Supplementary Materials for 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

We report more details and analyses here:

1. Category/Attribute/Affordance Selection

We choose affordances, categories, and attributes, considering their causal relations. Their word clouds are shown in Fig. 1. The complete lists can be found in Suppl. Sec. 12.

(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 [4], 160 from [11], 146 from [3], 97 from [12], 41 from [45], 21 from [31] (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 [4], 78 from [11], 127 from [3], 53 from [12], 13 from [45], 11 from [31], with *overlaps*).

(2) **Category**: Considering the taxonomy (WordNet [9]), we collect a pool with over 1,742 object categories from previous datasets: 32 from [8], 28 from [33], 717 from [42], 1,000 from [25] (with *overlaps*). Then we merge the similar categories according to WordNet [9] 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. 2.

(3) **Attribute**: We extract the attributes from several large-scale attribute datasets: 64 from [8], 203 from [33], 66 from [42], 25 from [25], top 500 from [15]), 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.

2. Annotation Details

2.1. Attribute Annotation

(1) **Category-level attribute** (A). Following [32], 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.

2.2. Affordance Annotation

(1) **Category-level affordance** B. Following [4], 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

[*] Corresponding author.

(a) Category (b) Attribute (c) Affordance

Figure 1: Word clouds of object categories, attributes, and affordance (by positive frequencies in OCL).

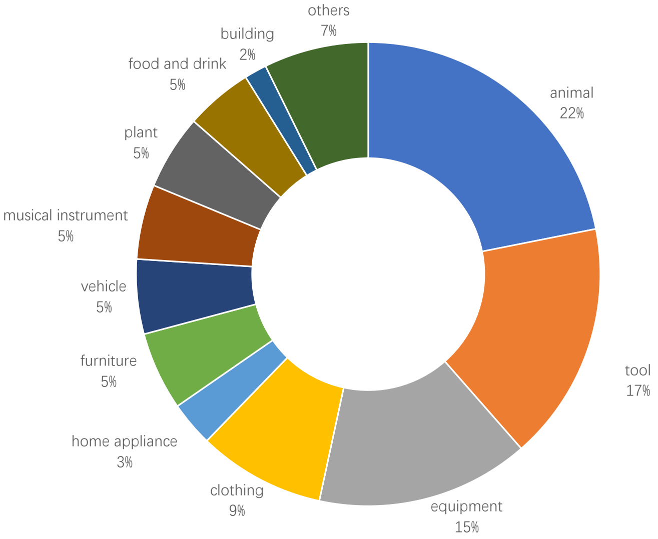


Figure 2: Super-categories of objects in OCL.

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* [14]. 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.

2.3. 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 114×170 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. 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

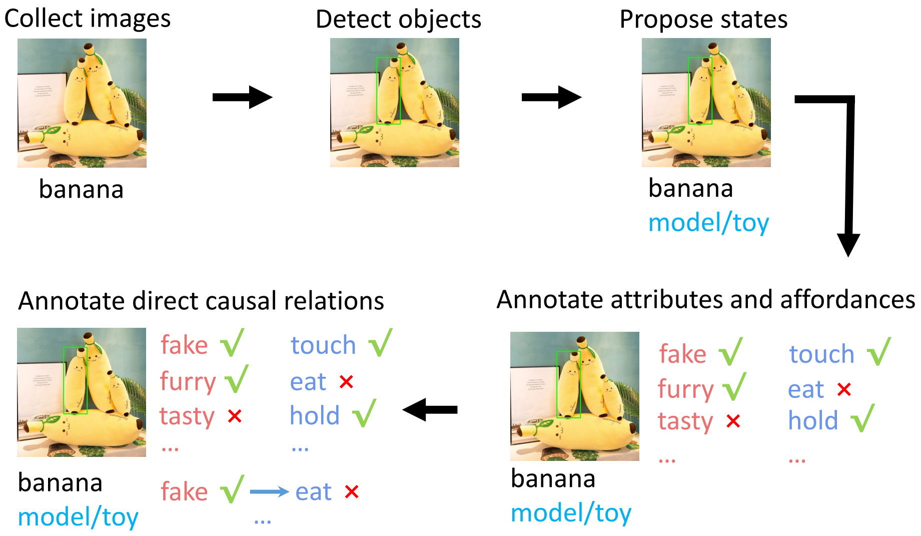


Figure 3: A running example of dataset construction.

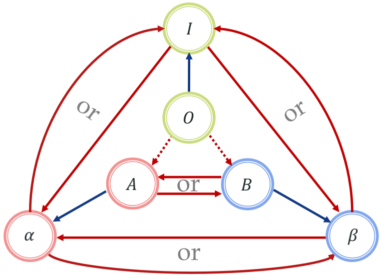


Figure 4: A more complex causal graph of our knowledge base. A, B, O are the object category and category-level attribute and affordance. I is the object appearance, α, β are the instance-level attribute and affordance. Note that "or" indicates that the arcs between A, B, α, β, I, α, and I, β indicate that either $A \leftarrow B$ or $A \rightarrow B$ (the others are similar) is considering in the setting.

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 *unambiguous* 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.

2.4. A Running Example of Dataset Construction.

A running example is shown in Fig. 3 to show the process of annotations clearly.

3. 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.

3.1. Basics of Causal Inference and Causal Graph

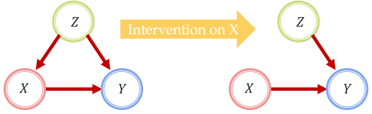


Figure 5: An example of the causal graph and causal intervention. We study the causal relation $X \rightarrow Y$ while confounder Z exists and brings bias. After the intervention on variable X, the poisonous relation $Z \rightarrow X$ is eliminated.

A causal graph is a DAG that describes the causal relations between multiple factors. Each directed edge points from the "cause" to its "effect", *e.g.* in Fig. 5, 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. 5 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. 5, 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. [34] *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. 5, the confounder Z can be eliminated with **Back-door Adjustment**:

$$P(Y|do(X)) = \sum_z P(Y|X, Z=z)P(Z=z), \qquad (1)$$

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.

3.2. Causal Graph of Our Knowledge Base

A more complicated causal graph considering more arcs between nodes is shown in Fig. 4. The causal relations between nodes or arcs in Fig. 4 are determined as follows:

Firstly, we introduce two kinds of special arcs.

$O \rightarrow A$, $O \rightarrow B$ (dotted arcs): in OCL, A and B are defined as the category-level annotations. Given O, A, and B are strictly determined. In Fig. 4, 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 instance-level 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 concept-level 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 [13].

$A \leftarrow B$ or $A \rightarrow B$, $\alpha \leftarrow \beta$ 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`.

3.3. 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. 6).

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. 7.

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. 8.

4. OCL Characteristics

4.1. Object Box Size

We visualize the distribution of normalized object box size in Fig. 9, 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.

4.2. Annotator Information

Annotators' age, major, and education degree are presented in Fig. 10, 11, and 12.

4.3. Matrix Samples

The category-level attribute and affordance (A, B) matrices are detailed in Fig. 13, 14 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.

4.4. 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. 15 shows the state distribution per object category.

