Deciphering Object Concepts: Hierarchical Cross-Modal Relational Reasoning for Mining Object-Attribute-Affordance Associations

Yuxuan Wang, Muli Yang, Yihang Zhu, Jiexi Yan, Fen Fang,
Xulei Yang, *Senior Member, IEEE*, and Cheng Deng, *Senior Member, IEEE*

Abstract—Object Concept Learning (OCL) aims to recognize high-level attributes and affordances of objects and to infer the causal relationships between them. The key is to accurately model the many-to-many mapping between objects and concepts: While an object may possess multiple concepts, a concept can also belong to multiple objects. Existing methods primarily rely on attention mechanisms to capture label correlations, which limits their ability to comprehend high-level concepts and to perform effective causal reasoning. Inspired by the human cognitive process of progressive understanding, a Hierarchical Cross-Modal Relational Reasoning (CORE) framework is proposed to enhance the understanding of object concepts through hierarchical interaction and reasoning between visual and textual modalities. Specifically, a coarse-to-fine relational reasoning module is developed, in which multi-step learnable prompts are employed to progressively localize the conceptual information of objects, thereby improving the accuracy of object-concept mapping. Subsequently, to facilitate the modeling of causal relationships between object attributes and affordances, a counterfactual reasoning mechanism is introduced. By constructing counterfactual samples and distinguishing the predictive outputs of factual and counterfactual parts, the model's ability to capture causality among concepts is enhanced. Significant performance gains and extensive visualization analysis demonstrate the superiority of our method.

Index Terms—Multi-Label Image Recognition, Object Concept Learning, Cross-Modal Reasoning, Prompt Learning.

1 Introduction

The development of deep neural networks and their wide applications, such as object recognition [1], [2], [3], have achieved many progresses. Understanding object concepts as human beings is an essential pursuit of artificial intelligence. Most existing approaches [4], [5] leverage specialized neural architectures to extract discriminative representations and establish mappings between these representations and their corresponding categories. However, such methods often fall short in enabling models to comprehend higher-level concepts, such as attributes and affordances (*e.g.*, furry, pourability), and they are typically sensitive to changes in environmental conditions. To overcome these limitations, the challenging task of Object Concept Learning (OCL) [6] has been introduced, which aims to reason out the attributes and affordances associated with the object and reveal the causality between attribute and affordance. It is worth noting that affordance learning has been studied from different perspectives in computer vision and robotics. In robotics and embodied settings, affordances are often modeled as task- or interaction-conditioned properties that depend on an agent's behavior and the surrounding environment. By contrast, OCL follows a more object-centric visual formulation, where attributes and affordances are inferred directly from visual observations [7]. In this way, visual affordances—such as graspability or cuttability—can be predicted from geometric and physical cues even in the absence of explicit interaction signals or action annotations. Importantly, this affordance notion is not contradictory to the task-aware definition, but rather represents a perceptual prior for downstream affordance reasoning.

A common approach to OCL is to follow traditional object classification paradigms to mine discriminative representations corresponding to object-related attributes and affordances. For example, the OCRN model [6] aims to transfer category-level knowledge of attributes and affordances to the instance level, thereby establishing mappings between objects and their corresponding attribute and affordance labels. However, the mappings between objects and concepts cannot be fully captured through category-level associations alone. For instance, as shown in Fig. 1, a single object often exhibits multiple attributes and affordances simultaneously (*e.g.*, a cake can be both "round" and "hold"), while multiple objects may share the same attributes and affordances (*e.g.*, "round" and "hold" can describe cake, vegetable, and bowl). Thus, for OCL, it is necessary that the model is capable of accurately capturing many-to-many relationships between objects and higher-level attributes and affordances. Existing multi-label image recognition methods [8], [9], [10], as shown in Fig. 2(a), typically rely on attention mechanisms to model the correlations between image regions and labels, as well as to capture

This work was supported in part by the National Key R&D Program of China (No. 2023YFC3305600) and the National Natural Science Foundation of China (U25B2048, 62132016).(Corresponding author: Cheng Deng.)

- *Yuxuan Wang, Yihang Zhu, and Jiexi Yan are with the School of Electronic Engineering, Xidian University, Xi'an 710071, China. (E-mail: {wangyuxuan, zhuyihang, jxyan}@stu.xidian.edu.cn)*
- *Muli Yang, Fen Fang, and Xulei Yang are with the Institute for Infocomm Research (I²R), A*STAR, Singapore. (E-mail: {yangml, fangf, yangx}@a-star.edu.sg)*
- *Cheng Deng is with the School of Electronic Engineering, Xidian University, Xi'an 710071, China. (E-mail: chdeng.xd@gmail.com)*

label dependencies. However, while attention mechanisms are effective in extracting label correlations, their ability to understand high-level object concepts remains limited. To accurately model the many-to-many mappings between objects and concepts and to infer the causal relationships among concepts, a deeper understanding of the object semantics and contextual information is required.

In recent years, vision-language models (VLMs) such as CLIP [11] and Qwen [12] have demonstrated great potential in understanding object information [13] and addressing many-to-many mapping problems [14]. This potential can be attributed to the rich knowledge acquired through large-scale image-text pretraining, which enables the association of visual content with deeper semantic concepts. To leverage such foundation models for recognizing object concepts, several tuning approaches [15], [16], [17], [18] have been proposed. These methods enhance the representational capacity of pretrained networks by learning a set of continuous vectors and injecting them into the input space, thereby facilitating the understanding of label correlations within images. Nevertheless, when multiple object concepts are present in an image, the concept information extracted by the visual encoder from global representations often becomes entangled. As a result, directly using concept names as textual inputs usually fails to yield accurate predictions of the corresponding visual concepts, leading to degraded performance. Recent studies [19], [20], [21], [22], [23] have shown that humans employ hierarchical and counterfactual strategies to solve complex multi-step problems. Inspired by these findings, we explore simulating human cognitive processes by introducing a learnable coarse-to-fine multi-step prompting combined with counterfactual reasoning to address the many-to-many mapping problem.

Specifically, a Hierarchical **C**ross-M**o**dal **Re**lational Reasoning (CORE) approach is proposed, mainly consisting of a coarse-to-fine relational reasoning (CRR) and a counterfactual relation enhancing (CRE) module. To effectively capture the complex relationships between concepts and objects and to improve the accuracy of concept and object mapping, the CRR module is introduced. For objects in an image, a given attribute or affordance may be associated with multiple instances, indicating that a concept can exist across different objects. To enhance the accuracy of concept-object mapping, a category-agnostic prompting is first designed. The CRR module then operates in two stages: in the first stage, global visual semantic information is utilized to refine the learnable prompts, aligning text embeddings with the overall visual context; in the second stage, local instance features are exploited to further adjust the prompt representations, progressively guiding the embeddings toward the target instance concepts and thereby enabling precise object-concept mappings. Building on this, we further introduce the CRE module to move beyond correlations and uncover causal dependencies among concepts. CRE constructs counterfactual samples and distinguishes the predictive outputs of factual and counterfactual parts to discard spurious correlations, thereby revealing the causal relationships between attributes and affordances. This causal reasoning not only strengthens the semantic alignment between objects and their corresponding concepts but also leads to significant performance improvements.

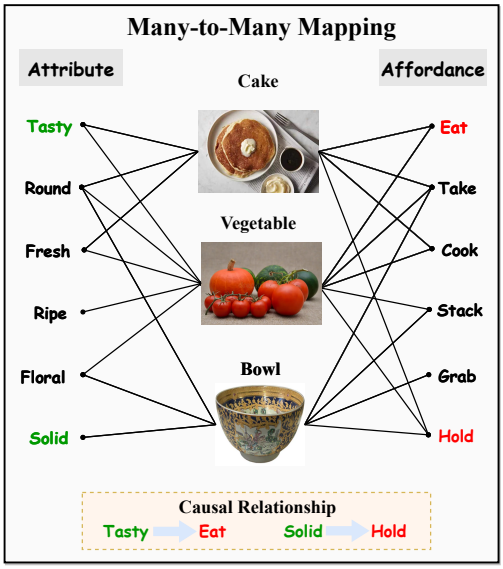


Fig. 1. The many-to-many issue in Object Concept Learning (OCL). Embodied agents require a deep understanding of objects, which goes beyond recognizing their categories to also identifying their attributes and affordances. To this end, the OCL task has been introduced. The core challenge of OCL lies in constructing accurate many-to-many mappings between objects and concepts: an object may possess multiple attributes and affordances, while a single attribute or affordance may apply to multiple objects. Moreover, OCL further demands reasoning over causal relationships among attributes and affordances to achieve a deeper level of object understanding.

The contributions are summarized as follows:

(1) We first formulate OCL as a many-to-many mapping problem. To this end, we propose a novel Hierarchical Cross-Modal Relational Reasoning (CORE) approach, which constructs semantic-aware object-concept mappings and incorporates counterfactual reasoning, thereby significantly enhancing the model's reasoning accuracy.

(2) To precisely localize concepts associated with objects, a coarse-to-fine relational reasoning (CRR) module is introduced. CRR conducts reasoning that progressively integrates visual and textual information, thereby enabling more effective identification of object attributes and affordances, achieving precise concept localization.

(3) To capture the causal relationships between attributes and affordances, a counterfactual relation enhancing (CRE) module is designed. By incorporating counterfactual reasoning, CRE learns how affordance features change when attribute states are altered, thereby strengthening the model's causal reasoning capability.

(4) Extensive experimental results and visualization analyses demonstrate the effectiveness of our approach. Notably, compared with the state-of-the-art method [6], our method achieves improvements of **8.0** mAP and **4.0** mAP in attribute and affordance prediction, respectively.

2 Related Work

2.1 Attribute and Affordance Recognition

Attribute recognition shares common background with other popular topics in research such as object detection [24], image segmentation [25] and classification [5]. It typically

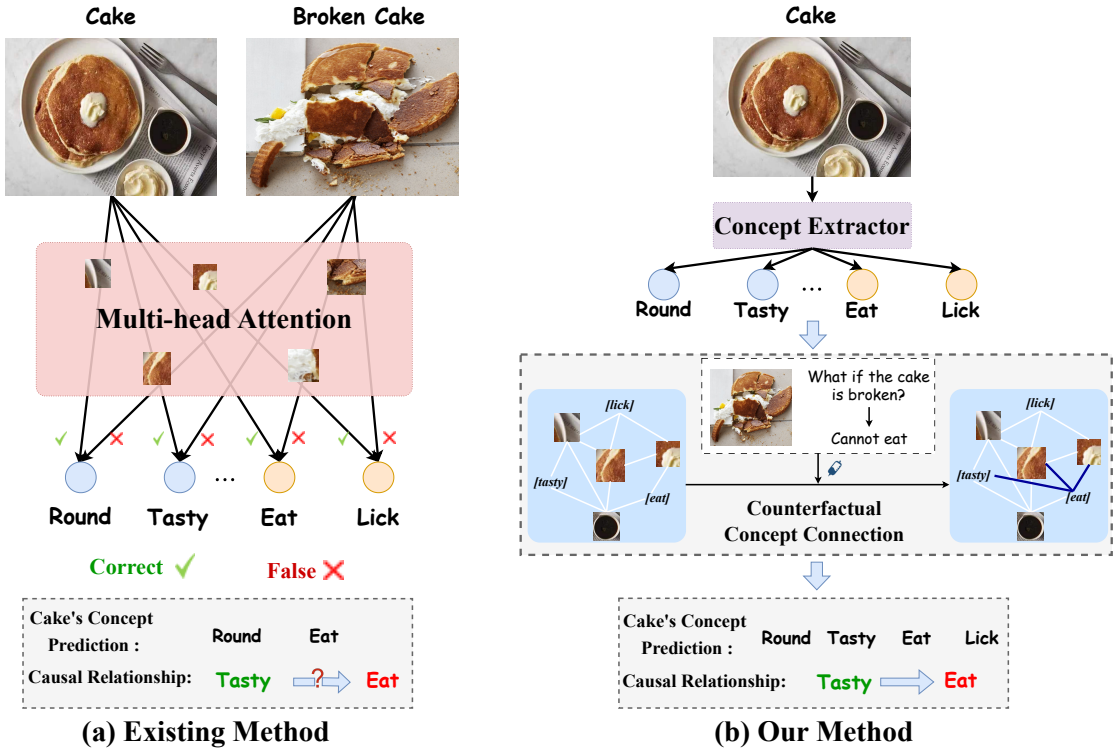


Fig. 2. Comparison between existing method and our proposed approach. (a) Existing methods typically rely on multi-head attention mechanisms, which model the correlations between image regional features and labels to recognize the multi-label content of objects. However, such methods primarily emphasize label correlations and fall short in capturing a deeper understanding of high-level object concepts. (b) We explore designing a hierarchical relational reasoning method. We first exploit a series of learnable coarse and fine prompts to progressively focus on concept-relevant information. Then, multiple counterfactual samples are selected to strengthen the causal relationships between attributes and affordances, which improves the accuracy of the learned object concepts.

plays the role of mediator between pixels and higher-level concepts. However, visual attribute recognition has its unique challenges that set it apart from other visual tasks. The examples of these challenges include the large number of attributes that need to be predicted, and there exist many-to-many mapping rules between attributes and categories. Consequently, for attribute recognition, besides classifying the attribute directly, other methods incorporate both global and local information [26] or excavate the intrinsic properties [27] to enhance the performance.

