figure 25: Clustering using attribute and attribute-affordance labels.

(4) Linguistic Correlation (LingCorr): measuring the correlation between object and α/β classes via their Bert [66] cosine similarity. P_{α} , P_{β} are given by multiplying P(O|I) to correlation matrices.

(5) Kernelized Probabilistic Matrix Factorization (KPMF) [67]: calculating the Softmax normalized cosine similarity between each testing instance and all training samples as weights. Then P_{α} or P_{β} is generated as the weighted sum of GT α or β of training samples.

(6) A&B Lookup: returning the expectation of category-level attribute or affordance vectors A_i, B_i w.r.t $P(O_i|I)$. In detail, seen category probabilities are obtained from GT prior M_A, M_B . Unseen category probabilities are voted by the top 3 most similar seen categories according to the cosine similarity of category Word2Vec [80] vectors. Then, we generate category-level attribute and affordance matrices M'_A, M'_B given the GT prior (seen) and similarity-based probabilities (unseen). Finally, we multiply P(O|I) with M'_A, M'_B to predict P_A, P_B and assign them to P_α, P_β respectively.



Figure 26: Top-50 object categories with the largest ratio of the difference between category- and instance-level labels.

(7) Hierarchical Mapping (HMa): first mapping f_I to category-level attribute or affordance space by an MLP supervised by GT A or B. Then the mapped features are fed to an MLP-Sigmoids to predict P_{α} or P_{β} .

Fold II. Directed arc from β to α .

(8) DM from β to α (DM- $\beta \rightarrow \alpha$): training a β classifier with f_I same with DM-V, but using the concatenated representation of affordance as f_β to train the α classifier.

(9) **DM from** β and I to α (**DM**- $\beta I \rightarrow \alpha$): training a β classifier with f_I same with DM-V, but using the concatenated representation of attributes f_β and objects f_I to train the α classifier.

Fold III. Directed arc from α to β .

(10) DM from α to β (DM- $\alpha \rightarrow \beta$): training an α classifier with f_I same with DM-V, but using the concatenated representation of attributes as f_{α} to train the β classifier.

(11) DM from α and I to β (DM- $\alpha I \rightarrow \beta$): training an α classifier with f_I same with DM-V, but using the concatenated representation of attributes f_{α} and objects f_I to train the β classifier.

(12) Ngram [68]: adopting Ngram to retrieve the relevance between α and β and generating an association matrix $M_{\alpha-\beta}$. Then we multiply DM predicted P_{α} with $M_{\alpha-\beta}$



Figure 27: More OCL samples of object categories.

to estimate P_{β} .

(13) Markov Logic Network (MLN-GT) [69]: adopting MLN to model the $\alpha - \beta$ relations following [16]. After training on OCL, we infer β with GT α to estimate its *performance upper bound*.

(14) Instantiation with attention (Attention): feeding $[f_{\alpha}, f_I]$ to an MLP-Sigmoid to generate attentions and predicting P_{β} by multiplying the attentions with P_B .

We operate baselines with a directed arc from α to β (Fold III) to perform ITE. The ITE calculation needs **feature zero-masking** to eliminate the effect of specific attributes [65]. These methods (DM-At, DM-AtO, Attention, OCRN) follow the same ITE calculation (feature masking). Two unique cases are Ngram and MLN-GT. Ngram uses attribute probabilities to infer affordance. Thus, we randomize the specific attribute probabilities for Ngram to operate the ITE calculation. And MLN-GT must use GT attribute labels to distinguish the "positive" and "negative" causes and then reason out the effect affordance. Thus, in ITE, we directly eliminate its corresponding attribute input.

G. Detailed Result Analysis

G.1. Detailed Attribute and Affordance Performances

We compute and analyze the performance (AP) of OCRN on each attribute or affordance class in Fig. 35 and Fig. 36, which suggest that visually abstract concepts like fake are more difficult to model than concrete ones like metal, breakable. The performance of attribute classes



Figure 28: More OCL samples of attributes and affordances.

is lower than affordance classes. This is mainly because the attributes have more diversity. Thus the *positive* instances of each attribute class are **less** than the affordance class.