4.5. Attribute-Affordance Relation

We analyze the instance-level attribute-affordance relations in our knowledge base under three criteria. (1) **At-**

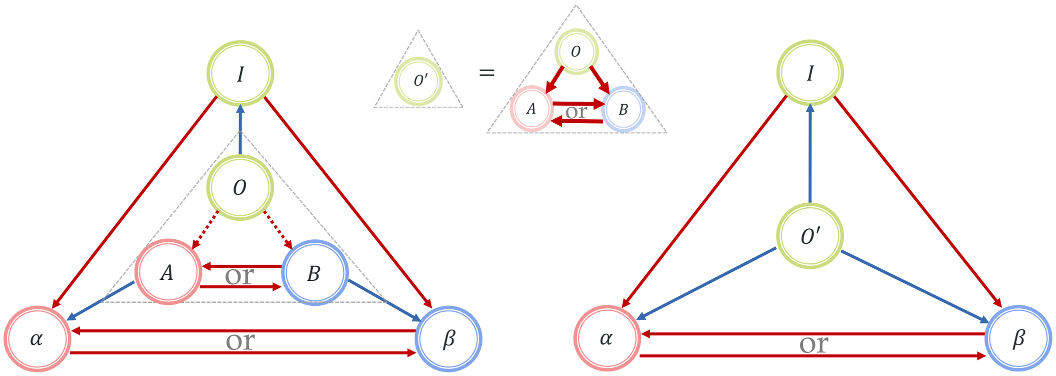
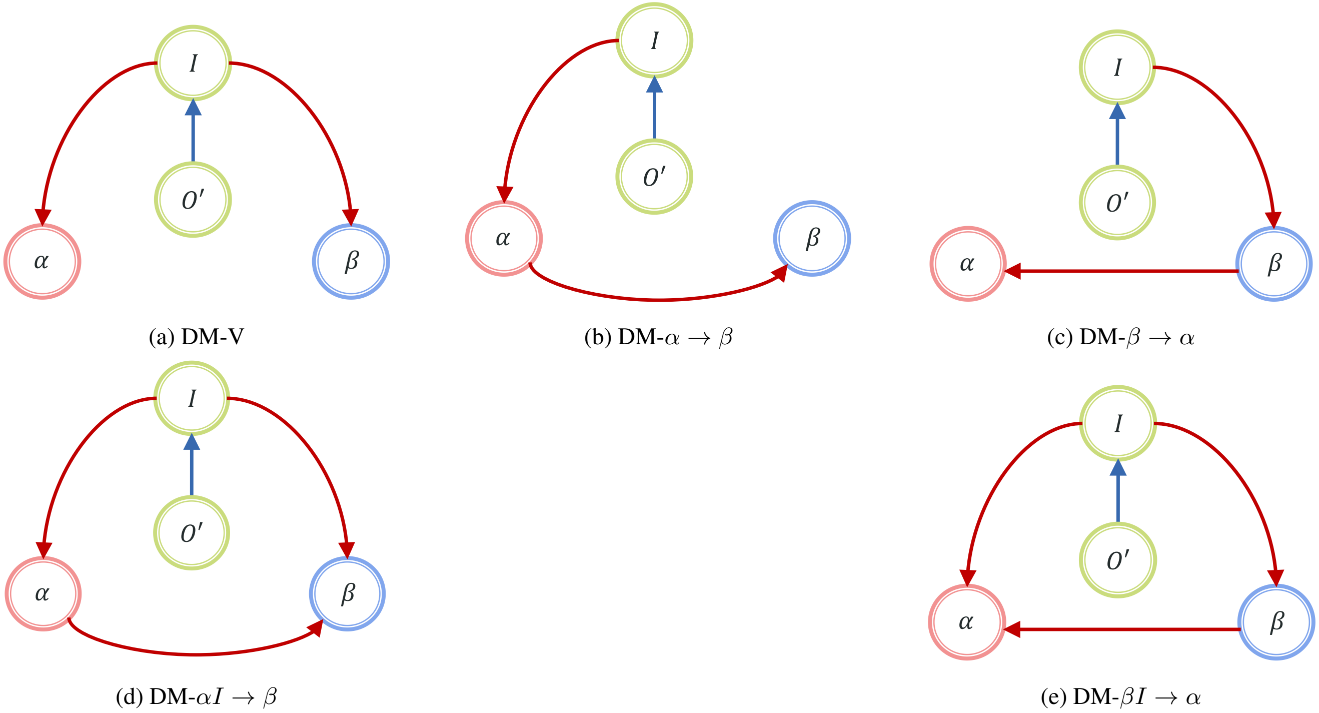


Figure 6: Simplified causal graph for OCL task. Note that "or" indicates that the arcs between A, B and α, β are either $A \leftarrow B$ or $A \rightarrow B$ ($\alpha \leftarrow \beta$ or $\alpha \rightarrow \beta$), instead of concurrence.



(a) DM-V

(b) DM-$\alpha \rightarrow \beta$

(c) DM-$\beta \rightarrow \alpha$

(d) DM-$\alpha I \rightarrow \beta$

(e) DM-$\beta I \rightarrow \alpha$

Figure 7: Causal graphs of the baselines.

tribute 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. 16, 17, 18 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. 16, 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 the former two matrices in Fig. 16, 17, they all point out the relation between *tasty* and *pick*, since most *tasty* objects

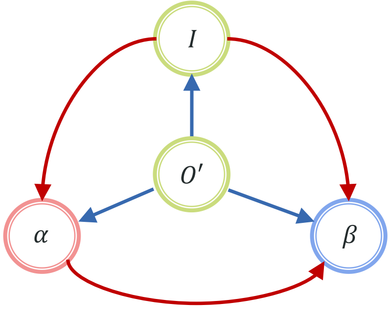


Figure 8: Causal graph of OCRN.

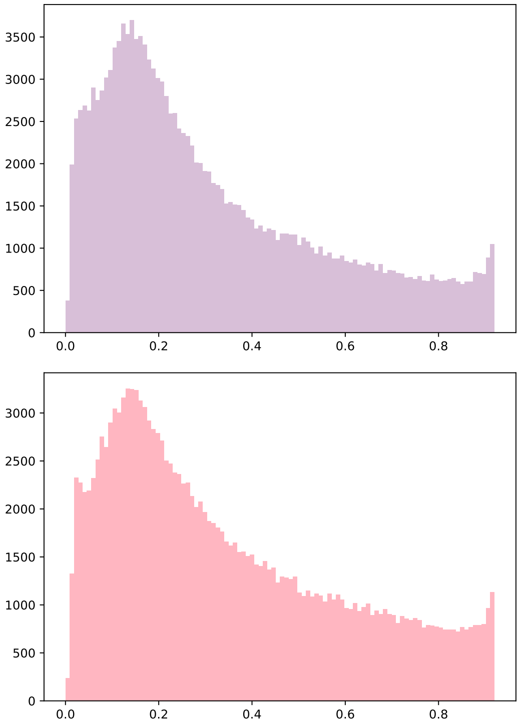


Figure 9: Distribution of normalized object box width (left) and height (right).

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. 18 is more sparse and clear of causality.

4.6. 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. 19 via t-SNE [29]. 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.

4.7. 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 attribute/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. 20. 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.

4.8. 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 $[\alpha_p, \beta_q]$ pairs to avoid the biased influence of very rare $[\alpha_p, \beta_q]$ pairs that include less than 35 object instances.

4.9. 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. 1. 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 1: Detailed data split of our knowledge base.

4.10. Images and Instances

Some additional data samples of our knowledge base are shown in Fig. 21, 22a, 22b, 23, 24, and 25, 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. 26, 27, and 28.

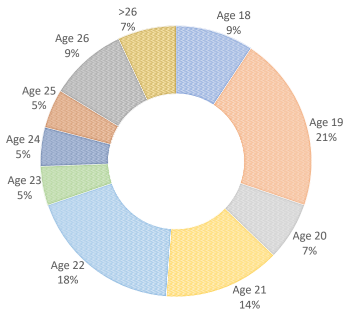


Figure 10: Age information of annotators.

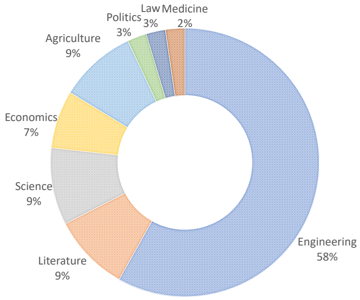


Figure 11: Major information of annotators.