Affordance recognition [28], aiming to reason about the objects' affordances in a scene through the input, leads to multifaceted applications in scene understanding [29], human-object interaction [30], and so on. Most traditional affordance recognition methods rely on a Bayesian network [31] or Support Vector Machine [32] to encode the dependencies between the object's global features and the affordance characteristics [33], [34]. Deep learning-based methods learn the information of different modalities as prior knowledge and compensate for them with the traditional methods to improve accuracy [28]. For instance, Pinto et al. [35] utilized the multi-stage learning approach to collect affordances. Dadure et al. [36] discuss knowledge representation and reasoning the target object itself. However, these methods often focus solely on affordance recognition. In practical scenarios, people often infer affordance based on observed attributes. For example, if we need to drink water but do not have a cup, we may find another hollow, hard object to hold the water. This reflects the importance of perceiving the relationship between attributes and affordances. Our work is built upon this visual affordance perspective: rather than modeling full task-conditioned interactions, we focus on understanding object concepts through attribute recognition, affordance recognition, and the structured dependencies between them. Effectively addressing OCL from this perspective holds potential for promoting the development of embodied artificial intelligence.

2.2 Multimodal Prompting Methods

Recently, many works [37], [38], [39], [40] have focused on prompting large pre-trained vision-language models to adapt to specific downstream tasks. The key idea of prompt engineering is to provide hints and other textual information to guide the pre-trained model in leveraging its existing knowledge to solve new tasks. The hints can be in the form of continuous vector representations, referred to as prompt tuning [37]. This approach directly optimizes prompts within the embedding space of the model. The related work, such as [41], uses prompt tuning to improve the adaptation of pre-trained Vision Transformers to image and video understanding tasks. Additionally, CoOp [42] introduces prompt tuning for visual tasks. They achieve this by converting context words into a set of learnable vectors to adapt them to the pre-trained vision-language model. CoCoOp [43] further transforms static prompts into dynamic prompts to better handle category shifts. Chain-of-Thought Prompting [15] is a method to prompt the model

by adding a series of intermediate reasoning steps. Each prompt in the chain incorporates contextual information, enabling the model to generate more coherent and contextually appropriate responses. Inspired by the progressive nature of human cognitive understanding, we explore hierarchical cross-modal relational reasoning as a bridge to leverage the powerful capabilities of large pre-trained models, thereby addressing the reasoning challenges in OCL.

2.3 Cross-Modal Causal Reasoning

Recent studies have explored cross-modal causal reasoning to mitigate spurious correlations in visual–language learning. For example, Liu et al. [23] introduce causal relational reasoning for event-level VideoQA by applying front-door and back-door interventions to disentangle visual and linguistic biases. Wei et al. [44] further propose visual causal scene refinement to identify key causal segments that support reliable answers in video question answering. In the context of video grounding, Chen et al. [45] develop a cross-modal causal relation alignment framework that jointly improves answer prediction and temporal grounding consistency. Beyond video understanding, cross-modal causal representation learning has also been explored for radiology report generation [46], where causal interventions are used to reduce visual–language confounders. These works mainly focus on event-level reasoning or multimodal generation, where causal modeling improves vision–language alignment for downstream tasks. While these aforementioned studies focus on event-level or scene-level causalities, we specifically investigate the triadic associations among objects, attributes, and affordances. By framing these associations within CORE, we bridge the gap between low-level visual perception and high-level object understanding, ensuring that the identified attributes and affordances are grounded in intrinsic object properties rather than coincidental context.

3 METHODOLOGY

Fig. 3 shows the framework of the CORE model. Our work focuses on how to enhance the model's reasoning ability to understand object concepts. In this section, we present our method including coarse-to-fine relational reasoning and counterfactual relation enhancing modules.

3.1 Coarse-to-Fine Relational Reasoning

Compared with concrete object categories, e.g., cake, object concepts are much more abstract and involves plentiful information. To deepen the understanding of object concepts, we explore imitating human beings to perform coarse-to-fine reasoning to progressively localize concept-relevant object content, which improves the OCL performance.

3.1.1 Coarse-Grained Relational Reasoning

The goal of coarse-grained relational reasoning is to create continuous vector representations as input prompts (see *Category-Agnostic Prompting*) and obtain more accurate attribute and affordance semantic descriptions through integrating global visual contexts and prompts (see *Contextual Relational Reasoning*), which is helpful for a thorough understanding of visual content.

Category-Agnostic Prompting. To adapt the large pre-trained vision-language model to the downstream recognition tasks, a common way is to use text prompt templates in CLIP, like "a photo of a [cls]", which primarily focuses on category semantics. Nevertheless, utilizing such hard text prompt templates presents challenges in generating generic attribute and affordance textual embeddings. This is because the original pretraining of CLIP focused on aligning with categorical semantics rather than high-level attributes and affordances concepts of images. To overcome this limitation, we aim to construct a set of learnable text prompts incorporating the prior knowledge of concepts. Recent works [6], [7] reveal that the attributes and affordances are shared across objects. For objects in an image, an attribute and affordance could belong to multiple different objects, which means concepts exist across categories. Consequently, we construct a category-agnostic model and optimize prompts focusing on enhancing concepts-object mapping precision. We employ the prompt tuning [37] to construct a set of learnable text prompts h incorporating the knowledge of attributes and affordances as:

$$\begin{aligned} h_\alpha &= [T_1]\,[T_2]\ldots[T_n]\,[\text{is}][\text{attribute}], \\ h_\beta &= [P_1]\,[P_2]\ldots[P_n]\,[\text{afford to}][\text{affordance}], \end{aligned} \tag{1}$$

where $[T_i]$ and $[P_i](i \in 1,\ldots,n)$ are learnable token embeddings in attribute and affordance text prompt templates, respectively. These learnable tokens adapt the prompt representation to the downstream object concept learning task while preserving the semantic structure introduced by the fixed tokens. The tokens "[is]" and "[afford to]" are used as fixed semantic anchors to explicitly distinguish attribute descriptions from affordance descriptions. This design helps the model interpret attribute concepts as object properties (*e.g.*,"... is metal") and affordance concepts as potential actions (*e.g.*, "... afford to hold"). This design ensures a category-agnostic prompt template to learn the shared patterns of different categories.

Contextual Relational Reasoning. Since the nearby environment affects the recognition of attributes and affordances [7], we devise a contextual relational reasoning approach that uses global coarse-grained visual contexts to reason and optimize the prompt features, making the generated textual embeddings capable of aligning visual content. Specifically, given an input image X_i, we extract the visual embedding $F_g \in \mathbb{R}^d$ from CLIP visual encoder $v(\cdot)$ and feature $F_M \in \mathbb{R}^{p\times d}$ from the M-th intermediate layer of CLIP visual encoder following [47], where p and d separately denote the number of patches and feature dimension. And the input text h_α and h_β are sent to the text encoder $f(\cdot)$, obtaining the text feature embeddings $H_\alpha \in \mathbb{R}^{N_\alpha\times d}$ and $H_\beta \in \mathbb{R}^{N_\beta\times d}$. N_α is the number of attributes and N_β is the number of affordances. To encourage the text features to align with related visual elements, we design the context decoder, where the features F_M are used as the keys and values, and the text features H_α and H_β are used as the queries:

$$\begin{aligned} \hat{H}_\alpha &= \Theta_g(\Phi_g(H_\alpha), F_M) + H_\alpha, \\ \hat{H}_\beta &= \Theta_g(\Phi_g(H_\beta), F_M) + H_\beta, \end{aligned} \tag{2}$$

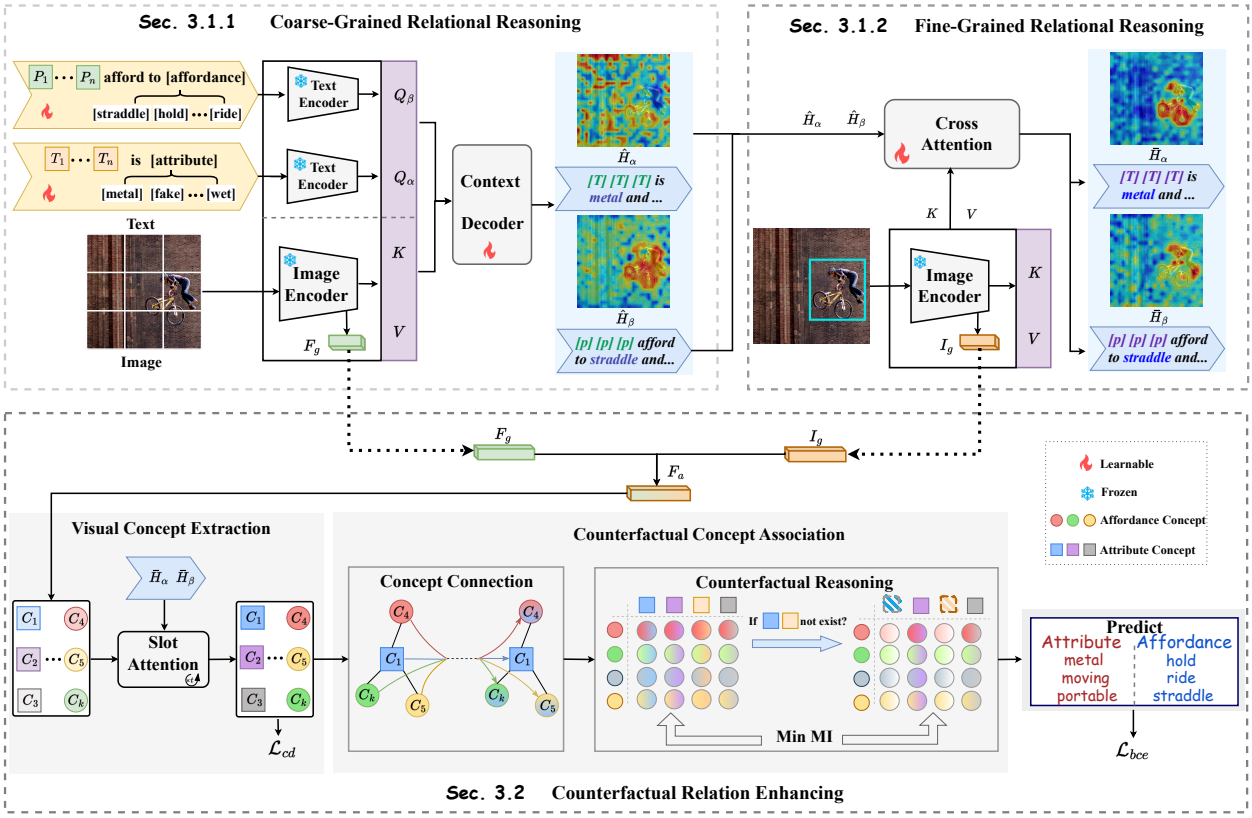


Fig. 3. The details of Hierarchical Cross-Modal Relational Reasoning (CORE). This method mainly consists of two components: coarse-to-fine relational reasoning (CRR) and counterfactual relation enhancing (CRE). Concretely, CRR first extracts visual tokens from the global image and generates coarse-grained attribute and affordance prompts. Subsequently, CRR refines the prompts by combining coarse-grained text prompts with localized visual information. The heatmaps illustrate the prompt learning process from coarse to fine. Finally, we design the CRE to further build accurate relations among objects and concepts, which improves the performance of concept prediction and causal inference.