G.2. Visualization of ITE Result

In Fig. 37, we show the correct instance proportions (%) of OCRN and Attention after ITE. (a) randomly chosen causal pairs $[\alpha_p, \beta_q]$ with ground truth $\beta_q = 1$, expecting $\hat{\beta}_q > \hat{\beta}_q |_{do(\mathfrak{A}_{\mathbb{R}})}$. (b) randomly chosen causal pairs $[\alpha_p, \beta_q]$ with ground truth $\beta_q = 0$, expecting $\hat{\beta}_q < \hat{\beta}_q |_{do(\mathfrak{A}_{\mathbb{R}})}$. The higher proportions indicate that OCRN performs better on ITE.

G.3. Attribute and Affordance Recognition Given Detected Boxes

Though OCL is a high-level concept learning task with object boxes as inputs, we can also consider object detection in evaluation for practical applications. We adopt Swin Transformer (Swin) [71] as the detector. It is pretrained on COCO [2] and finetuned on the OCL train set with GT boxes of 381 categories. On the OCL test set, it achieves 22.9 AP_{50} on object detection. Subsequently, it will provide detected box b_o for all models in inference. We can consider the detection effect in the attribute and affordance recognition metric to build a more strict criterion. Namely, all false positive detections (IoU<0.3 with referring to GT boxes) as the *false positives* of α and β recognition too. Moreover, ITE calculation needs to construct the counterfactual of an object instance. If the inaccurately detected object box shifts according to the GT box, it is difficult to know whether the counterfactual comes from the attribute masking or visual content change, using the corresponding attribute-affordance causal relation labels of this GT box. Thus, considering the unique property of causal inference different from common recognition, here we do not report the ITE score. Tab. 5 shows the results given detected boxes. Due to the more strict criterion and detection quality, the



Figure 29: More OCL samples. We present objects in different states, together with their key attributes and affordances.



Figure 30: More OCL samples of causal relations.

performances of all methods degrade greatly. But OCRN still holds the superiority on two tracks.

G.4. OCL-Based Image Retrieval

We visualize the OCL reasoning performance by retrieving the top-score instances with OCRN. Some results are shown in Fig. 38 and Fig. 39. The model can correctly retrieve the related images, especially on some common concepts e.g., columnar, sit.

H. Application on Human-Object Interaction (HOI) Detection

To further verify the generalization ability, we apply OCL to Human-Object Interaction (HOI) detection [81, 82, 83] and help HOI methods boost their performances. HOI



Figure 31: More OCL samples in the same category but different states.



Figure 32: Counts of object categories.

detection recently attracts a lot of attention and makes progresses [44, 84, 85, 86, 87, 88, 89, 90] thanks to the success of deep learning and large-scale HOI datasets [44, 56, 91, 92]. mans. Usually, an object has multi-affordance, *i.e.*, a person can perform different actions upon it. But in an image, just one or several actions/affordances are usually happening/**activated**. Without object knowledge, previous methods [93, 94, 95] can find the activated affordances from hun-

HOI depicts the actions performed upon objects by hu-





dreds of actions [44]. For example, for each human-object pair in HICO-DET [44], a model has to select one or several actions from the defined 116 actions. With OCL, things are different. OCL covers many actions, so we can use OCRN to infer P_{β} of an object to narrow the solution space. Thus, we propose two ways:

(1) **OCL Filtering**: We use P_{β} to narrow the action space with a threshold γ and generate P_{β}^{γ} . Affordances with probabilities higher than γ are kept and others are set to *zero* ($\gamma = 0.5$). Then, the HOI model only needs to predict in a narrowed action space. In practice, we multiply the prediction P_{HOI} from HOI model with P_{β}^{γ} element-wisely to





Figure 36: AP of affordance classes.

| Method | α | β |
|------------|------------|------|
| DM-V | <u>7.4</u> | 11.0 |
| DM-L | 4.6 | 9.1 |
| MM | 5.4 | 9.9 |
| LingCorr | 1.7 | 5.6 |
| KPMF | 6.4 | 10.5 |
| A&B-Lookup | 4.1 | 5.8 |
| HMa | 6.5 | 10.9 |
| DM-At | 6.8 | 10.5 |
| DM-AtO | 6.6 | 10.8 |
| Ngram | 5.1 | 10.2 |
| MLN-GT | - | - |
| Attention | 5.5 | 10.1 |
| OCRN | 7.9 | 11.3 |

Table 5: Attribute and affordance recognition results given detected boxes from Swin Transformer [71].

obtain the final prediction $P'_{HOI} = P_{HOI} * P^{\gamma}_{\beta}$.