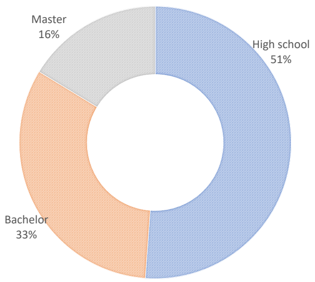


Figure 12: Degree information of annotators.

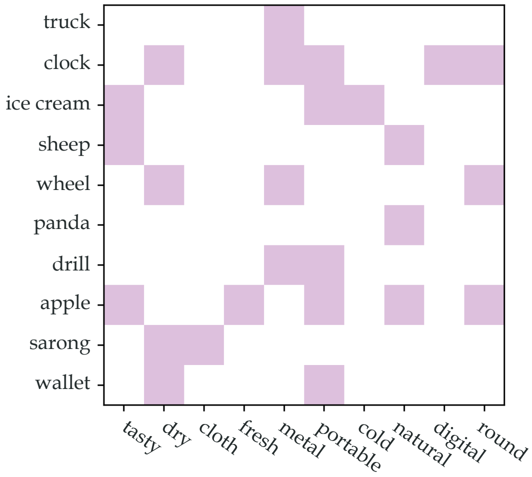


Figure 13: Category-level attribute (A) matrix.

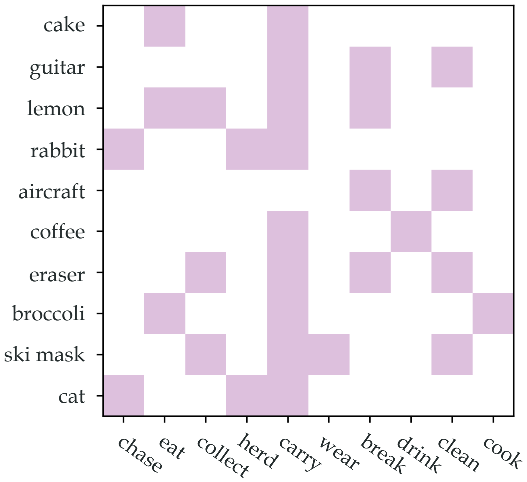


Figure 14: Category-level affordance (B) matrix.

4.11. 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 α, 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.

4.12. Potential Bias

We have considered the bias issue in the construction of our dataset. (1) In our dataset, the existing datasets (ImageNet [6], COCO [23], aPY [8], SUN [42]) are open-sourced 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 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.

5. ITE Metric Details

ITE (**Individual Treatment Effect (ITE)** [36]) 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

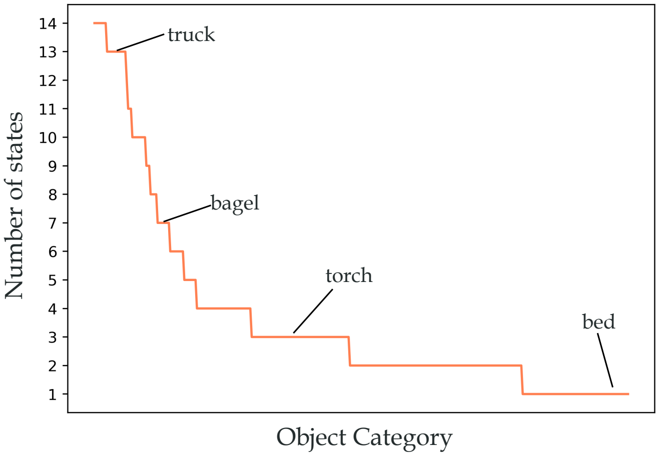


Figure 15: State distributions of different object categories.

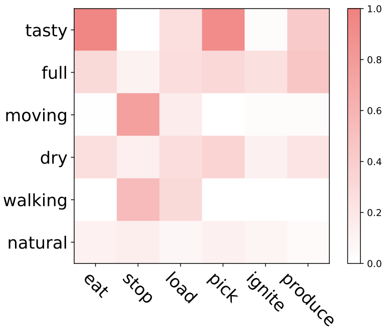


Figure 16: Attribute **conditioned** affordance matrix.

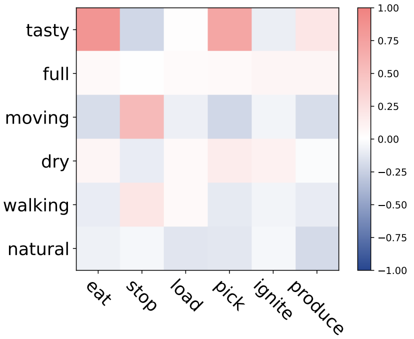


Figure 17: Attribute-affordance **correlation**.

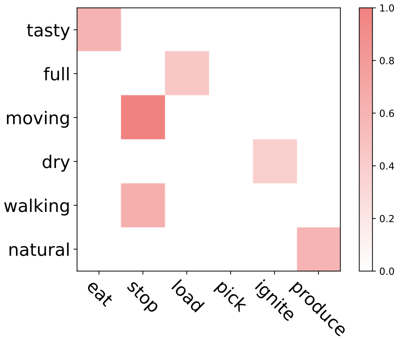


Figure 18: Attribute-affordance **causality**.

score:

$$\mathcal{S}_{\text{ITE}} = \begin{cases} \max(\Delta\hat{\beta}_q,\ 0), & \beta_q = 1, \\ \max(-\Delta\hat{\beta}_q,\ 0), & \beta_q = 0, \end{cases} \tag{2}$$

where

$$\Delta\hat{\beta}_q = \hat{\beta}_q|_{do(\alpha_p)} - \hat{\beta}_q|_{do(\cancel{\alpha_p})} == \hat{\beta}_q - \hat{\beta}_q|_{do(\cancel{\alpha_p})}, \tag{3}$$

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 $[\alpha_p, \beta_q]$, 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:

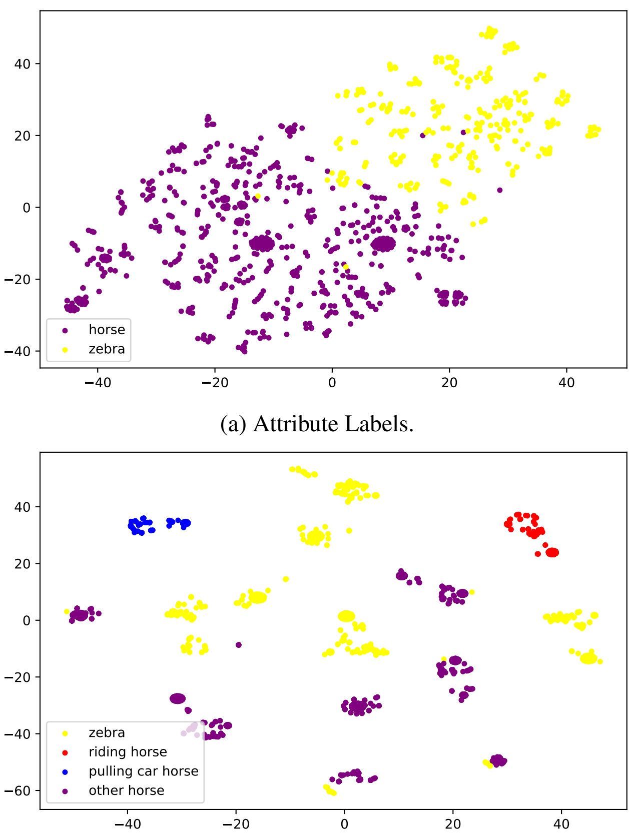
$$\begin{aligned} P(\hat{\alpha_p} = \alpha_p) &= \begin{cases} \hat{\alpha_p}, & \alpha_p = 1, \\ 1 - \hat{\alpha_p}, & \alpha_p = 0, \end{cases} \\ P(\hat{\beta_q} = \beta_q) &= \begin{cases} \hat{\beta_q}, & \beta_q = 1, \\ 1 - \hat{\beta_q}, & \beta_q = 0. \end{cases} \end{aligned} \tag{4}$$

And the overall metric is:

$$\mathcal{S}_{\alpha\text{-}\beta\text{-ITE}} = \mathcal{S}_{\text{ITE}} P(\hat{\alpha_p} = \alpha_p) P(\hat{\beta_q} = \beta_q) \tag{5}$$

The factors measure the correctness of attributes and affordances. Hence a model achieves a high $\mathcal{S}_{\alpha\text{-}\beta\text{-ITE}}$ 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



(a) Attribute Labels.