where the $\Phi_g(\cdot)$ and the $\Theta_g(\cdot)$ represent the self-attention and cross-attention operation respectively. The self-attention mechanism allows for focusing on important contextual information for each word while reducing attention to irrelevant information. The cross-attention helps the model understand the semantic alignment between images and language, thereby providing a more accurate and meaningful joint representation. Through the residual connection "+", the language priors from the text features are preserved. Based on this, we can store the extracted text features $\hat{H}_\alpha$ and $\hat{H}_\beta$ with sufficient global visual information.

3.1.2 Fine-Grained Relational Reasoning

As our task is to identify the instance object's attribute and affordance, to further map object-concepts (align the instance visual feature to attributes and affordances prompts), we employ ground-truth bounding boxes to crop the objects and compute visual features $I_g \in \mathbb{R}^{1\times d}$ through the CLIP visual encoder. Then, the text features $\hat{H}_\alpha$ and $\hat{H}_\beta$ defined as the queries are refined with the help of instance visual features I_g defined as keys and values by cross-attention $\Theta_I(\cdot)$ as follows:

$$\bar{H}_\alpha = \Theta_I(\hat{H}_\alpha, I_g), \bar{H}_\beta = \Theta_I(\hat{H}_\beta, I_g), \tag{3}$$

where $\bar{H}_\alpha \in \mathbb{R}^{N_\alpha \times d}, \bar{H}_\beta \in \mathbb{R}^{N_\beta \times d}$. From the above two steps, we first obtain category-agnostic text features with the global coarse-grained visual content, which helps to capture the attribute and affordance semantics comprehensively and enhance concepts-objects mapping precision. Then, the fine-grained prompt is introduced to enable a mapping between object-concepts to concentrate on fine-grained visual contents, which guides the following visual concept reasoning.

3.2 Counterfactual Relation Enhancing

Upon obtaining fine-grained textual prompts, we extract visual features that correspond to the associated concepts. To further reason about the causal relationships among these concepts, we propose a counterfactual concept association network, which explicitly models causal dependencies and enhances reasoning accuracy.

3.2.1 Prompt-Guided Visual Concept Extraction

We define a set of attribute concepts $C_\alpha = \{c_i \in \mathbb{R}^D, i = 1, ..., k\}$ and affordance concepts $C_\beta = \{c_i \in \mathbb{R}^D, i = 1, ..., k\}$, where k denotes the number of concept and D is the dimension of each concept. Each concept $c_i^{(t)}$ is initialized by visual feature F_a and updated through attention and Gated Recurrent Unit (GRU) [48] operation over t iterations, where F_a is aggregated by the global image feature F_g and instance image feature I_g. We project the F_a and the text prompt features $\bar{H}_\alpha$, $\bar{H}_\beta$ dimension to D by nonlinear transformations Q , V and K

respectively. Dot-product is applied to generate an attention matrix $attn^{(t)}$:

$$\begin{aligned} attn_{\alpha}^{(t)} &= \mathrm{Softmax}(\frac{1}{\sqrt{d}}Q(C_{\alpha}^{(t)}) \cdot K(\bar{H}_{\alpha})), \\ attn_{\beta}^{(t)} &= \mathrm{Softmax}(\frac{1}{\sqrt{d}}Q(C_{\beta}^{(t)}) \cdot K(\bar{H}_{\beta})), \end{aligned} \tag{4}$$

where attention matrix $attn_{\alpha} \in \mathbb{R}^{k \times N_{\alpha}}$, $attn_{\beta} \in \mathbb{R}^{k \times N_{\beta}}$.

To aggregate the input values V to their assigned concepts, we use cross product operation and get the updates feature $U_{\alpha}^{(t)}$ and $U_{\beta}^{(t)}$:

$$U_{\alpha}^{(t)} = attn_{\alpha}^{(t)} \cdot V(\bar{H}_{\alpha}), U_{\beta}^{(t)} = attn_{\beta}^{(t)} \cdot V(\bar{H}_{\beta}), \tag{5}$$

where the aggregated updates feature $U_{\alpha}^{(t)}, U_{\beta}^{(t)} \in \mathbb{R}^{k \times D}$. The concept code C_{α} and C_{β} are eventually updated with a GRU as $c_i^{(t)} = \mathrm{GRU}\left(c_i^{(t-1)}, U^{(t)}\right)$, separately. In our experiment, the concepts are updated for $t = 3$ times.

3.2.2 Counterfactual Concept Association Network

As is shown in Fig. 3, the object rider affords to ride and hold because the bicycle is moving and portable. Obviously, there are causal relationships between some attributes and affordances. In order to enable the model not only recognizing these concepts but also learning the causal relationships between specific concepts, we design the concept association network with counterfactual, including *Concept Connection Network* and *Counterfactual Reasoning*.

Concept Connection Network. After obtaining the concepts $C_{\alpha} \in \mathbb{R}^{k \times D}$ and $C_{\beta} \in \mathbb{R}^{k \times D}$, we construct connections among concepts to reason out the specific attributes and affordances labels. Specifically, we seek to model an undirected attribute-affordance graph $G_a = \{V, \xi, \mathbf{A}\}$, where ξ is the set of graph edges to learn and $\mathbf{A} \in \mathbb{R}^{k \times k}$ is the corresponding adjacency matrix. Each node $\nu_{\alpha}, \nu_{\beta} \in V$ corresponds to one element of the visual concept C_{α} and C_{β}. The size of V is set to $2k$. We do not explicitly instantiate a full $2k \times 2k$ adjacency matrix. Instead, we focus on modeling attribute-to-affordance relations, which align with the causal relationships of interest in the OCL setting. We define an adjacency matrix for the graph as $\mathbf{A} = \mathrm{softmax}_c\left(C_{\alpha}C_{\beta}^T\right) + I_d$, where I_d indicates the identity matrix and softmaxc indicates we make softmax operation across the column direction.

$$M = \mathbf{A}C_{\beta}, \quad \widetilde{M} = \tanh\left(w_f^c * M + b_f^c\right), \tag{6}$$

where $w_f^c \in \mathbb{R}^{k \times k}$, $b_f^c \in \mathbb{R}^D$ indicate the trainable parameters. $\widetilde{M} \in \mathbb{R}^{k \times D}$ is the output of the concept connection network. "$*$" indicates the multiplication operation. Each row of the affordance matrix M represents a feature vector of a node, which is a weighted sum of the neighboring node features of the current node.

Subsequently, we design a fusion operation to obtain the attribute and affordance classification feature. The fusion feature $\bar{F} \in \mathbb{R}^k$ is obtained by taking the dot-product of feature F_a and matrix $\widetilde{M}$. The Softmax function $\phi(.)$ is used to generate a probability simplex over the $\bar{F}$, *i.e.*, $\phi(\bar{F}) = [p_i]_{i=1}^k$. Next, the affordance concept representation F_{β} is derived by using a convex combination of the affordance features $\widetilde{M}$ weighted by their corresponding p_i, *i.e.*, $F_{\beta} = \sum_{i=1}^k p_i \cdot \widetilde{M}$. The attribute concept features F_{α} is getting from the mean operation of the concepts C_{α}.

Counterfactual Reasoning. After obtaining attribute and affordance concepts, relying solely on the CRR module may lead to spurious associations, where non-causal attributes are incorrectly linked to affordances. In the OCL task, we consider a causal structure: $O \to \alpha$, $O \to \beta$, $\alpha \to \beta$, where the O acts as confounders that induce correlations between attributes α and affordances β. To mitigate the adverse effects of confounders, a widely adopted strategy is the Total Direct Effect (TDE) [49]. It removes the unwanted effect that will lead to shortcut bias in learning and inference, by subtracting the unwanted predictions and retaining the wanted ones. To achieve this, we introduce a counterfactual reasoning module to capture the underlying causal relationships between attributes and affordances. We implement interventions via attribute masking [50] and reveal causal dependencies by minimizing the mutual information (MI) between affordance representations before and after intervention, cutting off the link $O \to \alpha$ completely. The core intuition is that if an attribute is a genuine causal factor of a particular affordance, intervening on this attribute should induce substantial changes in the corresponding affordance representation. In such cases, a clear discrepancy will emerge between the factual and counterfactual representations, accompanied by a significant reduction in their mutual information.

To realize this intervention mechanism, we follow the strategy adopted in OCL and perform masking-based interventions on attribute representations that correspond to causal factors. The masked attribute text prompt embedding is formulated as $\bar{H}_{\alpha mask} = \bar{H}_{\alpha} \odot Mask$, where $Mask$ is generated following the causal attribute–affordance pairs provided in the OCL dataset [6] and apply masking only to these attributes. Then, we send the masked prompts to the visual concept extraction module (3.2.1) and obtain the masked attributes $F_{\alpha mask}$ and counterfactual affordance results $F_{\beta mask}$ from the concept connection network. When the $F_{\alpha mask}$ are cause factors of the affordance, masking them effectively severs the causal links between these attributes and the affordance. This leads to a significant shift in the distribution of affordance features, resulting in a large discrepancy between F_{β} and $F_{\beta mask}$. Therefore, to obtain causality between attribute and affordance, the correlation between the predicted affordance features $F_{\beta mask}$ and the original affordance features F_{β} should be minimized. As training progresses, the model gradually discovers the underlying associations between certain attributes and specific affordances through this optimization process.