(2) **Human-as-Probe**: Another more straightforward way is to predict HOI via OCL directly. We treat the human paired with the object as a **probe**. Assuming the human feature is f_h and human-object spatial configuration feature is f_{sp} (from [93, 94]). As P_β indicates all possible affordances, the ongoing actions can be seen as the **instantiation** of P_β , *i.e.*, they are activated by the "probe" f_h and f_{sp} . So we use f_h and f_{sp} to generate attention A_{h+sp} via MLP-Sigmoid. Then we operate $P_\beta * A_{h+sp}$ and late fusion to get the final prediction $P'_{HOI} = (P_\beta * A_{h+sp} + P_{HOI})/2$.



Figure 37: ITE given different $[\alpha_p, \beta_q]$.

Concretely, we use OCRN to enhance HOI detection models (iCAN [94], TIN [93], IDN [95]) on HICO-DET [44]. As OCL merely contains 15 object categories in HICO-DET [44], the rest 65 object categories are **unseen**. We embed OCRN into three HOI models according to OCL filtering and Human-as-Probe, and the public model checkpoints of [94, 93, 95] are used.

The results are shown in Tab. 6. With OCL filtering, iCAN [94], TIN [93], and IDN [95] achieve a gain of mAP by 0.65%, 0.90%, and 0.77% respectively. The Human-as-Probe is more suitable for HOI detection and contributes a performance boost of 1.50%, 1.46%, and 0.98% to three models. These strongly verify the efficacy and generalization ability of OCL.

I. Comparison on Imbalance Learning

I.1. Debiasing Learning

The motivation of the OCRN is to follow the prior knowledge of the three levels of objects with a deep learning-based causal graph model, to pursue the object understanding beyond the common direct mapping from pixTop-5 Attribute Retrieval with OCRN



Figure 38: Top-5 attribute retrievals on the OCL test set.

Top-5 Affordance Retrieval with OCRN

Capture Eat Ignite Sit

Figure 39: Top-5 affordance retrievals on the OCL test set.

| Method | Full | Rare | Non-Rare |
|----------------|-------|-------|----------|
| iCAN | 14.84 | 10.45 | 16.15 |
| iCAN+Filtering | 15.49 | 8.76 | 17.50 |
| iCAN+Probe | 16.34 | 11.66 | 17.74 |
| TIN | 17.03 | 13.42 | 18.11 |
| TIN+Filtering | 17.93 | 13.79 | 19.17 |
| TIN+Probe | 18.49 | 15.02 | 19.58 |
| IDN | 23.36 | 22.47 | 23.63 |
| IDN+Filtering | 24.13 | 23.74 | 24.24 |
| IDN+Probe | 24.34 | 24.03 | 24.43 |

Table 6: Results of HOI detection (using detected object boxes).

els to labels, and to avoid the bias estimation such as in the Simpson's paradox [20]. Thus, we use intervention to deconfound the confounder *category* and exclusive the possible spurious bias and correlation imported bias from im-

| Model | Test Inference | α Amp. | β Amp. |
|------------------------------------|---|---------------|--------------|
| OCRN | $\operatorname{argmax}_{y} P(y x)$ | 0.127 | 0.112 |
| DM-V + Joint ND-way Softmax | $\operatorname{argmax}_{y} \max_{d} P_{te}(y, d x)$ | 0.151 | 0.158 |
| DM-V + Joint ND-way Softmax | $\operatorname{argmax}_{u} \sum_{d} P_{te}(y, d x)$ | 0.148 | 0.154 |
| DM-V + N-way classifier per domain | $\operatorname{argmax}_{y} P_{te}(y d^*, x)$ | 0.135 | 0.112 |
| DM-V + N-way classifier per domain | $\operatorname{argmax}_{y} \sum_{d} s(y, d, x)$ | 0.147 | 0.145 |

Table 7: Comparison with debiasing models.

balanced object categories. Overall, we propose our OCRN in a causal inference perspective instead of the pure classification viewpoint, which also suits our causal graphical model well. Similar cases are also proposed in recent works like [53, 78, 65]. Moreover, to better compare our method with the common debiasing methods, we further conduct the experiments as follows.

We regard α, β recognition as multiple independent binary classification tasks and implement some methods introduced in [96] on our strong baseline DM-V to reduce bias from object categories. We use mean bias amplification (Amp) in [97] as bias evaluation metric: small Amp means model suffers less from data category bias. The test results are shown in Tab. 8. The proposed OCRN has comparable or smaller bias amplification than the variants of DM-V since our model follows the causal graph and exploits the tools of causal inference, while most methods for category bias are from the view of classification.