(b) Attribute-Affordance Labels.

Figure 19: Clustering using attribute and attribute-affordance labels.

ability. Similar to our OCRN, some works [39, 37, 38] also try to marry supervised deep learning and causal inference.

6. Baseline Details

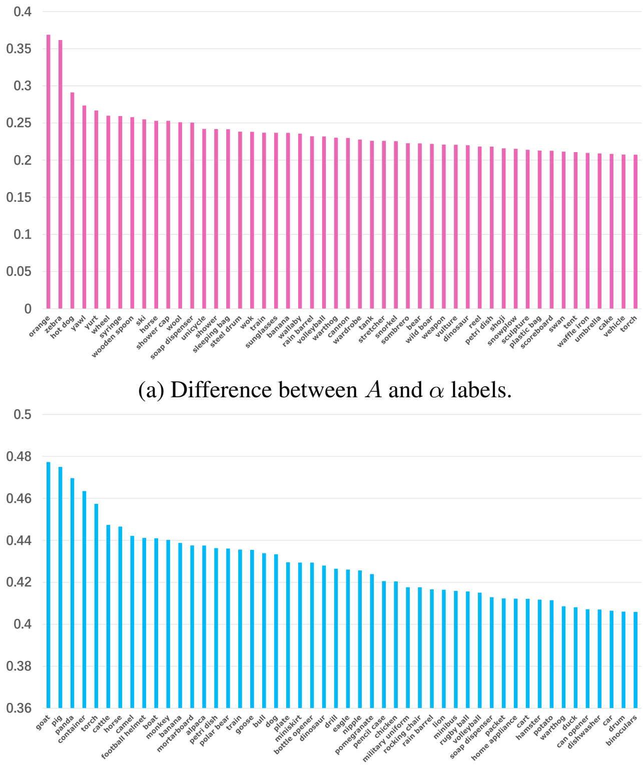
We introduce the details of all baselines here:

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 LayerNorm [1] 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 [7] 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 multi-modal aligned f_I is fed to an MLP-Sigmoids to predict P_α, P_β.



(a) Difference between A and α labels.

(b) Difference between B and β labels.

Figure 20: Top-50 object categories with the largest ratio of the difference between category- and instance-level labels.

(4) Linguistic Correlation (LingCorr): measuring the correlation between object and α/β classes via their Bert [7] cosine similarity. P_α, P_β are given by multiplying $P(O|I)$ to correlation matrices.

(5) Kernelized Probabilistic Matrix Factorization (KPMF) [44]: 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 [30] 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.

(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

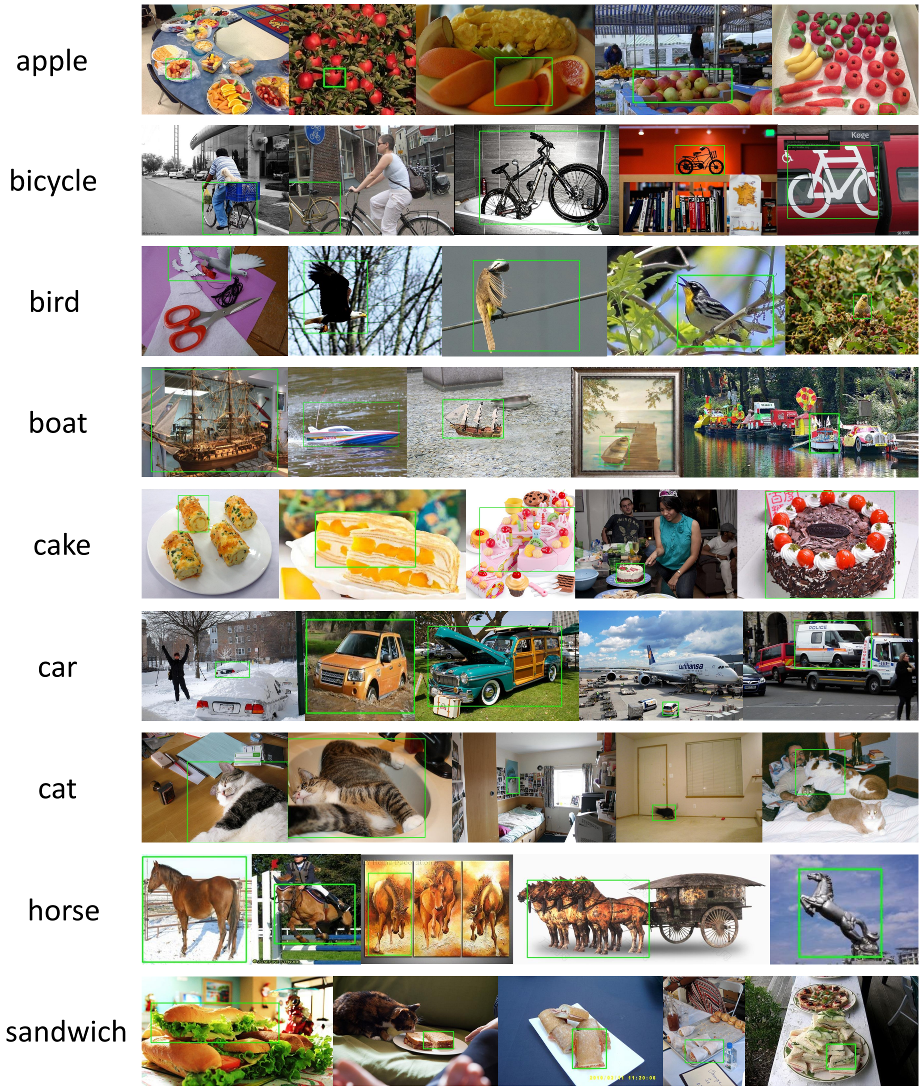


Figure 21: More OCL samples of object categories.

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.