Specifically, mutual information (MI), denoted as:

$$I\left(F_{\beta}; F_{\beta mask}\right) = \mathbb{E}_{p(F_{\beta}, F_{\beta mask})}\left[\log \frac{p\left(F_{\beta mask} \mid F_{\beta}\right)}{p\left(F_{\beta mask}\right)}\right], \tag{7}$$

is used to quantify the dependency between F_{β} and $F_{\beta mask}$. Unfortunately, the conditional distribution $p\left(F_{\beta mask} \mid F_{\beta}\right)$ are not directly accessible in our task. Therefore, we employ a variational distribution $q_{\theta}\left(F_{\beta mask} \mid F_{\beta}\right)$ to approximate the true posterior, which is implemented with neural net-

works. Then, the MI is extended to $\hat{I}\left(F_\beta; F_{\beta mask}\right)$:

$$\hat{I}\left(F_\beta; F_{\beta mask}\right) = \mathbb{E}_{p(F_\beta, F_{\beta mask})}\left[\log q_\theta(F_{\beta mask} \mid F_\beta)\right] - \mathbb{E}_{p(F_\beta)}\mathbb{E}_{p(F_{\beta mask})}\left[\log q_\theta(F_{\beta mask} \mid F_\beta)\right]. \tag{8}$$

By minimizing Eq.8, the intervened affordance distribution $F_{\beta mask}$ and the original affordance distribution F_β are enforced to be independent. This independence compels the model to explicitly account for the contribution of masked attributes to affordance prediction, thereby capturing their underlying relationships. Considering $\hat{I}\left(F_\beta; F_{\beta mask}\right)$ is no longer an upper bound for MI, we follow the log-likelihood function in CLUB [51] and minimize the negative log-likelihod between F_{β_i} and $F_{\beta mask_i}$:

$$\mathcal{L}_{\text{likel}} = -\frac{1}{N}\sum_{i=1}^{N}\log q_\theta\left(F_{\beta mask_i} \mid F_{\beta_i}\right), \tag{9}$$

where N is the number of batch size. Therefore, we make any two dimensions of the affordance representations independent of each other. Based on the above operation, we connect the attribute features C_α with the affordance features C_β. To promise the final predction results, we consider the following optimization strategy.

3.3 Optimization

The attribute and affordance features F_α and F_β are sent to the different classifier, obtaining the predicted probability $\hat{y}_\alpha$ and $\hat{y}_\beta$ and calculating binary cross-entropy losses:

$$\mathcal{L}_{bce} = BCE\left(y_\alpha, \hat{y}_\alpha\right) + BCE\left(y_\beta, \hat{y}_\beta\right), \tag{10}$$

where y_α and y_β are the attribute and affordance label. To capture diverse visual features, different concepts should correspond to distinct image regions. Therefore, each concept is enforced to keep it far from any other concept. We define a concept distinctiveness loss to achieve as:

$$\mathcal{L}_{\text{cd}} = \frac{1}{k(k-1)}\sum_{i,j}^{k}\frac{\langle U_i, U_j\rangle}{\|U_i\|_2^2\,\|U_j\|_2^2}, \tag{11}$$

where $\|\cdot\|_2$ denotes L2-norm and $\langle\cdot,\cdot\rangle$ denotes the inner product operation. U_i means the i-th updated concept feature, U_j represents any other updated feature different from U_i. In this way, the concepts can capture different aspects of the image. The final loss function is $\mathcal{L}_{total} = \mathcal{L}_{bce} + \lambda_1\mathcal{L}_{cd} + \lambda_2\mathcal{L}_{likel}$. In the experiment, the $\lambda_1 = 0.1$ and $\lambda_2 = 1$. For clarity, we summarize the main notations used in this paper in Table 14 to facilitate readability.

4 Experiments

We evaluate our method on the OCL datasets. We demonstrate that CORE improves attributes and affordances recognition performance and effectively enhances the causal effect between them. We also conduct experiments on Multi-Task Indoor Scene Understanding and Open-Vocabulary Object Attribute Recognition tasks to demonstrate that CORE can also perform well.

4.1 Datasets and Baselines.

We consider three different tasks ranging from object concept learning, multi-task indoor scene understanding and open-vocabulary object attribute recognition.

Object Concept Learning. We consider OCL [6] dataset, which is the first attribute-affordance reasoning dataset comprising 185,941 instances of 381 categories, 114 attributes, and 170 affordances. The state-of-the-art competing methods include DM [6], HMa [52], Attention [4], DM-att [6], OCRN [6], CLIP [11] and Qwen2-VL [12].

Multi-Task Indoor Scene Understanding. We consider NYUd2 [53] dataset including 1449 RGB-D images of indoor scenes with 40 object categories, 5 affordances and 11 attributes labels. The SOTA competing methods include PSPNet [54], FastFCN [55], DeepLab V3 [56], VarReg [57] and Cerberus [58].

Open-Vocabulary Object Attribute Recognition. We also evaluate our method on the OVAD [59] and VAW [60] datasets. For OVAD, it introduces a clean and densely annotated attribute evaluation benchmark for open-vocabulary attribute detection. This benchmark defines 117 attribute categories, encompassing over 14,300 object instances. VAW is constructed based on VGPhraseCut [61] and GQA [62]. VAW contains a rich vocabulary of 620 attributes, including properties such as color, material, shape, size, texture, and action. Each object instance is annotated with positive, negative, and missing attribute labels. We compare our approach with several state-of-the-art methods, including CLIP [11], Open CLIP [63], OvarNet [64], OV-DAR [65] and OVA_t^{II} [65].

4.2 Implementation Details

We use the CLIP model VIT-L/14@336px as our backbone. The length of learnable attribute and affordance prompt embeddings n is set to 12 and the CLIP visual encoder parameters are frozen. For OCL, the concept number k in the best experiment results is 10. The ablation experiments setting are based on the $k = 10$. The model learns with batch size 128. We adopt SGD with learning rate 1, which provides stable optimization when training relatively small parameter sets and is commonly used in similar vision tasks. For NYUd2 benchmark, the counterfactual reasoning cannot be used since there is no causality annotation. Thus, we add our concept extraction module to the baseline network. The experiments use the standard SGD optimizer with a learning rate of 7e-3, momentum 0.9, and batch size 2. For VAW and OVAD, the names of the attributes corresponding to the image labels have been added to the prompt template. And we set $k = 30$ and batch size 8. The training involves a larger parameter space and follows the optimization settings widely adopted in prior work based on vision–language models. Therefore, we employ AdamW optimizer with learning rate 1e-4 for parameter optimization In addition, the metrics in OCL also include the $\mathcal{S}_{\text{ITE}}$ and $\mathcal{S}_{\alpha\text{-}\beta\text{-ITE}}$, which combine the actual affordance probability and counterfactual output following [6] to verify the performance of reasoning. All experiments are conducted with PyTorch-1.10 using two NVIDIA RTX A6000 GPUs. Our method can be considered an independent module that can be flexibly integrated into existing methods. For NYUd2,

TABLE 1
OCL accuracies (mAP). The baselines in the "$\alpha \rightarrow \beta$" fold means α and β are calculated without counterfactual reasoning process. "$\alpha \rightarrow \beta$ w/$\mathcal{L}_{likel}$" means the α and β are calculated with counterfactual reasoning process.

Fold	Method	α	β	$\mathcal{S}_{\text{ITE}}$	$\mathcal{S}_{\alpha\text{-}\beta\text{-ITE}}$	Fold	Method	α	β	$\mathcal{S}_{\text{ITE}}$	$\mathcal{S}_{\alpha\text{-}\beta\text{-ITE}}$
$\alpha \rightarrow \beta$	DM	28.7	52.6	7.6	6.7	$\alpha \rightarrow \beta$ w/$\mathcal{L}_{likel}$	DM	28.8	52.7	7.8	6.9
	Attention	24.1	48.9	8.1	7.1		Attention	24.0	49.0	8.5	7.6
	OCRN	31.6	53.3	9.5	9.2		OCRN	31.6	53.5	9.9	9.5
	CLIP†	33.1	53.7	10.2	9.8		CLIP§	33.5	54.2	10.6	10.0
	Qwen2-VL†	35.9	55.7	10.3	10.1		Qwen2-VL§	36.2	56.0	10.9	10.5
	CORE	**38.2**	**56.6**	**10.5**	**10.1**		CORE	**39.6**	**57.5**	**11.1**	**10.7**

TABLE 2
Quantitative results on NYUd2 for Attribute, Affordance, and Semantic tasks.

Attribute		Affordance		Semantic		
Method	**mIoU (%)**	**Method**	**mIoU (%)**	**Method**	**Input**	**mIoU (%)**
PSPNet	36.7	PSPNet	60.4	FastFCN	RGB	45.4
DeepLab V3	38.1	DeepLab V3	61.4	VarReg	RGB	50.7
Cerberus	45.3	Cerberus	66.3	Cerberus	RGB	50.4
Cerberus + CORE	**46.0**	Cerberus + CORE	**67.2**	Cerberus + CORE	RGB	**50.6**

since the original method involves joint training of attributes and affordances, we can incorporate our method into the original approach by constructing learnable prompts using the labels of attributes and affordances. For VAW and OVAD, we directly use the base and extra datasets to train our method. Additionally, incorporating our module does not alter the optimization process of the original network.

4.3 Main Results

Tables 1, 2, 3 and 4 show the comparison of CORE with the SOTA models on object concept learning, multi-task indoor scene understanding and open-vocabulary object attribute recognition benchmarks. CORE consistently achieves superior performance compared to previous methods.

Object Concept Learning. According to the experiment in [6], we evaluate the affordance (β), attributes (α), $\mathcal{S}_{\text{ITE}}$ and the $\mathcal{S}_{\alpha\text{-}\beta\text{-ITE}}$ performance. The mean Average Precision (mAP) is the evaluation metric for α and β. In addition, we evaluate causal reasoning using the Individual Treatment Effect (ITE) [66], which measures the change in affordance prediction after intervening on an attribute. Based on this value, the benchmark defines S_{ITE} to assess reasoning correctness and $\mathcal{S}_{\alpha\text{-}\beta\text{-ITE}}$ to jointly measure recognition and reasoning performance. We follow the OCL and use S_{ITE} and $\mathcal{S}_{\alpha\text{-}\beta\text{-ITE}}$ as the causal relevant metrics. For fair comparison, we combine the CLIP and Qwen2-VL baseline methods with the classifier, denoted as "CLIP†", "CLIP§", "Qwen2-VL†", and "Qwen2-VL§" in the two folds respectively, to investigate the effect of our method. For the Qwen2-VL, the learnable embeddings are trained with a learning rate of 0.01. As shown in Table 1, our approach outperforms the published OCRN method [6] by **8.0** mAP on attribute and **4.0** mAP on affordance, respectively. Compared with the "Qwen2-VL§", which is a powerful multimodal large model, our method improves the performance, attaining a 3.4 mAP enhancement on attribute and a 1.5 mAP enhancement on affordance. These suggest that our method can effectively capture the object's attributes and affordances characteristics and own the ability to reason out the multi-label affordances from attributes. Furthermore, we observe that the performance of affordance (β) is better than attribute (α). As mentioned in [6], the possible reason is that one object contains richer attribute information than affordance information. In addition, the reasoning scores $\mathcal{S}_{\text{ITE}}$ and $\mathcal{S}_{\alpha\text{-}\beta\text{-ITE}}$, which incorporate recognition probabilities, are higher than those of the baseline methods. This demonstrates that our approach effectively facilitates reasoning about the relationships between attributes and affordances. As shown in Fig. 9, compared with OCRN [6], our method not only predicts attributes and affordances more accurately but also correctly recognizes the causality pairs, which further demonstrates the superiority of CORE. To provide a more rigorous evaluation of attribute and affordance recognition, we complement the qualitative visualizations with quantitative analysis. Notably, the OCL dataset does not provide ground-truth concept localization annotations for attributes and affordances. To address this limitation, we manually annotate a subset of images with ground-truth regions corresponding to attribute and affordance concepts, and adopt IoU as a metric to evaluate localization accuracy. Specifically, we select 500 images from the OCL dataset, covering diverse 20 object categories, 10 attributes, and 10 affordances. For each image, concept-relevant regions are annotated according to the semantic definition of the corresponding attribute or affordance. For example, attribute annotations focus on the visual regions that directly exhibit the target property, while affordance annotations highlight the object parts responsible for the corresponding functionality. The results, as shown in Fig. 4, indicate that CORE achieves more accurate and spatially consistent localization than baseline method OCRN, as evidenced by consistently higher IoU scores.

