Open Ad-hoc Categorization with Contextualized Feature Learning

Zilin Wang [*1] Sangwoo Mo[*1] Stella X. Yu[1,2] Sima Behpour[3] Liu Ren[3]
[1]University of Michigan [2]UC Berkeley [3]Bosch Center for AI
{zilinwan, swmo, stellayu}@umich.edu sima.behpour@gmail.com liu.ren@us.bosch.com

Abstract

Adaptive categorization of visual scenes is essential for AI agents to handle changing tasks. Unlike fixed common categories for plants or animals, ad-hoc categories, such as things to sell at a garage sale, are created dynamically to achieve specific tasks. We study open ad-hoc categorization, where the goal is to infer novel concepts and categorize images based on a given context, a small set of labeled exemplars, and some unlabeled data.

We have two key insights: 1) recognizing ad-hoc categories relies on the same perceptual processes as common categories; 2) novel concepts can be discovered semantically by expanding contextual cues or visually by clustering similar patterns. We propose OAK, a simple model that introduces a single learnable context token into CLIP, trained with CLIP's objective of aligning visual and textual features and GCD's objective of clustering similar images.

On Stanford and Clevr-4 datasets, OAK consistently achieves the state-of-art in accuracy and concept discovery across multiple categorizations, including 87.4% novel accuracy on Stanford Mood, surpassing CLIP and GCD by over 50%. Moreover, OAK generates interpretable saliency maps, focusing on hands for Action, faces for Mood, and backgrounds for Location, promoting transparency and trust while enabling accurate and flexible categorization.

1. Introduction

The concept of *ad-hoc categories* in cognitive science differs from *common categories* for plants or animals [2]. For example, *things to sell at a garage sale* (Fig. 1) may not share visual or semantic similarity but are grouped to achieve the goal of selling unwanted items. Unlike common categories, ad-hoc categories are less established in memory, often lack clear labels, require explicit naming of exemplars, and depend on context. Recognizing them relies on the same perceptual processes as common categories but requires contextualization to adapt to varying tasks [24, 41].

[*]Equal Contribution. https://github.com/Wayne2Wang/OAK

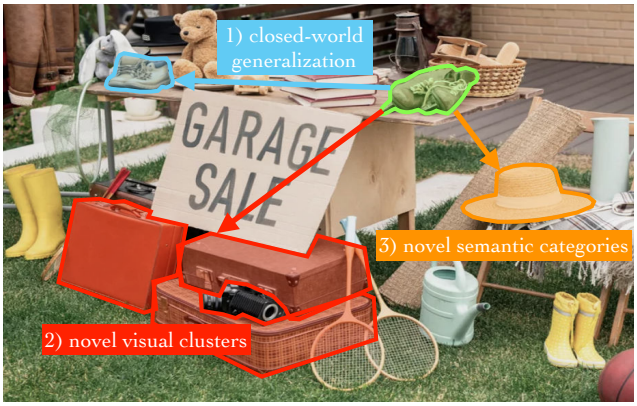


Figure 1. **We study open ad-hoc categorization** such as *things to sell at a garage sale* to achieve a specific goal (*selling unwanted items*). Given the context *garage sale*, labeled exemplars such as *shoes*, we need to both infer novel categories and recognize everything in the scene that *can be sold at the garage sale*. Supervisedly trained models like CLIP focus on 1) closed-world generalization, recognizing other *shoes*. Unsupervisedly trained methods like GCD discover 2) novel visual clusters, identifying *suitcases*. Intuitively, we can also discover 3) novel semantic categories by contextual expansion from *shoes* to *hats*.

We are inspired to study a novel problem setting called *open ad-hoc categorization*, which learns categorization rules under varying contexts to predict predefined categories and discover new ones. Fig. 1 illustrates three types of concepts we seek to uncover. **1)** Closed-world generalization applies known concepts to new instances, such as CLIP [39] recognizing unseen *shoes*. **2)** Novel visual clusters can emerge through data-driven image clustering, as in GCD [49] identifying *suitcases* as a new category. **3)** Novel semantic categories may not resemble known ones visually can arise contextually, such as from *shoes* to *hats*.