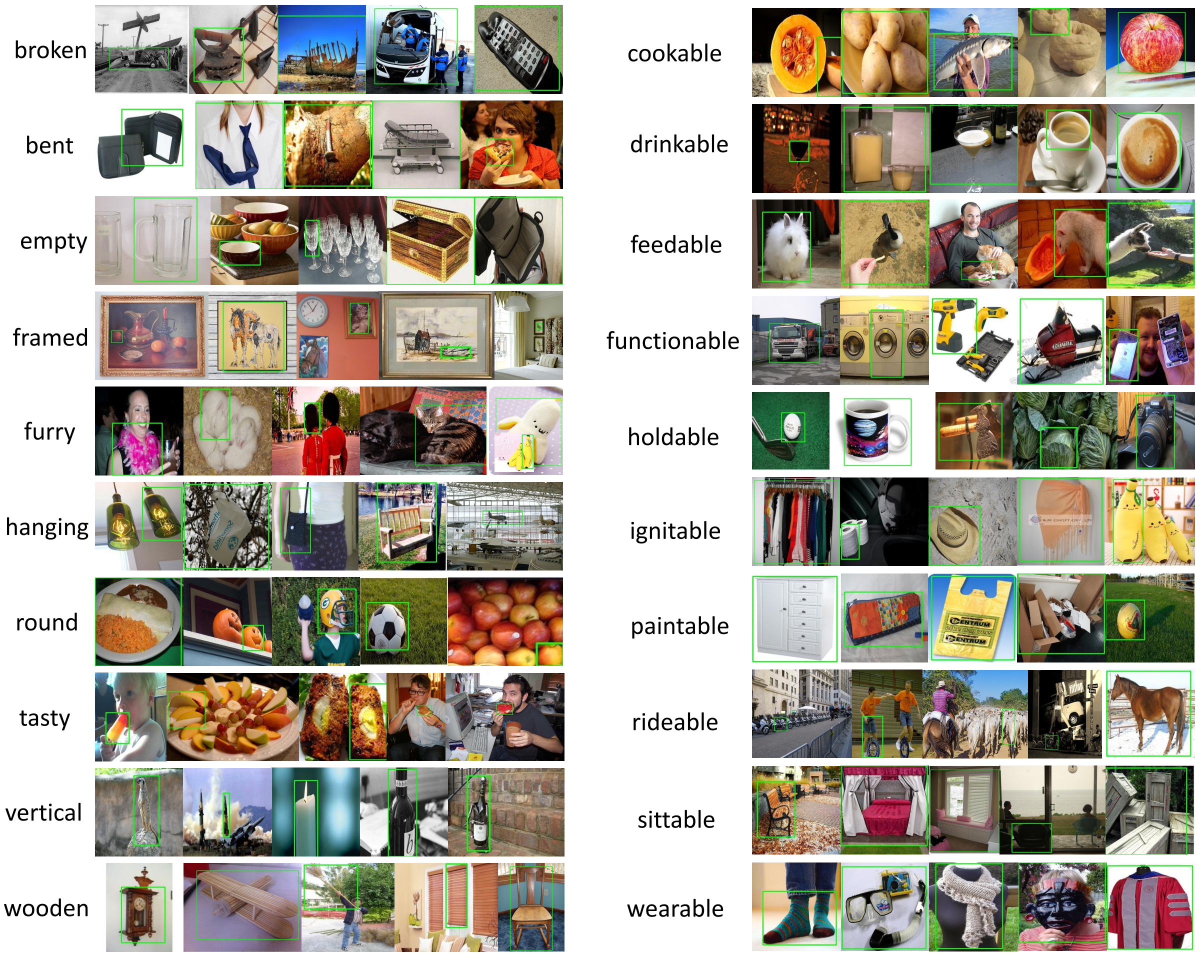
(11) DM from α and I to β (DM-$\alpha I \rightarrow \beta$): training an α classifier with f_I same with DM-V, but using the concatenated representation of attributes f_α and objects f_I to train the β classifier.

(12) Ngram [24]: 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 $M_{\alpha-\beta}$ to estimate P_β.

(13) Markov Logic Network (MLN-GT) [35]: adopting MLN to model the $\alpha - \beta$ relations following [45]. 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 **fea-**



(a) Attributes.

(b) Affordances.

Figure 22: More OCL samples of attributes and affordances.

ture zero-masking to eliminate the effect of specific attributes [38]. 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.

7. Detailed Result Analysis

7.1. Detailed Attribute and Affordance Performances

We compute and analyze the performance (AP) of OCRN on each attribute or affordance class in Fig. 29 and Fig. 30, which suggest that visually abstract concepts like `fake` are more difficult to model than concrete ones like `metal`, `breakable`. The performance of attribute classes 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.

7.2. Visualization of ITE Result

In Fig. 31, 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 $\hat{\beta}_q > \hat{\beta}_q|_{do(\cancel{\alpha_p})}$. (b) randomly chosen causal pairs $[\alpha_p, \beta_q]$ with ground truth $\beta_q = 0$, expecting $\hat{\beta}_q < \hat{\beta}_q|_{do(\cancel{\alpha_p})}$. The higher proportions indicate that OCRN performs better on ITE.

7.3. 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 detec-

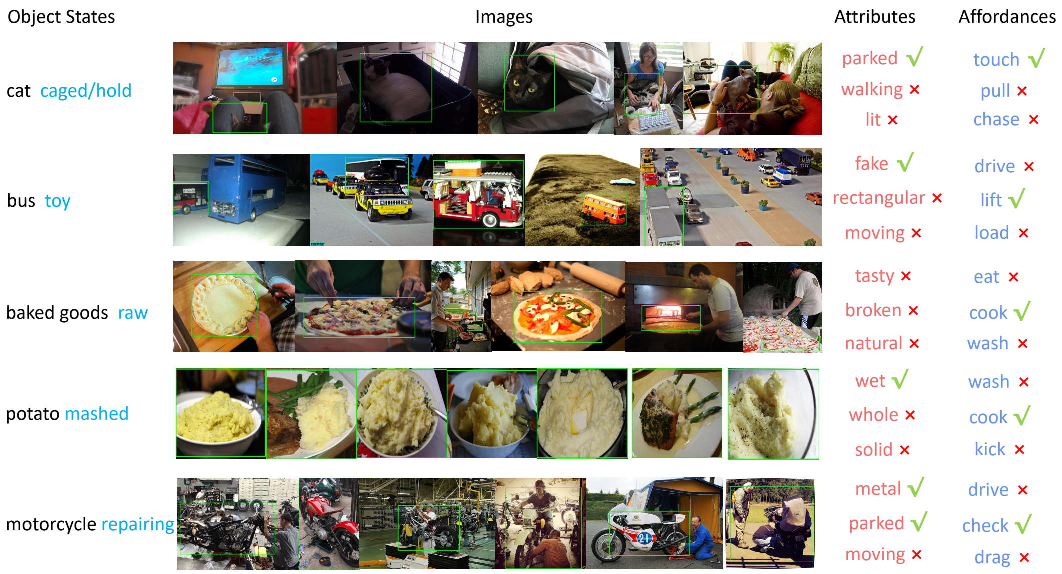


Figure 23: More OCL samples. We present objects in different states, together with their key attributes and affordances.

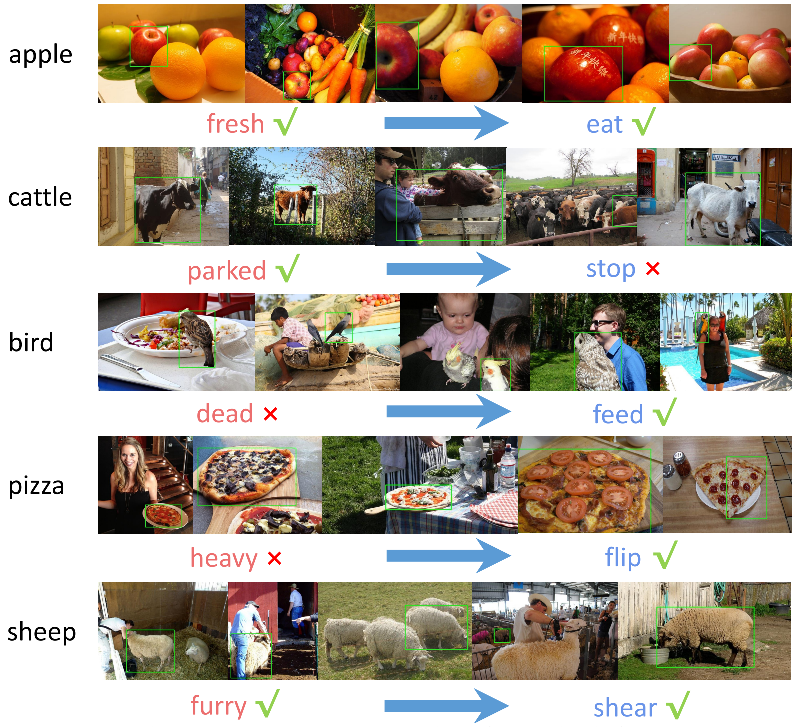


Figure 24: More OCL samples of causal relations.

tion in evaluation for practical applications. We adopt Swin Transformer (Swin) [28] as the detector. It is pretrained on COCO [23] 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.