Multi-Task Indoor Scene Understanding. Multi-task indoor scene understanding is a task to parse attribute, affordance and semantic from a single image. The mean intersection over union (mIoU) score is the evaluation metric. To evaluate the scene understanding qualities and generalization ability of our proposed method, we add our

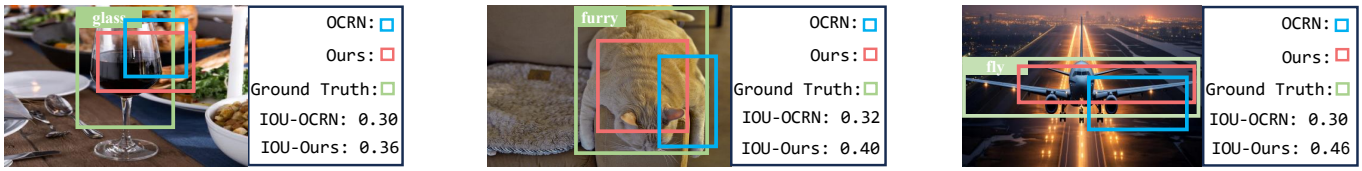


Fig. 4. IoU-based visualization results on the OCL dataset. We present IoU evaluations across diverse objects with different attributes and affordances. Compared with the baseline method OCRN, our approach achieves higher IoU scores, indicating more accurate localization of concept-relevant regions. These results demonstrate the effectiveness of our method in learning precise concept regions.

TABLE 3
Evaluation of cross-dataset transfer on the OVAD benchmark, encompassing all, head, medium, and tail attributes.

Method	$\mathbf{AP_{all}}$	$\mathbf{AP_{head}}$	$\mathbf{AP_{medium}}$	$\mathbf{AP_{tail}}$
CLIP RN50 [11]	15.8	42.5	17.5	4.2
CLIP VIT-B16 [11]	16.6	43.9	18.6	4.4
Open CLIP RN50 [63]	11.8	41.0	11.7	1.4
Open CLIP ViT-B16 [63]	16.0	45.4	17.4	3.8
CORE	**16.9**	44.0	**18.9**	**5.1**

TABLE 4
Comparative analysis of open-vocabulary attribute recognition on the VAW test set.

Method	Extra data	$\mathbf{AP_{base}}$	$\mathbf{AP_{novel}}$	$\mathbf{AP_{all}}$
OvarNet [64]	COCO Cap-sub	69.80	56.43	68.52
OV-DAR-query [65]	COCOCap	72.19	58.88	70.49
OV-DAR-anchor [65]	COCOCap	72.78	59.41	71.05
OVA_t^{II} [65]	COCOCap	77.47	65.74	75.97
CORE	COCOCap	**78.21**	**66.33**	**76.70**

TABLE 5
Effect of the proposed modules on the OCL benchmark.

Method	α	β	$\mathcal{S}_{\mathrm{ITE}}$	$\mathcal{S}_{\alpha\text{-}\beta\text{-ITE}}$
Base	33.5	54.2	10.6	10.0
+CRR	38.0	56.1	10.4	10.1
+CRE	**39.6**	**57.5**	**11.1**	**10.7**

TABLE 6
Effect of the prompt step on the OCL benchmark.

Prompt	α	β	$\mathcal{S}_{\mathrm{ITE}}$	$\mathcal{S}_{\alpha\text{-}\beta\text{-ITE}}$
global	37.5	55.2	10.7	10.2
local	24.8	40.6	9.8	9.5
global+local	**39.6**	**57.5**	**11.1**	**10.7**

method to the baseline method Cerberus [58]. We set the number of attribute concepts k_α equals 6 and the number of affordance concepts k_β equals 3. In Table 2, CORE is added to the Cerberus and improves the performance significantly. Besides, in the semantic parsing task, although the results do not perform well compared with VarReg, it still improves upon the Cerberus baseline in accuracy. These indicate that CORE not only enhances the model's reasoning ability but also helps to achieve joint inference in multi-task prediction.

Open-Vocabulary Object Attribute Recognition. To further examine CORE's transferability and zero-shot reasoning capability, we perform open-vocabulary attribute prediction under the box-supervised setting on the VAW [60] and OVAD [59] benchmarks. The evaluation is conducted using Average Precision (AP) as the metric for multi-label attribute classification performance. As shown in Table 3, the proposed method is evaluated on the OVAD benchmark, following the testing protocol specified in OVAD. The AP accuracy is measured across different attribute categories ("head", "middle", and "tail") under varying frequency distributions. In Table 4, comparisons are made with other attribute prediction approaches on the VAW test set following OVA_t^{II}. Under the box-supervised setting, CORE improves accuracy across all attribute classes, indicating that it can benefit open-vocabulary attribute recognition.

4.4 Ablation Study

Module Ablation. We validate the effectiveness of different high-level modules of our CORE, including CLIP (Base), coarse-to-fine relational reasoning (CRR) and counterfactual relation enhancing (CRE). As shown in Table 5, each component contributes to the superior performance of CORE. Compared with CLIP (Base), CRR improves the recognition of object concepts by leveraging coarse-to-fine prompt learning, which progressively incorporates both global and local contextual information. This hierarchical reasoning process allows the model to more accurately align visual features with corresponding concepts, thereby reducing ambiguities in object–concept mappings. However, it is worth noting that CRR primarily focuses on enhancing perceptual concept extraction rather than explicitly modeling causal relations. The reasoning metric $\mathcal{S}_{\mathrm{ITE}}$ evaluates causal consistency by measuring how affordance predictions change under counterfactual interventions. Since CRR does not introduce any intervention mechanism, the learned representations contain correlations between attributes and affordances rather than causal dependencies. As a result, while recognition accuracy improves significantly, the causal reasoning score may exhibit slight fluctuations. Furthermore, CRE enables the model to suppress spurious correlations and capture causal dependencies, which not only strengthens the semantic association between objects and their concepts but also improves the accuracy of reasoning scores $\mathcal{S}_{\mathrm{ITE}}$ and $\mathcal{S}_{\alpha\text{-}\beta\text{-ITE}}$. Together, these components demonstrate that both hierarchical relational reasoning and counterfactual enhancement are essential for object concept learning.

Analysis of the Contextual Prompt. Table 6 presents the effects of prompt flow. We decompose the prompt reasoning into a two-step refining process, where the first step is to

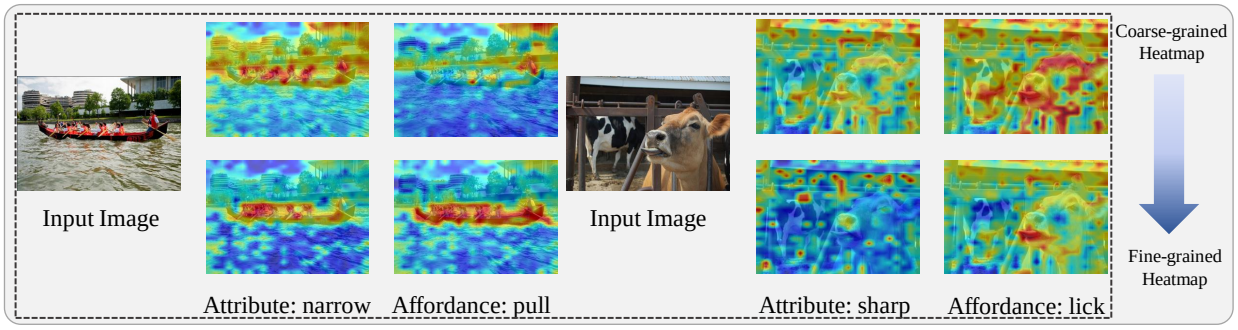
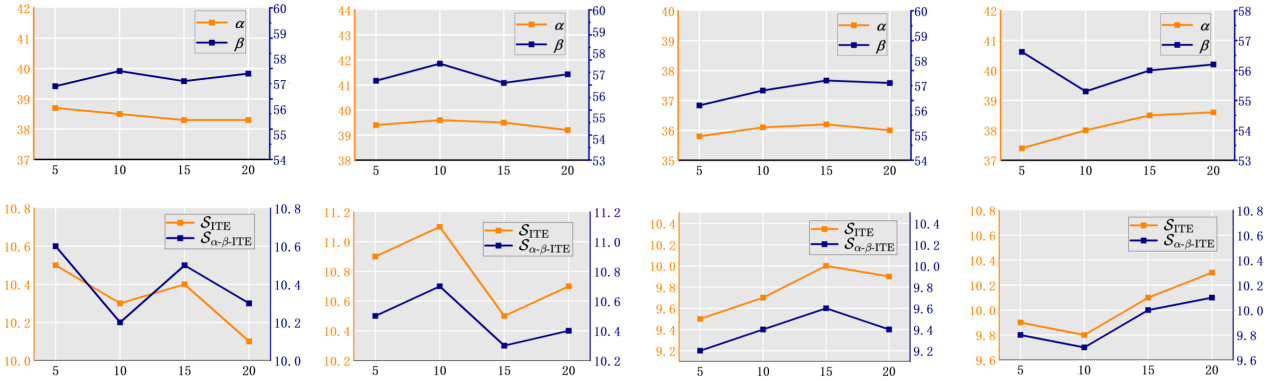


Fig. 5. The heatmaps of CORE. For each image, the top side indicates the heatmaps from the coarse-grained relational reasoning module, and the bottom side indicates the heatmaps from the fine-grained relational reasoning module.



(a) Ablation on the k_β when k_α equals 5.

(b) Ablation on the k_β when k_α equals 10.

(c) Ablation on the k_β when k_α equals 15.

(d) Ablation on the k_β when k_α equals 20.

Fig. 6. The ablation results of the concept number on the OCL benchmark. When the values of k_α and k_β are the same, the recognition accuracy is higher. The significant differences in visual concepts k_α and k_β may lead to inaccurate mapping relationships.

generate prompts containing the global image information. The second row in Table 6 shows the results only by fusing the global image context with text prompts. The second step is to relay the prompt from the previous step and deepen the local instance corresponding prompts. The fourth row results indicate the performance of the second step, surpassing the performance of the global contextual prompt. However, only the instance-specific contextual prompt leads to poor performance, as shown in the third row. The results suggest that the model can not align the attribute and affordance prompts with image features directly, solely from local regions. The reason may be that most objects interact with the nearby environment. Thus, the model is difficult to comprehend instance information without global image guidance. Additionally, we present the prompt heatmaps in Fig. 5. Our method could focus on concept-related object regions progressively by means of coarse-to-fine prompts, which improves the accuracy of recognition concepts.

Analysis of the Number of Concepts. We study the impact on recognition results when the numbers of attributes k_α and affordances k_β in visual concepts differ. In Fig. 6, the results are reported by changing $k \in [1, 20]$ with a 5 interval. As can be observed in the OCL benchmark, when the values of k_α and k_β are the same, the recognition accuracy is higher. This indicates that significant differences in visual concepts may lead to inaccurate mapping relationships. The recognition accuracy reaches its peak when k equals 10.

A reasonable explanation is that the learning process may become overly complicated due to the greater number of attributes and affordances, as well as the complex image content in OCL. We also conduct experiments where the number of k samples is equal to the total number of attributes (114) and affordances (170) in the entire dataset. The results are shown in Table 7. When the number of k equals the number of attributes and affordances, the model's performance could be improved compared with the CLIP baseline. However, the model's performance decreases compared with the "attr(10)-aff(10)". The experimental results reveal that an excessive number of k samples decreases performance, and a significant difference between attributes and affordances also results in poor performance. Learning too many visual concepts will increase the network's complexity and introduce information redundancy, thus degrading performance.

Analysis of the Prompt Token Number n. From the Table 8, we can find that the performance initially improves with an increase in the value of n. However, within the range of lengths from 12 to 16, we notice a decline in performance, which suggests that excessively long learnable prompts could introduce redundant or noisy information, thereby reducing the effectiveness of feature representation. Therefore, an appropriate value is $n = 12$, which achieves the best performance.