We formulate the problem as predicting both known and novel concepts from unlabeled images using a few labeled examples within a given context. Fig. 2 illustrates how the same set of images can belong to different ad-hoc categories, such as action, location, or mood. For instance, in the context of *action*, given labeled examples like *drinking* and *reading*, the task is to identify novel classes such

Figure 2. **Open ad-hoc categorization** (OAK) learns diverse categorization rules, dynamically adapting to varying user needs at hand. The same image should be recognized differently depending on context, such as *drinking* for action and *residential* for location. We emphasize the ability to switch between multiple contexts in OAK. Specifically, given 1) a context defined by classes, 2) a few labeled images, and 3) a set of unlabeled images, OAK holistically reasons over labeled and unlabeled images, spanning both known and novel classes, to infer novel concepts and propagate labels across the entire dataset. We show the class names of labeled images in the color box and unlabeled images inside the parenthesis, reflecting the unlabeled class names are not available, only the images. OAK introduces unique challenges beyond generalized category discovery (GCD), requiring adapting to diverse ad-hoc categorization rules based on context.

as *riding* and *climbing* and categorize all unlabeled images. When the context shifts to *location*, labeled examples like *residential area* and *natural environment* require inferring novel types such as *sports field* and *store*.

We have two insights: **1)** recognizing ad-hoc categories relies on the same perceptual processes as common categories; **2)** novel concepts can be discovered semantically by expanding contextual semantics or visually by clustering similar patterns. We capture these ideas by adapting the pre-trained CLIP with the given context and integrating CLIP's semantic classification with GCD's visual clustering.

We propose *Open Ad-hoc Categorization with Contextualized Feature Learning* (OAK for short, with K from Kategorisierung: Categorization in German), a simple model that introduces a single learnable context token into CLIP, trained with CLIP's objective of aligning visual and textual features and GCD's objective of clustering similar images. Combining CLIP's top-down open-vocabulary classification with GCD's bottom-up data-driven visual clustering, OAK can discover novel ad-hoc categories both semantically and visually. By switching the single global context token fed into the otherwise fixed CLIP, OAK derives contextualized image features for each ad-hoc categorization.

We evaluate performance using *Omni accuracy*, which measures the rate of fully correct predictions across all contexts, along with accuracies for known and novel classes. This is crucial for AI agents that must adapt seamlessly to diverse tasks [30], such as a household robot switching its context across actions for assistance, locations for navigation, and moods for emotional responses.

We benchmark OAK on the Stanford [23] (Action, Location, Mood) and Clevr-4 [51] (Texture, Color, Shape, Count) datasets. OAK outperforms all baselines, achieving higher novel accuracy, particularly in the Omni context. For example, it reaches 87.4% novel accuracy on Stanford Mood, significantly outperforming CLIP-ZS + LLM vocab (35.4%) and GCD (40.6%). Moreover, saliency maps from OAK offer interpretable insights, highlighting hands for Action, backgrounds for Location, and faces for Mood. Finally, OAK can name discovered novel clusters by aligning them with CLIP's text embeddings.

To summarize, our **key contributions** are as follows.

- We propose the open ad-hoc categorization task which unifies feature learning with context switching.
- We develop OAK, a simple yet effective method that contextualizes features integrating visual and textual cues.
- OAK outperforms baseline methods while also producing interpretable saliency maps and class names.

2. Related Work

Categorization is a core problem in computer vision, typically assuming common categories like object species [43]. We instead aim to learn *open ad-hoc categorization*, where category rules can be defined arbitrarily based on purpose. This context-dependent setting is often framed as meta-learning [14] or few-shot learning [45]. Our work adopts a similar setup but extends it to handle both known and novel classes, unlike conventional closed-world approaches.

Open-vocabulary classification models like CLIP [39] have emerged as universal classifiers, capable of handling diverse contexts by adjusting their class vocabulary. Although originally designed for closed-world scenarios with predefined labels, recent efforts have extended CLIP to open-world settings. These include retrieving terms from a

database [8, 12], prompting large language models (LLMs) to generate class descriptions [25, 29, 37, 63], and using vision-language models (VLMs) to analyze images and infer class names [21, 23, 26, 28, 60]. While these methods can assign names to potential novel classes, they remain limited to the vocabulary known to CLIP. In contrast, OAK adapts the image encoder to context, enabling the discovery of novel concepts from visual clusters.

Generalized category discovery (GCD) [3, 49] aims to identify both known and novel classes in unlabeled images by learning clusters from labeled data. We extend GCD to support multiple ad-hoc categorization rules. Vaze et al. [51] explored a related setting by evaluating visual backbones across contexts, observing that no single representation rules them all. In contrast, we propose a unified method that dynamically adapts features to context, performing well across all contexts. Our approach builds on GCD and remains compatible with recent techniques [6, 38, 40, 54, 61]. Some works apply GCD to CLIP [35, 52, 64] by combining caption and image features, whereas we tackle multi-context GCD without relying on captions. SPTNet [53] tunes visual prompts with a focus on efficiency, while we emphasize contextualizing visual features. GPC [62] infers pseudo-labels using only visual cues, whereas we incorporate semantic knowledge from a text encoder.