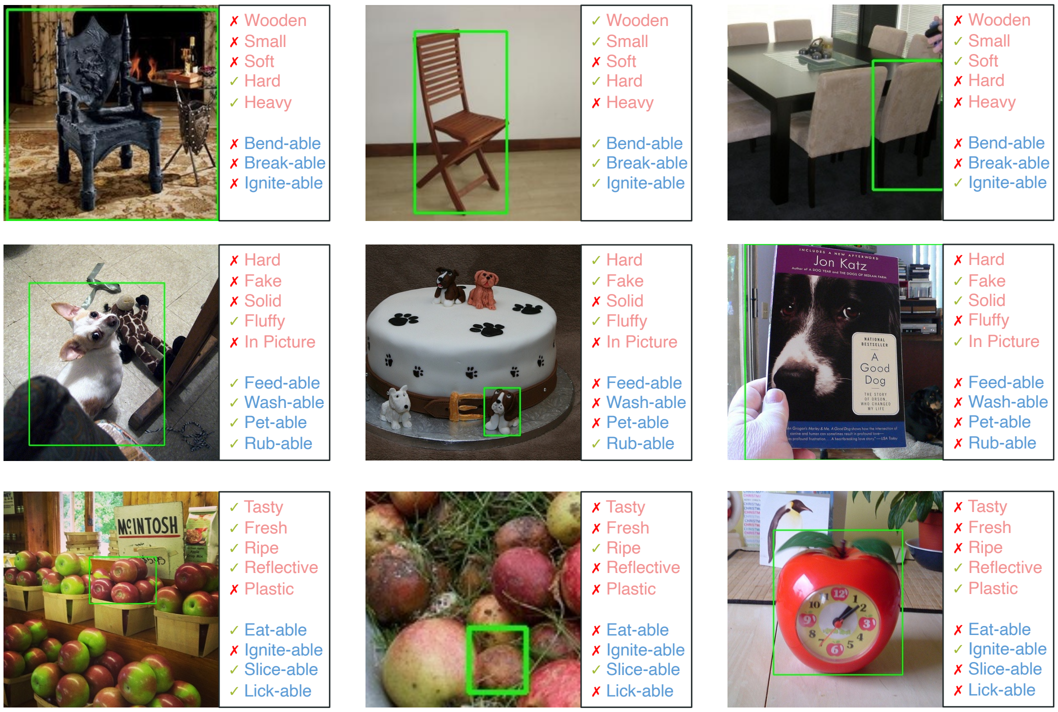


Figure 25: More OCL samples in same category but different states.

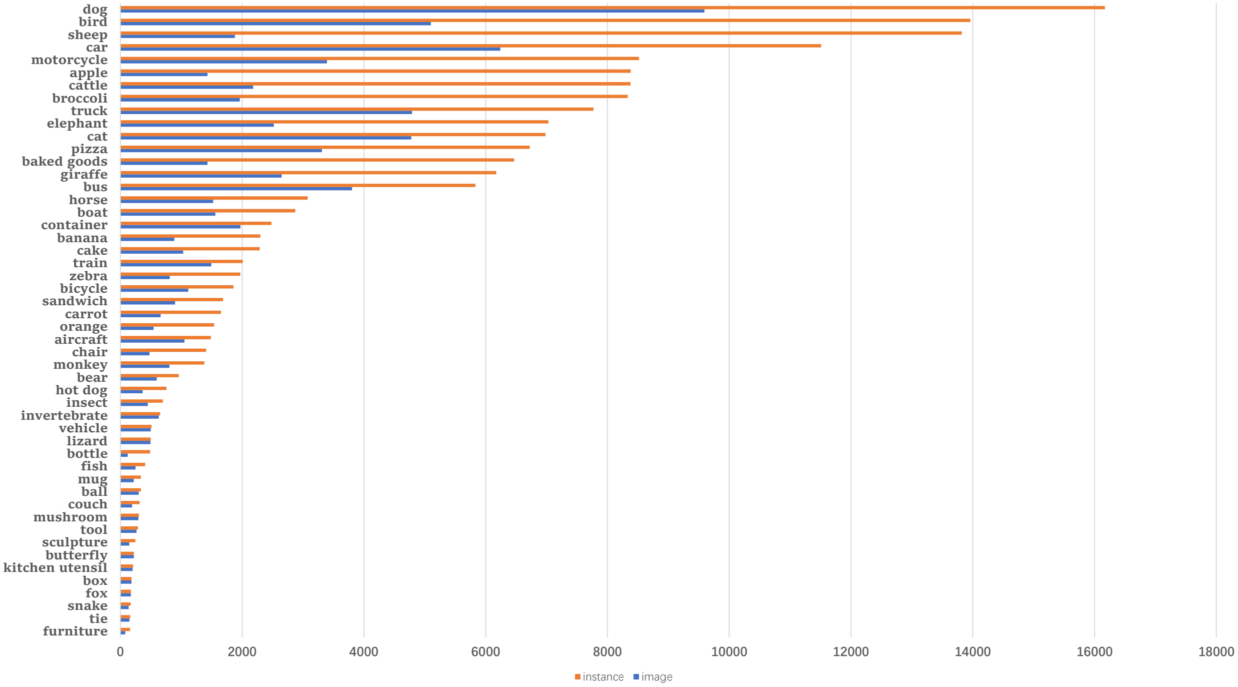


Figure 26: Counts of object categories.

Thus, considering the unique property of causal inference different from common recognition, here we do not report the ITE score. Tab. 2 shows the results given detected boxes. Due to the more strict criterion and detection quality, the performances of all methods degrade greatly. But OCRN still holds the superiority on two tracks.

7.4. OCL-Based Image Retrieval

We visualize the OCL reasoning performance by retrieving the top-score instances with OCRN. Some results are

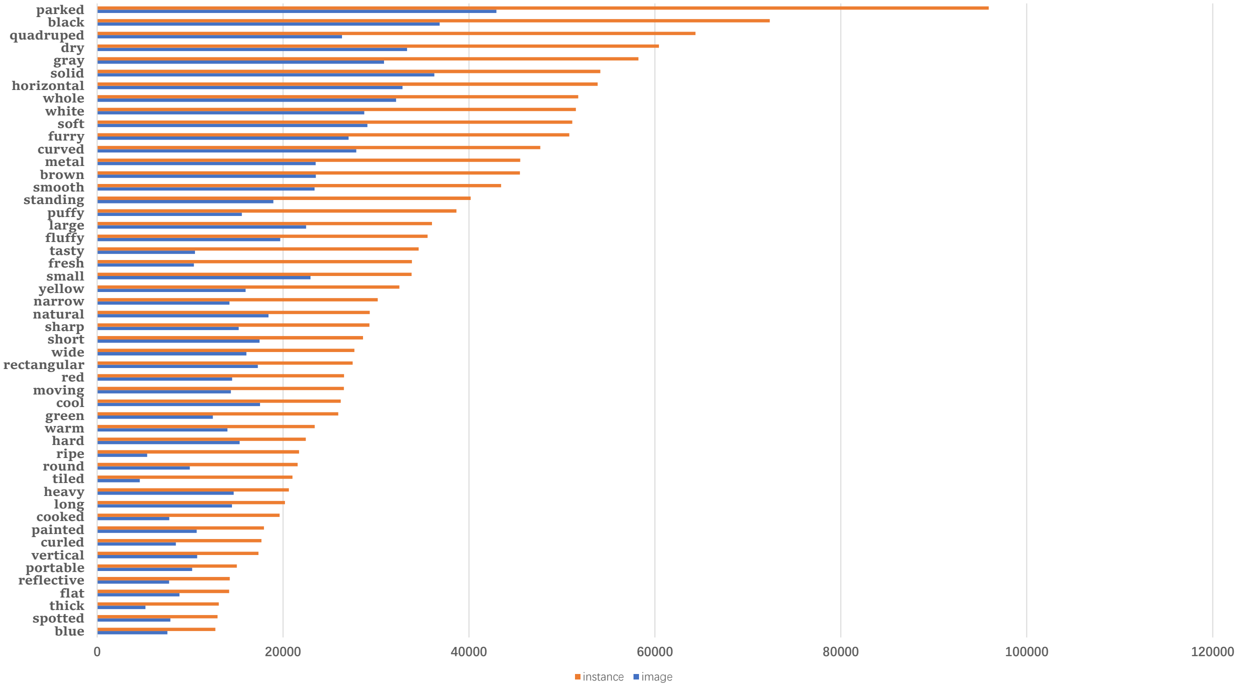


Figure 27: Counts of attribute classes.