Analysis of the Concept Connection Network. The goal of the concept connection network is to construct connections among concepts to reason out the specific affordances labels.

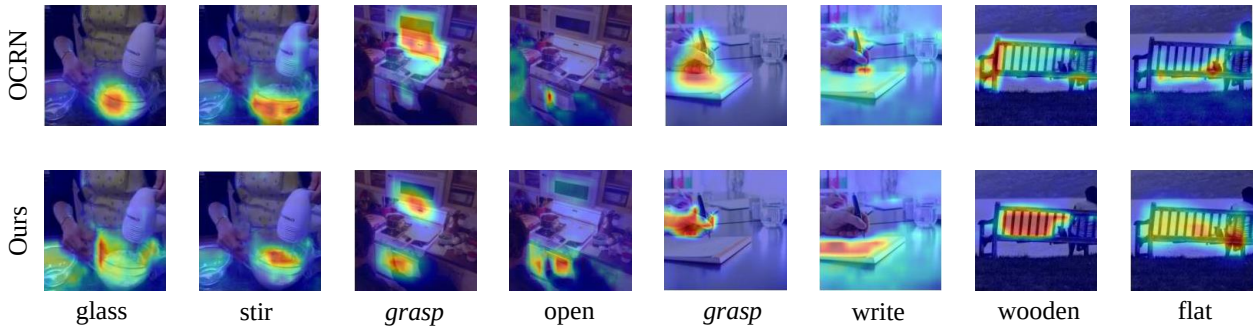


Fig. 7. Many-to-many attribute–affordance heatmaps. The upper row shows the attribute and affordance heatmaps produced by the baseline method OCRN, while the lower row presents the corresponding results of our method. Compared with OCRN, our method yields more spatially precise and semantically disentangled activation patterns. Notably, the visualizations highlight the inherent many-to-many relationships between objects, attributes, and affordances: an object possess multiple concepts, a concept (*e.g.*, grasp) can also belong to multiple objects.

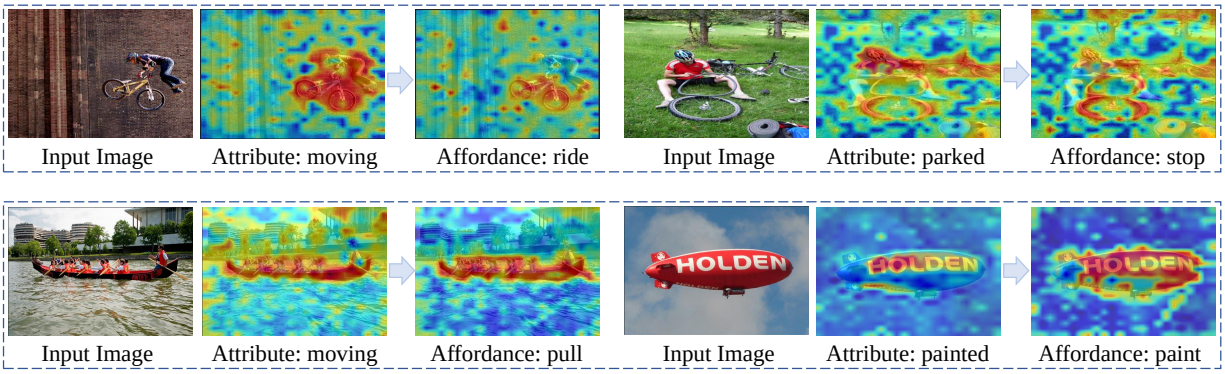


Fig. 8. The visualization of causal relations. The results show that the regions of attributes and affordances with causal relationships are largely aligned, suggesting that specific attributes can be used to infer the corresponding affordances.

TABLE 7
Effect of concept number k on OCL. "att" and "aff are the abbreviations for attributes and affordances, respectively. The values in "()" represent the number of k.

k	α	β	$\mathcal{S}_{\text{ITE}}$	$\mathcal{S}_{\alpha\text{-}\beta\text{-ITE}}$
att(114)-aff(114)	36.7	55.9	10.3	10.0
att(114)-aff(170)	35.9	55.1	10.1	9.9
att(10)-aff(10)	**39.6**	**57.5**	**11.1**	**10.7**
CLIP[§]	33.5	54.2	10.6	10.0

TABLE 8
Effect of learnable prompts n on OCL.

n	α	β	$\mathcal{S}_{\text{ITE}}$	$\mathcal{S}_{\alpha\text{-}\beta\text{-ITE}}$
10	38.9	57.0	10.8	10.5
12	**39.6**	**57.5**	**11.1**	**10.7**
14	39.2	57.1	10.9	10.5
16	38.7	56.8	10.7	10.4

To validate the effectiveness of our designed network, we replace matrix $\mathbf{A}$ in the network with a linear network for experimental analysis. The results in Table 9 shows that replacing the adjacency matrix $\mathbf{A}$ with a linear network reduces performance. This indicates that our concept connection network is better at improving performance.

Analysis of the Bounding Boxes. We replace the bounding boxes with the predicted boxes extracted by Faster R-CNN. The results of our experiments are shown in Table 10.

TABLE 9
Effect of concept connection network on OCL.

Method	α	β	$\mathcal{S}_{\text{ITE}}$	$\mathcal{S}_{\alpha\text{-}\beta\text{-ITE}}$
w/ Linear Network	39.2	56.0	10.6	10.2
w/o Linear Network	**39.6**	**57.5**	**11.1**	**10.7**

TABLE 10
Effect of Ground-truth Bounding Boxes (bbox) on OCL.

Method	α	β	$\mathcal{S}_{\text{ITE}}$	$\mathcal{S}_{\alpha\text{-}\beta\text{-ITE}}$
OCRN w/ bbox	31.6	53.5	9.9	9.5
OCRN w/o bbox	28.6	49.8	9.4	9.1
CORE w/o bbox	**37.5**	**54.4**	**10.6**	**10.0**

TABLE 11
Results for OCL dataset on random vs manual prompts.

Method	α	β	$\mathcal{S}_{\text{ITE}}$	$\mathcal{S}_{\alpha\text{-}\beta\text{-ITE}}$
$[T_i]_{i=1}^n/[P_i]_{i=1}^n$	39.5	57.3	11.1	10.6
$[T_i]_{i=1}^n$[has]/$[P_i]_{i=1}^n$[can]	39.5	57.5	11.1	10.7
$[T_i]_{i=1}^n$[is]/$[P_i]_{i=1}^n$[afford to]	**39.6**	**57.5**	**11.1**	**10.7**

From the experimental results, it can be observed that there is a decline in performance after using the pre-trained Faster R-CNN to extract the prediction boxes. However, our method can still enhance the performance of the baseline.

Analysis of the Hand-Crafted Prompts. To clarify the rationale behind the choice and optimization status of these

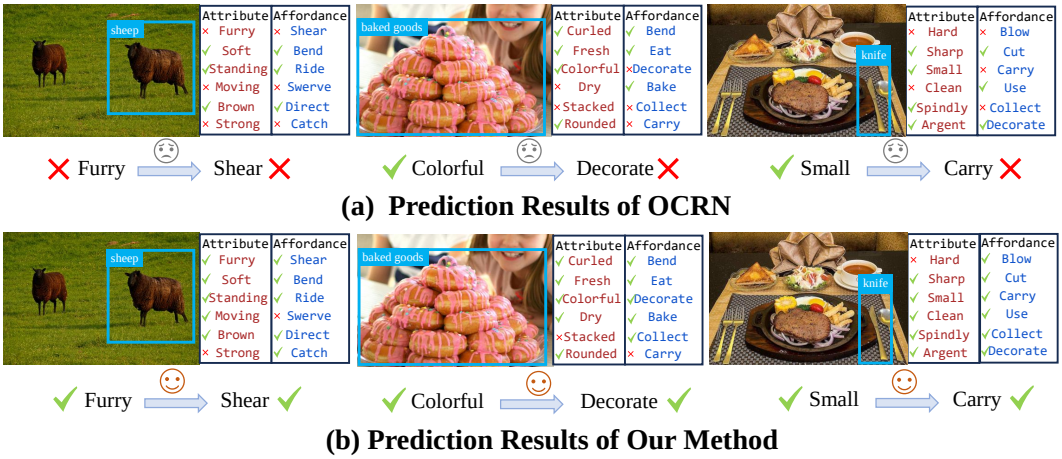


Fig. 9. The visualization of the baseline method (OCRN) and CORE. The attributes and affordances prediction results are shown in the right part of the image. The causal relation from dataset annotation is presented below each image. It is clearly shown that our method can recognize the causality pairs more accurately.

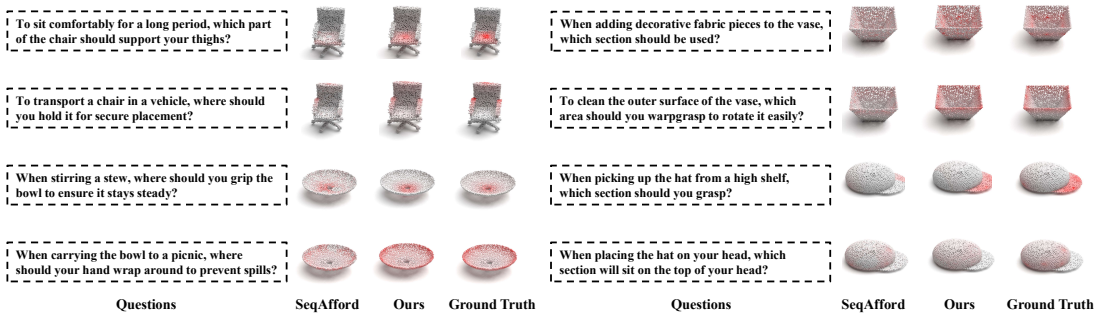


Fig. 10. Qualitative comparison of 3D affordance reasoning. For each query, we visualize the predicted affordance regions from SeqAfford, Ours and the Ground Truth. This highlights the superiority of our method in accurately grounding affordance semantics in 3D scenes.

TABLE 12
Results for the OCL dataset on zero-shot and fine-tune setting.

	Method	α	β	$\mathcal{S}_{\text{ITE}}$	$\mathcal{S}_{\alpha\text{-}\beta\text{-ITE}}$
zero-shot	CLIP	23.6	49.6	5.2	4.9
	Qwen2-VL-7B	30.0	52.2	8.1	7.8
	InterVL-2-4B	32.6	53.1	8.6	8.2
fine-tune	CLIP + CORE (Ours)	39.6	57.5	11.1	10.7
	Qwen2-VL + CORE	43.5	60.0	12.0	11.3
	InterVL-2 + CORE	44.6	62.0	12.2	11.5

TABLE 13
Quantitative results on 3DAffordSplat for affordance reasoning task.

Setting	Method	mIoU↑	AUC↑	SIM↑	MAE↓
Seen	IAGNet [67]	14.63	56.67	0.35	0.41
	PointRefer [68]	18.40	78.50	0.43	0.20
	AffordSplatNet [69]	30.25	83.85	0.44	0.21
	AffordSplatNet+CORE	**31.00**	**84.98**	**0.45**	**0.20**
Unseen	IAGNet [67]	4.70	40.77	0.24	0.43
	PointRefer [68]	15.90	67.00	0.31	0.29
	AffordSplatNet [69]	17.31	67.18	0.32	0.31
	AffordSplatNet+CORE	**17.92**	**67.22**	**0.32**	**0.30**

specific tokens, we have conducted an ablation study on the OCL datasets. We compared our current tokens ("[is]" and "[afford to]") with "[has]" and "[can]" and randomly initialized tokens. As shown in the Table 11, the natural language tokens consistently outperform random ones. This demonstrates that using semantically meaningful tokens effectively activates the prior information embedded in the pretrained vision language model.