To verify the debiasing of OCRN, we compare the model bias of OCRN w/ or w/o deconfounding. The bias of category O upon an attribute α is measured following [97], by $b(O,\alpha) = c(O,\alpha) / \sum_{\alpha'} c(O,\alpha')$. When measuring **data** bias, $c(O, \alpha)$ is the number of co-occurrence of O and α in OCL, and when it comes to **model** bias, $c(O, \alpha)$ is the sum of probabilities that O are predicted positive with α . The bias of β is measured in the same manner. Fig. 40 and 41 show some examples of the biases of training data and models, indicating that OCRN deconfounding effectively prevents the model from bias toward the train set.

I.2. Long-tailed Learning

Besides the debiasing learning techniques, we also apply longtailed learning methods on our baseline method DM- $\alpha \rightarrow \beta$ for comparison, including class-balanced sampling [98] and focal loss [99]. The models with additional re-balancing modules suffer from minor accuracy degradation, mainly for OCL is long-tailed on object class while we infer α , β , so the gap minimizes the effect of long-tailed learning.

J. Discussion about States

We did not use object states in our model because there is also a **compositional zero-shot problem** and object-state pairs, *i.e.*, there can be **unseen** states in real-world data.



Figure 40: Attribute bias (w/ and w/o deconfounding) for category frying pan.

| Method | α | β |
|--|-------|-------|
| $DM-\alpha \rightarrow \beta$ | 28.7% | 52.6% |
| $DM-\alpha \rightarrow \beta$ +Class balance sampling [98] | 27.3% | 52.1% |
| $DM-\alpha \rightarrow \beta$ +Focal loss [99] | 27.6% | 51.2% |

Table 8: Comparison with debiasing models.

Differently, affordances are more general. The models explicitly incorporating object states will fail to generalize to these zero-shot states and it adds to the object category bias. In experiments, the state supervision during training would indeed *slightly improve* the affordance recognition performance, since instances in the **same state** lie in a tight cluster in affordance label space. But this will hurt the ITE performance greatly.

K. Discussion about Causality and Causal Graph

Annotating causality in the real world is extremely difficult. In data annotation, we have met numerous ambiguities and difficulties to confirm the "right" causal relations. To address these challenges, we follow the following principles: (1) Firstly, we only emphasize clear and strong causal relations via crowdsourcing, but omit the vague ones. (2) Second, we take an object affordance-centric viewpoint to look at the possible causal relations. (3) We would rather discard than condone the controversial situation in the annotation. (4) We only focus on the simple relations between one attribute and one affordance, instead of the very complex compositions of multiple attributes and affordances which are almost impossible to annotate. Therefore, we finally find that we can label a very small percentage of all arcs with the whole causal graph consisting of so many nodes (category, attributes, affordances, contexts, etc.) while keeping the quality.

Our causal graph follows the human priors from our ex-



Figure 41: Affordance bias (w/ and w/o deconfounding) for category giraffe.

perts and crowdsourcing annotators. Some previous works also follow this before designing the method, such as [16]. From the viewpoint of causal discovery [20, 78, 65, 53], the above arcs (*e.g.*, the inverted arc from attribute to category in the causal graph directed acyclic graph, DAG) are indeed possible. However, here, we mainly study the object concept learning problem, especially attribute and affordance learning for intelligent robots and embodied AI. Thus, from the perspective of affordance learning, we think the arcs from category to attribute and affordance are more vital and meaningful to us.

Causality can also be confused with enabling condition. In OCL, the affordance of an object indicates what human can do to/with it. In this case, "fresh" causes "eatable" (rather than causes "eat" action). As causality is discussed in the view of embodied agents, this rule can hold. In modern causal inference models like structured causal models (SCM), causality and enabling conditions are not strictly distinguished. As stated by Cheng et al. [100], causes and enabling conditions hold the same logical relation to the effect in those terms and the methods that explain their distinction come from the subject judgment of humans. The distinction can be explained based on the nor*mality* of potential factors, or considering the existing assumption of the inquirer. They proposed an approach by measuring the covariation between potential factors to the effect over a set of questions. So in SCM, both will be represented as nodes and involved in causal mechanisms. OCL follows the "open" setting: affordance is a subjective property of the object, so all reasons given by humans/robots (including enabling conditions) are regarded as causal factors.

L. Detailed Lists

The detailed object categories, attributes, and affordances are listed on our website: https:// mvig-rhos.com/ocl.