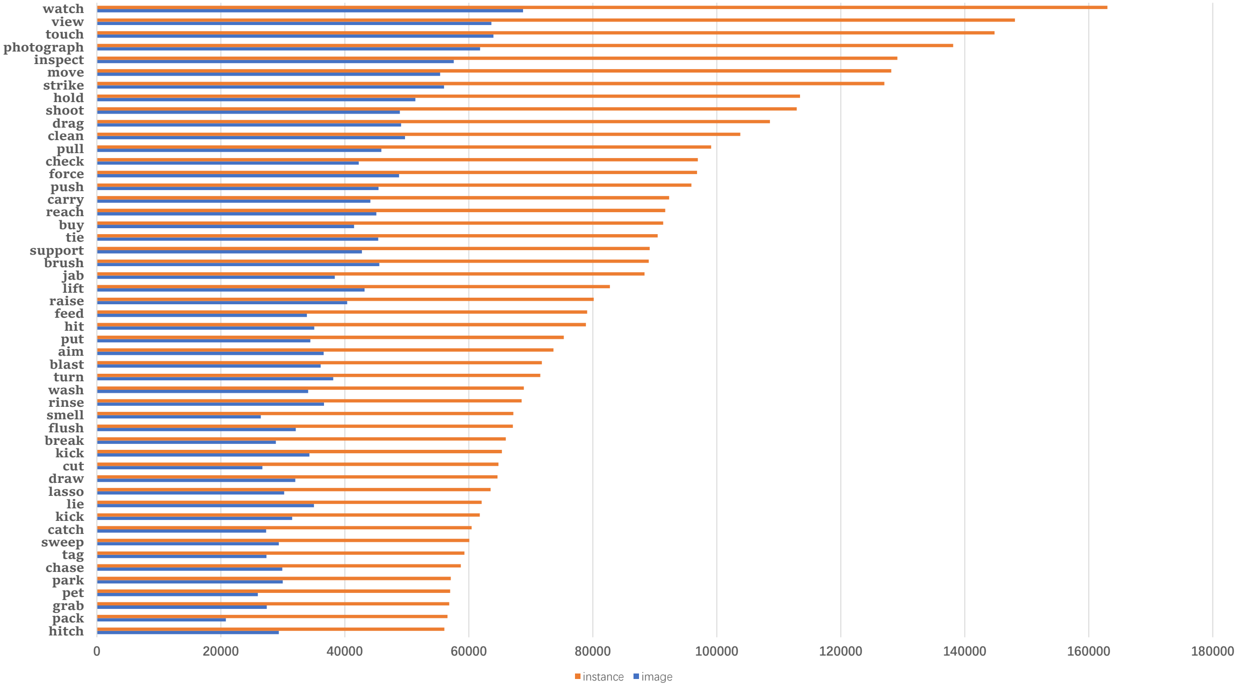


Figure 28: Counts of affordance classes.

shown in Fig. 32 and Fig. 33. The model can correctly retrieve the related images, especially on some common concepts *e.g.*, `columnar`, `sit`.

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

To further verify the generalization ability, we apply OCL to Human-Object Interaction (HOI) detection [20, 3, 16] and help HOI methods boost their performances. HOI detection recently attracts a lot of attention and makes pro-

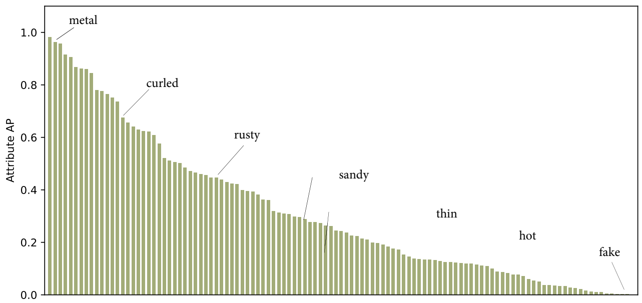


Figure 29: AP of attribute classes.

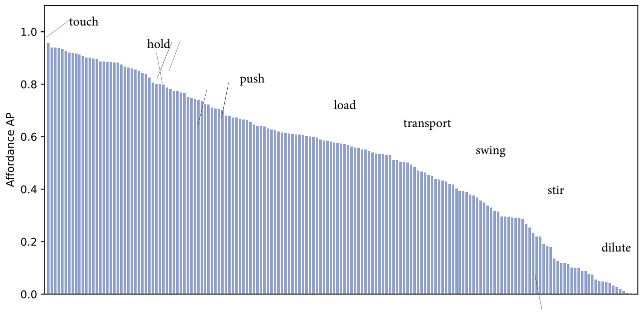


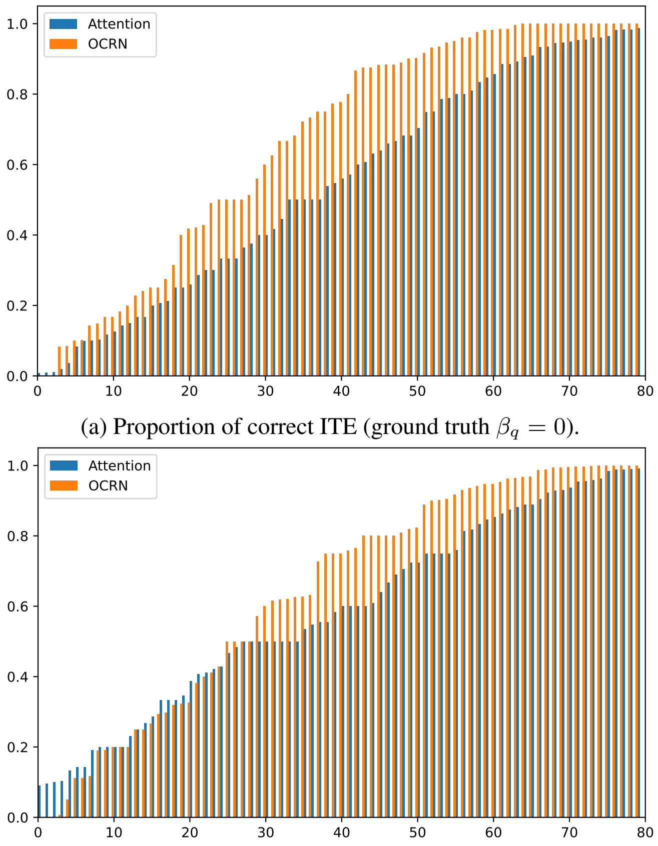
Figure 30: AP of affordance classes.

Method	α	β
DM-V	7.4	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 2: Attribute and affordance recognition results given detected boxes from Swin Transformer [28].

gresses [26, 27, 41] thanks to the success of deep learning and large-scale HOI datasets [3, 12, 19, 18].

HOI depicts the actions performed upon objects by humans. 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 [21, 10, 17] can find the activated affordances from hundreds of actions [3]. For example, for each human-object pair in HICO-DET [3], 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:



(a) Proportion of correct ITE (ground truth $\beta_q = 0$).

(b) Proportion of correct ITE (ground truth $\beta_q = 1$).

Figure 31: ITE given different $[\alpha_p, \beta_q]$.

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

(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 [21, 10]). 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 $P'_{HOI} = (P_\beta * A_{h+sp} + P_{HOI})/2$.