4.5 Visualization for Object-Attribute-Affordance

To gain deeper insights into object–attribute–affordance associations, we extend our visualization analysis to a broader range of object categories, including diverse interaction types such as utensils and office supplies, and incorporate multi-attribute and multi-affordance heatmaps. As illustrated in Fig. 7, a single object instance simultaneously exhibits multiple attributes and affordances, while these concepts are also being associated with multiple object categories. These examples explicitly reveal the many-to-many relationships between object attributes and affordances, consistent with the conceptual formulation presented in Fig. 1. Furthermore, we provide side-by-side comparisons with the baseline method OCRN. The resulting heatmaps demonstrate that our approach produces more spatially precise and semantically disentangled activation patterns. These observations provide stronger evidence for the effectiveness of our proposed reasoning framework. To further illustrate that our method CORE can learn the causal relationships

TABLE 14
Summary of Notations used in CORE.

Symbol	Dimension	Description	Symbol	Dimension	Description
X_i	–	Input image	k	–	Number of concept slots
$v(\cdot)$	–	Visual encoder	$f(\cdot)$	–	Text encoder
N_α	–	Number of attribute categories	N_β	–	Number of affordance categories
F_g	$\mathbb{R}^d$	Global visual feature	F_M	$\mathbb{R}^{p\times d}$	Patch features
I_g	$\mathbb{R}^d$	Instance-level feature	F_a	$\mathbb{R}^d$	Aggregated visual feature
H_α	$\mathbb{R}^{N_\alpha\times d}$	Attribute embeddings	H_β	$\mathbb{R}^{N_\beta\times d}$	Affordance embeddings
$\hat{H}_\alpha$	$\mathbb{R}^{N_\alpha\times d}$	Coarse attribute features	$\hat{H}_\beta$	$\mathbb{R}^{N_\beta\times d}$	Coarse affordance features
$\bar{H}_\alpha$	$\mathbb{R}^{N_\alpha\times d}$	Fine attribute features	$\bar{H}_\beta$	$\mathbb{R}^{N_\beta\times d}$	Fine affordance features
C_α	$\mathbb{R}^{k\times D}$	Attribute concepts	C_β	$\mathbb{R}^{k\times D}$	Affordance concepts
$C_\alpha^{(t)}$	$\mathbb{R}^{k\times D}$	Attribute concepts at iteration t	$C_\beta^{(t)}$	$\mathbb{R}^{k\times D}$	Affordance concepts at iteration t
$\mathrm{attn}_\alpha^{(t)}$	$\mathbb{R}^{k\times N_\alpha}$	Attribute attention	$\mathrm{attn}_\beta^{(t)}$	$\mathbb{R}^{k\times N_\beta}$	Affordance attention
$U_\alpha^{(t)}$	$\mathbb{R}^{k\times D}$	Attribute updates	$U_\beta^{(t)}$	$\mathbb{R}^{k\times D}$	Affordance updates
$\Phi_g(\cdot)$	–	Self-attention refinement	$\Theta_g(\cdot)$	–	Global cross-attention
$\Theta_I(\cdot)$	–	Instance-level cross-attention	$[T_i]/[P_i]$	–	Prompt tokens
M	$\mathbb{R}^{k\times D}$	Aggregated concepts	$\widetilde{M}$	$\mathbb{R}^{k\times D}$	Concept connection output
w_f^c	$\mathbb{R}^{k\times k}$	Concept weights	b_f^c	$\mathbb{R}^D$	Bias term
F_α	$\mathbb{R}^D$	Final attribute representation	F_β	$\mathbb{R}^D$	Final affordance representation
$\bar{H}_{\alpha\mathrm{mask}}$	$\mathbb{R}^{N_\alpha\times d}$	Masked attribute embeddings	$F_{\beta\mathrm{mask}}$	$\mathbb{R}^D$	Counterfactual affordance
$\bar{F}$	$\mathbb{R}^k$	Fusion feature	$\hat{y}_\alpha$	–	Predicted attribute probabilities
$\hat{y}_\beta$	–	Predicted affordance probabilities	y_α	–	Ground-truth attributes
y_β	–	Ground-truth affordances	L_{bce}	–	Binary cross-entropy loss
L_{cd}	–	Concept distinctiveness loss	$\mathcal{L}_{\mathrm{likel}}$	–	Log-likelihood loss
λ_1, λ_2	–	Loss weights	L_{total}	–	Total loss

between attributes and affordances, we visualize the features of attributes and affordances in the image that have causal annotation relationships, with the results shown in Fig. 8. Through the visualization, it can be observed that the regions of attributes and affordances with causal relationships are approximately similar. The visualization results indicate that specific attributes can be used to infer the corresponding affordances.

4.6 Results on Zero-shot and Fine-tuned Settings

To further demonstrate the effectiveness of our method, we conduct experiments with the Qwen2-VL and InterVL-2 [70] models on the zero-shot and fine-tuned settings, and the results are as shown in Table 12. Under the zero-shot setting, for attribute and affordance recognition, we provide the corresponding attribute and affordance categories as candidate concepts and ask the models to perform matching. For causal relation identification, we simulate attribute-level interventions by modifying the input representations. Specifically, for CLIP, we perform token-level masking by setting the corresponding attribute embeddings to zero. For Qwen2-VL and InternVL-2, we approximate the intervention by replacing each attribute with its negated form (*e.g.*, "sharp" $\rightarrow$ "not sharp"). Based on these factual and counterfactual predictions, we compute the causal effect following the same protocol as the OCL benchmark. Under the fine-tune setting, for Qwen2-VL and InternVL-2, we attach the CORE to their multimodal representations while keeping the backbone encoders frozen. This design is intended to explicitly learn causal associations among objects, attributes, and affordances. By fixing the parameters of the VLM encoders, the results in Table 12 show that our method still achieves consistent improvements. Therefore, the observed gains can be attributed to the reasoning capability introduced by the CORE method.

TABLE 15
Quantitative results on SeqAfford for affordance segmentation task. * means that baseline methods use ground-truth sequential order as all existing methods cannot predict sequential affordances.

	Method	mIoU↑	AUC↑	SIM↑	MAE↓
Seen	ReferTrans [71]	11.4	77.2	0.449	0.135
	ReLA [72]	12.1	76.3	0.480	0.130
	3D-SPS [73]	10.1	75.2	0.413	0.141
	IAGNet [67]	14.2	81.7	0.510	0.117
	PointRefer [68]	16.3	84.3	0.568	0.108
	SeqAfford [74]	19.5	86.9	0.594	0.098
	SeqAfford+CORE	**20.0**	**87.3**	**0.608**	**0.073**
Unseen	ReferTrans [71]	9.1	67.4	0.427	0.151
	ReLA [72]	9.3	68.2	0.423	0.147
	3D-SPS [73]	7.1	66.9	0.397	0.162
	IAGNet [67]	11.7	73.6	0.438	0.143
	PointRefer [68]	12.4	76.1	0.502	0.132
	SeqAfford [74]	13.8	82.4	0.518	0.128
	SeqAfford+CORE	**14.1**	**83.5**	**0.522**	**0.126**
Squential	ReferTrans* [71]	10.8	74.5	0.425	0.142
	ReLA* [72]	11.4	74.8	0.463	0.136
	3D-SPS* [73]	9.9	73.1	0.407	0.148
	IAGNet* [67]	13.5	78.2	0.496	0.131
	PointRefer* [68]	14.3	80.7	0.521	0.124
	SeqAfford* [74]	14.6	84.2	0.573	0.118
	SeqAfford*+CORE	**14.7**	**85.0**	**0.586**	**0.115**

4.7 Results on Affordance Reasoning

We further extend the CORE framework to the affordance reasoning task to demonstrate its applicability in action-aware and 3D interaction scenarios. The CRR module can be adapted to SeqAfford [74] by reformulating it as a hierarchical 3D affordance reasoning module. Specifically, the coarse-grained stage uses the global scene representation together with instruction embeddings to produce step-aware affordance queries, while the fine-grained stage refines these queries using local 3D region features. In this way, CRR serves as a hierarchical Language–Point alignment module

that improves sequential affordance grounding in 3D scenes. We apply the CRR module to the 3DAffordSplat [69] by reformulating it as a coarse-to-fine affordance query refinement mechanism over Gaussian representations. Given an affordance query, we first perform coarse-grained reasoning by aligning it with global scene features aggregated from all Gaussians. The refined query is then further aligned with local Gaussian features through fine-grained cross-attention, producing spatially consistent affordance predictions. Experimental results in Table 13 and Table 15 demonstrate that such integration improves spatial localization accuracy. To further demonstrate the advantages of our method, we provide qualitative visualizations on the SeqAfford dataset, as shown in Fig. 10. It can be observed that our approach more accurately localizes affordance regions, highlighting its effectiveness in capturing meaningful interaction areas.

5 Conclusion

In this work, we present the Hierarchical Cross-Modal Relational Reasoning (CORE) framework designed to address many-to-many object–concept mappings and causal reasoning in Object Concept Learning. To improve the accuracy of concept–object associations, object-agnostic prompts are employed together with the integration of global semantic cues and fine-grained local features. Furthermore, to make the mapping between objects and concepts more precise, we introduce counterfactual reasoning to identify causal relationships between certain attributes and affordances. This enhances the mapping connections between objects and these concepts, which in turn improves recognition performance. Extensive experiments on multiple benchmarks demonstrate that CORE delivers significant improvements over state-of-the-art methods, while comprehensive visualization analyses confirm its interpretability and effectiveness. The results highlight that the hierarchical reasoning paradigm not only strengthens semantic alignment between objects and high-level concepts but also facilitates more reliable generalization in open-vocabulary scenarios. Nevertheless, the current framework still relies on the representational capacity of pretrained vision-language models. Although strong semantic priors can be inherited from these models, their reasoning abilities remain limited when handling more complex concept compositions and long-range dependencies. In future work, we plan to investigate the integration of large multimodal models and advanced reasoning mechanisms to further enhance concept understanding and causal inference.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2012, pp. 1106–1114.

[2] X. Li, Z. Wang, B. Zhang, F. Sun, and X. Hu, "Recognizing object by components with human prior knowledge enhances adversarial robustness of deep neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, pp. 8861–8873, 2023.

[3] C. Pan, J. Peng, X. Bu, and Z. Zhang, "Large-scale object detection in the wild with imbalanced data distribution, and multi-labels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, pp. 9255–9271, 2024.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 5998–6008.

[5] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 16 519–16 529.

[6] Y.-L. Li, Y. Xu, X. Xu, X. Mao, Y. Yao, S. Liu, and C. Lu, "Beyond object recognition: A new benchmark towards object concept learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 20 029–20 040.

[7] M. Hassan and A. Dharmaratne, "Attribute based affordance detection from human-object interaction images," in *Image and Video Technology – PSIVT 2015 Workshops (PSIVT)*, 2016, pp. 220–232.

[8] C. Zhou, H. Chen, J. Zhang, Q. Li, D. Hu, and V. S. Sheng, "Multi-label graph node classification with label attentive neighborhood convolution," *Expert Syst. Appl.*, vol. 180, p. 115063, 2021.

[9] Q. Guan and Y. Huang, "Multi-label chest x-ray image classification via category-wise residual attention learning," *Pattern Recogn. Lett.*, vol. 130, pp. 259–266, 2020.

[10] G. Sumbul and B. Demİr, "A deep multi-attention driven approach for multi-label remote sensing image classification," *IEEE Access*, vol. 8, pp. 95 934–95 946, 2020.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.