Concretely, we use OCRN to enhance HOI detection models (iCAN [10], TIN [21], IDN [17]) on HICO-

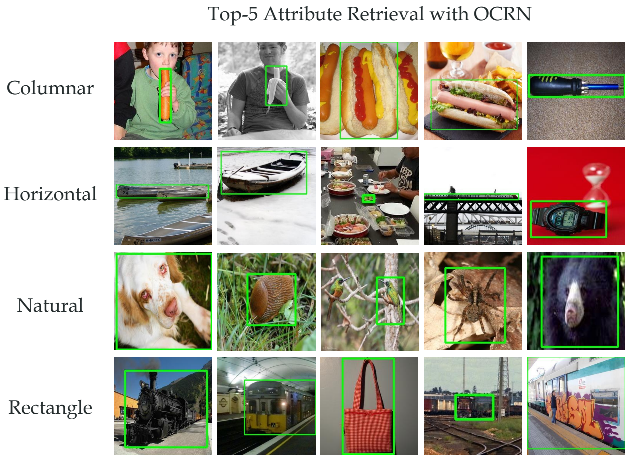


Figure 32: Top-5 attribute retrievals on the OCL test set.

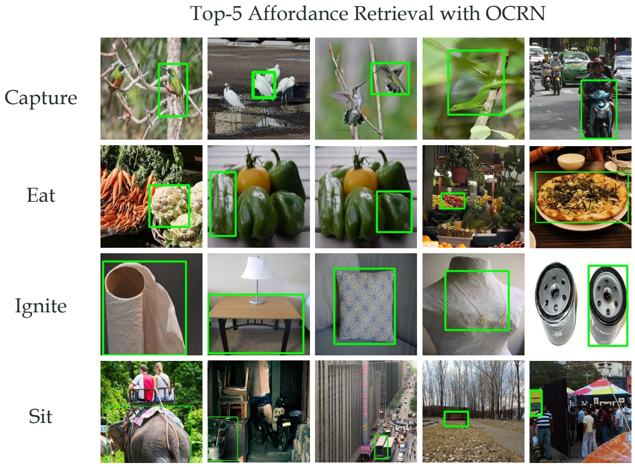


Figure 33: Top-5 affordance retrievals on the OCL test set.

DET [3]. As OCL merely contains 15 object categories in HICO-DET [3], 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 [10, 21, 17] are used.

The results are shown in Tab. 3. With OCL filtering, iCAN [10], TIN [21], and IDN [17] 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.

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

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

Table 4: Comparison with debiasing models.

9. Comparison on Imbalance Learning

9.1. 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 pixels to labels, and to avoid the bias estimation such as in the **Simpson's paradox** [34]. Thus, we use intervention to deconfound the confounder *category* and exclusive the possible spurious bias and correlation imported bias from imbalanced 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 [39, 37, 38]. 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 [40] on our strong baseline DM-V to reduce bias from object categories. We use **mean bias amplification** (**Amp**) in [43] as bias evaluation metric: small Amp means model suffers less from data category bias. The test results are shown in Tab. 5. 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 [43], 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

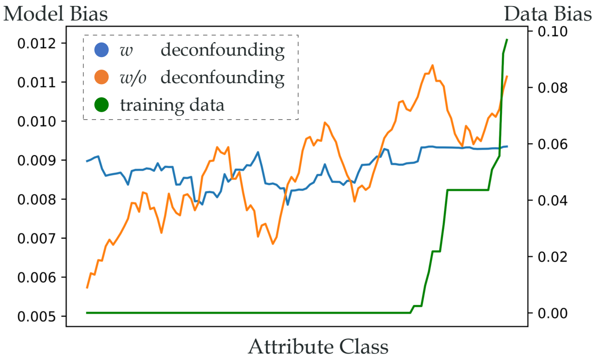


Figure 34: Attribute bias (w/ and w/o deconfounding) for category `frying pan`.

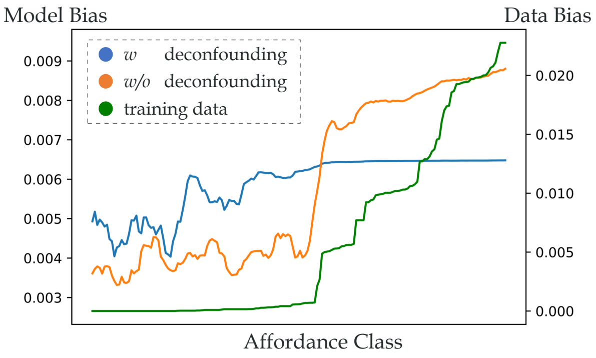


Figure 35: Affordance bias (w/ and w/o deconfounding) for category `giraffe`.

bias of β is measured in the same manner. Fig. 34 and 35 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.

9.2. 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 [2] and focal loss [22]. 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.

Method	α	β
DM-$\alpha \rightarrow \beta$	28.7%	52.6%
DM-$\alpha \rightarrow \beta$+Class balance sampling [2]	27.3%	52.1%
DM-$\alpha \rightarrow \beta$+Focal loss [22]	27.6%	51.2%

Table 5: Comparison with debiasing models.

10. 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. 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.

11. 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 experts and crowdsourcing annotators. Some previous works also follow this before designing the method, such as [45]. From the viewpoint of causal discovery [34, 37, 38, 39], 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 "**eat-able**" (**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.* [5], 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 *normality* 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.

12. Detailed Lists

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

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. 9

[2] 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. 17

[3] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018. 1, 14, 15, 16

[4] Yu-Wei Chao, Zhan Wang, Rada Mihalcea, and Jia Deng. Mining semantic affordances of visual object categories. In *CVPR*, 2015. 1

[5] Patricia W Cheng and Laura R Novick. Causes versus enabling conditions. *Cognition*, 40(1-2):83–120, 1991. 18

[6] 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. 7

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 9

[8] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 1, 7

[9] Christiane Fellbaum. Wordnet. *The encyclopedia of applied linguistics*, 2012. 1

[10] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. In *BMVC*, 2018. 15, 16

[11] 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. 1

[12] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 1, 15

[13] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *TIP*, 2019. 4

[14] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *CVPR*, 2015. 2

[15] 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. 1

[16] 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. 14

[17] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, and Cewu Lu. Hoi analysis: Integrating and decomposing human-object interaction. In *NeurIPS*, 2020. 15, 16

[18] 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. 15

[19] 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. 15

[20] 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. 14

[21] 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. 15, 16

[22] 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. 17

[23] 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. 7, 12

[24] 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. 10

[25] H. Liu, R. Wang, S. Shan, and X. Chen. Learning multifunctional binary codes for both category and attribute oriented retrieval tasks. In *CVPR*, 2017. 1

[26] Xinpeng Liu, Yong-Lu Li, and Cewu Lu. Highlighting object category immunity for the generalization of human-object interaction detection. In *AAAI 2022*, 2022. 15

[27] 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. 15

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin trans-

former: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 12, 15

[29] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008. 6

[30] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 9

[31] 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. 1

[32] Daniel N Osherson, Joshua Stern, Ormond Wilkie, Michael Stob, and Edward E Smith. Default probability. *Cognitive Science*, 15(2):251–269, 1991. 1

[33] Genevieve Patterson and James Hays. Coco attributes: Attributes for people, animals, and objects. In *ECCV*, 2016. 1

[34] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016. 4, 16, 17

[35] Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine learning*, 62(1-2):107–136, 2006. 10

[36] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005. 7

[37] 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. 9, 16, 17

[38] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *CVPR*, 2020. 9, 11, 16, 17

[39] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *CVPR*, 2020. 9, 16, 17

[40] 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. 16

[41] 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. 15

[42] 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, 7

[43] 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. 16

[44] Tinghui Zhou, Hanhuai Shan, Arindam Banerjee, and Guillermo Sapiro. Kernelized probabilistic matrix factorization: Exploiting graphs and side information. In *SDM*, 2012. 9

[45] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *ECCV*, 2014. 1, 10, 17