[12] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[13] C. Xie, S. Liang, J. Li, Z. Zhang, F. Zhu, R. Zhao, and Y. Wei, "Relationlmm: Large multimodal model as open and versatile visual relationship generalist," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, pp. 3515–3529, 2025.

[14] C.-M. Feng, K. Yu, X. Xu, S. Khan, R. S. M. Goh, W. Zuo, and Y. Liu, "Text to image for multi-label image recognition with joint prompt-adapter learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, pp. 7660–7674, 2025.

[15] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022, pp. 24 824–24 837.

[16] Y. Rao, W. Zhao, G. Chen, Y. Tang, Z. Zhu, G. Huang, J. Zhou, and J. Lu, "Denseclip: Language-guided dense prediction with context-aware prompting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 18 082–18 091.

[17] L. Wang, J. Xie, X. Zhang, H. Su, and J. Zhu, "Hide-pet: continual learning via hierarchical decomposition of parameter-efficient tuning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, pp. 6687–6702, 2025.

[18] X. Xing, Z. Xiong, A. Stylianou, S. Sastry, L. Gong, and N. Jacobs, "Vision-language pseudo-labels for single-positive multi-label learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 7799–7808.

[19] M. Ramadan, C. Tang, N. Watters, and M. Jazayeri, "Computational basis of hierarchical and counterfactual information processing," *Nat. Hum. Behav.*, pp. 1–15, 2025.

[20] J. A. Lorteije, A. Zylberberg, B. G. Ouellette, C. I. De Zeeuw, M. Sigman, and P. R. Roelfsema, "The formation of hierarchical decisions in the visual cortex," *Neuron*, vol. 87, pp. 1344–1356, 2015.

[21] N. Van Hoeck, P. D. Watson, and A. K. Barbey, "Cognitive neuroscience of human counterfactual reasoning," *Front. Hum. Neurosci.*, vol. 9, p. 420, 2015.

[22] A.-K. Dombrowski, J. E. Gerken, K.-R. Müller, and P. Kessel, "Diffeomorphic counterfactuals with generative models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, pp. 3257–3274, 2023.

[23] Y. Liu, G. Li, and L. Lin, "Cross-modal causal relational reasoning for event-level visual question answering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, pp. 11 624–11 641, 2023.

[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2015, pp. 91–99.

[25] C. Huynh, A. T. Tran, K. Luu, and M. Hoai, "Progressive semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 16 755–16 764.

[26] S. J. Hwang, F. Sha, and K. Grauman, "Sharing features between objects and their attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2011, pp. 1761–1768.

[27] Y.-L. Li, Y. Xu, X. Mao, and C. Lu, "Symmetry and group in attribute-object compositions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 11 316–11 325.

[28] D. Chen, D. Kong, J. Li, S. Wang, and B. Yin, "A survey of visual affordance recognition based on deep learning," *IEEE Trans. Big Data*, vol. 9, pp. 1458–1476, 2023.

[29] S. Aarthi and S. Chitrakala, "Scene understanding—a survey," in *IEEE International Conference on Comput., Commun., and Signal Process. (ICCCSP)*, 2017, pp. 1–4.

[30] M. Antoun and D. Asmar, "Human object interaction detection: Design and survey," *Image Vision Comput.*, vol. 130, p. 104617, 2023.

[31] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Mach. Learn.*, vol. 29, pp. 131–163, 1997.

[32] W. S. Noble, "What is a support vector machine?" *Nat. Biotechnol.*, vol. 24, pp. 1565–1567, 2006.

[33] L. Montesano, M. Lopes, A. Bernardino, and J. Santos-Victor, "Learning object affordances: from sensory–motor coordination to imitation," *IEEE Trans. Robot.*, vol. 24, pp. 15–26, 2008.

[34] E. Uğur and E. Şahin, "Traversability: A case study for learning and perceiving affordances in robots," *Aadpt. Behav.*, vol. 18, pp. 258–284, 2010.

[35] L. Pinto and A. Gupta, "Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours," in *Proc. IEEE Int. Conf. robot. autom. (ICRA)*, 2016, pp. 3406–3413.

[36] P. Dadure, P. Pakray, and S. Bandyopadhyay, "Challenges and opportunities in knowledge representation and reasoning," *Encyclopedia of Data Science and Machine Learning*, pp. 2464–2477, 2023.

[37] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," *arXiv preprint arXiv:2104.08691*, 2021.

[38] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, pp. 1–35, 2023.

[39] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022, pp. 23 716–23 736.

[40] A. J. Wang, P. Zhou, M. Z. Shou, and S. Yan, "Enhancing visual grounding in vision-language pre-training with position-guided text prompts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, pp. 3406–3421, 2023.

[41] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, and Z. Sui, "A survey on in-context learning," *arXiv preprint arXiv:2301.00234*, 2022.

[42] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vis.*, vol. 130, pp. 2337–2348, 2022.

[43] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 16 816–16 825.

[44] Y. Wei, Y. Liu, H. Yan, G. Li, and L. Lin, "Visual causal scene refinement for video question answering," in *ACM SIGMM Rec. (ACMMM)*, 2023, pp. 377–386.

[45] W. Chen, Y. Liu, B. Chen, J. Su, Y. Zheng, and L. Lin, "Cross-modal causal relation alignment for video question grounding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025, pp. 24 087–24 096.

[46] W. Chen, Y. Liu, C. Wang, J. Zhu, G. Li, C.-L. Liu, and L. Lin, "Cross-modal causal representation learning for radiology report generation," *IEEE Trans. Image Process.*, 2025.

[47] Q. Zhou, G. Pang, Y. Tian, S. He, and J. Chen, "Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection," *arXiv preprint arXiv:2310.18961*, 2023.

[48] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[49] J. Pearl, "Direct and indirect effects," in *Probabilistic and causal inference: the works of Judea Pearl*, 2022, pp. 373–392.

[50] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang, "Unbiased scene graph generation from biased training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 3716–3725.

[51] P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin, "Club: A contrastive log-ratio upper bound of mutual information," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 1779–1788.

[52] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, pp. 533–536, 1986.

[53] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 746–760.

[54] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2881–2890.

[55] H. Wu, J. Zhang, K. Huang, K. Liang, and Y. Yu, "Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation," *arXiv preprint arXiv:1903.11816*, 2019.

[56] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[57] H. Shi, H. Li, Q. Wu, and Z. Song, "Scene parsing via integrated classification model and variance-based regularization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 5307–5316.

[58] X. Chen, T. Liu, H. Zhao, G. Zhou, and Y.-Q. Zhang, "Cerberus transformer: Joint semantic, affordance and attribute parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 19 649–19 658.

[59] M. A. Bravo, S. Mittal, S. Ging, and T. Brox, "Open-vocabulary attribute detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 7041–7050.

[60] K. Pham, K. Kafle, Z. Lin, Z. Ding, S. Cohen, Q. Tran, and A. Shrivastava, "Learning to predict visual attributes in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 13 018–13 028.

[61] C. Wu, Z. Lin, S. Cohen, T. Bui, and S. Maji, "Phrasecut: Language-based image segmentation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 10 216–10 225.

[62] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 6700–6709.

[63] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev, "Reproducible scaling laws for contrastive language-image learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 2818–2829.

[64] K. Chen, X. Jiang, Y. Hu, X. Tang, Y. Gao, J. Chen, and W. Xie, "Ovarnet: Towards open-vocabulary object attribute recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 23 518–23 527.

[65] K. Chen, X. Jiang, H. Wang, C. Yan, Y. Gao, X. Tang, Y. Hu, and W. Xie, "Ov-dar: Open-vocabulary object detection and attributes recognition," *Int. J. Comput. Vis.*, vol. 132, pp. 5387–5409, 2024.

[66] D. B. Rubin, "Causal inference using potential outcomes: Design, modeling, decisions," *Journal of the American statistical Association*, vol. 100, no. 469, pp. 322–331, 2005.

[67] Y. Yang, W. Zhai, H. Luo, Y. Cao, J. Luo, and Z.-J. Zha, "Grounding 3d object affordance from 2d interactions in images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 10 905–10 915.

[68] Y. Li, N. Zhao, J. Xiao, C. Feng, X. Wang, and T.-s. Chua, "Laso: Language-guided affordance segmentation on 3d object," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 14 251–14 260.

[69] Z. Wei, J. Lin, Y. Liu, W. Chen, J. Luo, G. Li, and L. Lin, "3daffordsplat: Efficient affordance reasoning with 3d gaussians," in *ACM SIGMM Rec. (ACMMM)*, 2025, pp. 2821–2830.

[70] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu *et al.*, "Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 24 185–24 198.

[71] M. Li and L. Sigal, "Referring transformer: A one-step approach to multi-task visual grounding," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021, pp. 19 652–19 664.

[72] C. Liu, H. Ding, and X. Jiang, "Gres: Generalized referring expression segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 23 592–23 601.

[73] J. Luo, J. Fu, X. Kong, C. Gao, H. Ren, H. Shen, H. Xia, and S. Liu, "3d-sps: Single-stage 3d visual grounding via referred point progressive selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 16 454–16 463.
[74] C. Yu, H. Wang, Y. Shi, H. Luo, S. Yang, J. Yu, and J. Wang, "Seqafford: Sequential 3d affordance reasoning via multimodal large language model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025, pp. 1691–1701.

Yuxuan Wang received the B.E. degree from China University of Petroleum (East China), Qingdao, China, in 2021. She is currently working toward the Ph.D. degree with the School of Electronic Engineering, Xidian University, Xi'an, China. Her research interests include computer vision, causal inference and multimodal reasoning.

Muli Yang is a Research Scientist at the Institute for Infocomm Research (I^2R), A*STAR, Singapore. He received his PhD degree from Xidian University, Xi'an, China, in 2023, and was a visiting PhD student at Nanyang Technological University, Singapore, from 2022 to 2023. His research focuses on open-world learning and vision-language modeling. He has published more than 25 papers in leading conferences and journals, including CVPR, ICCV, ICLR, NeurIPS, ACL, IJCV and TIP.

Yihang Zhu received the B.E. degree from Xidian University, China, in 2023. He is currently pursuing a Ph.D. degree in the School of Electronic Engineering at Xidian University. His research interests include computer vision and machine learning.

Jiexi Yan received the B.E. and Ph.D. degrees from Xidian University, Xi'an, China, in 2017 and 2022. He is currently an Associate Professor with the School of Computer Science and Technology, Xidian University. His current research interests include machine learning and computer vision.

Fen Fang received the Ph.D. degree in computer science and engineering, majoring in computer graphics, from Nanyang Technological University, Singapore, in 2014. She is currently a Research Scientist with the Institute for Infocomm Research, A*STAR, Singapore. Her research interests include defect detection, object detection and segmentation, scene recognition, and reinforcement learning.

Xulei Yang (Senior Member, IEEE) received his PhD degree from Nanyang Technological University (NTU) in 2007. He is currently a principal scientist and group leader at Institute for Infocomm Research (I2R), A*STAR, previously the research head at YITU Technology Singapore, with more than 16 years of R&D experience in deep/machine learning for computer vision and healthcare. He has published more than 100 scientific papers and international patents in the fields of deep learning, 3D Vision and medical imaging. He is currently an IEEE Senior Member and Kaggle Competition Master.

Cheng Deng (Senior Member, IEEE) received the B.E., M.S., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China. He is currently a Professor with the Hohai University and Xidian University. He is the author and the co-author of more than 100 scientific papers at top venues, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CYBERNETICS, NeurIPS, ICML, CVPR, ICCV, AAAI, IJCAI, and KDD. His research interests include computer vision, pattern recognition, and information hiding.