Adaptation of neural networks has been widely studied. We contextualize visual features using visual prompt tuning (VPT) [1, 16], which introduces learnable tokens to the input of the ViT encoder [10]. This enables attention maps to highlight context-relevant regions [18, 19, 31, 44], such as hands for actions and backgrounds for locations. Other adaptation techniques [15, 65] or CLIP fine-tuning methods [11, 32, 56] could also be incorporated.

3. Open Ad-hoc Categorization

3.1. Problem setup

Open ad-hoc categorization aims to learn categorization rules from a few labeled and some unlabeled images, and then generalize these rules to discover novel classes. This setup extends generalized category discovery (GCD) [49] by incorporating multiple categorization rules. In the image domain $\mathcal{X}$, a user defines a context c using a labeled dataset $\mathcal{D}^c_{\mathcal{L}} = \{(\mathbf{x}_i, y^c_i) \in \mathcal{X} \times \mathcal{Y}^c_{\mathcal{L}}\}^N_{i=1}$, which contains samples from known classes $\mathcal{Y}^c_{\mathcal{L}}$, and an unlabeled dataset $\mathcal{D}^c_{\mathcal{U}} = \{\mathbf{x}_i \in \mathcal{X}\}^M_{i=1}$, consisting of both known classes $\mathcal{Y}^c_{\mathcal{L}}$ and novel classes $\mathcal{Y}^c_{\mathcal{N}}$. The goal is to classify both known and novel classes within the unlabeled set, maximizing accuracy on $\mathcal{D}^c_{\mathcal{U}_{\mathcal{L}}}$ and $\mathcal{D}^c_{\mathcal{U}_{\mathcal{N}}}$, the subsets of $\mathcal{D}^c_{\mathcal{U}}$ corresponding to known and novel samples, respectively.

Context c can be inferred through two main principles. The first is top-down text guidance, where the model uses known class names $\mathcal{Y}^c_{\mathcal{L}}$ and semantic knowledge to infer context. Open-vocabulary classifiers like CLIP [39] can adapt to arbitrary contexts by adjusting their vocabulary. To infer potential novel classes $\hat{\mathcal{Y}}^c_{\mathcal{N}}$, large language models (LLMs) can be prompted with known class names $\mathcal{Y}^c_{\mathcal{L}}$, while vision-language models (VLMs) can refine this inference using labeled dataset $\mathcal{D}^c_{\mathcal{L}}$. With this expanded vocabulary, CLIP can predict all classes in $\mathcal{Y}^c_{\mathcal{L}} \cup \hat{\mathcal{Y}}^c_{\mathcal{N}}$. However, while this text-guided approach accommodates various contexts, it is limited by the pretrained knowledge of CLIP and LLMs. It also does not adapt the image encoder to specific contexts, restricting its representational capacity.

The second is bottom-up image clustering, where the model infers context by grouping visually similar images from $\mathcal{D}^c_{\mathcal{L}} \cup \mathcal{D}^c_{\mathcal{U}}$. In this approach, GCD can be applied to each context to identify categories. However, GCD may struggle with complex ad-hoc categories that lack clear visual similarity. It also often requires extensive labeled data, which may be impractical for many ad-hoc categories.

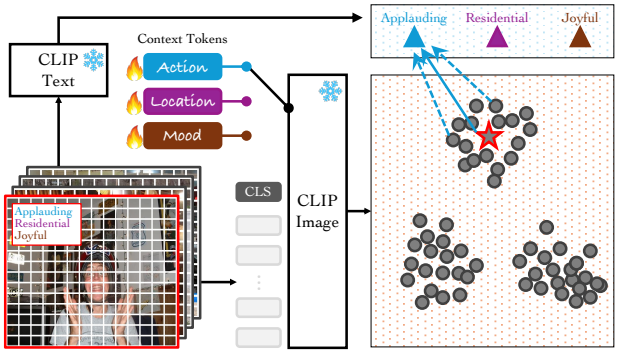


Figure 3. **OAK learns contextualized features** while preserving the foundations of perception of CLIP by introducing context tokens that modulate the frozen ViT encoder, achieving context-aware attention. This contextualized feature learning follows two key principles: 1) top-down text guidance, which leverages semantic knowledge from known class names, and 2) bottom-up image clustering, which captures visual similarity to infer categorization rules. OAK aligns visual clusters with semantic cues by inferring pseudo-labels using the text encoder and refining clusters accordingly. This unified approach outperforms individual methods such as CLIP and GCD, effectively combining their strengths.

3.2. Our OAK framework

We propose *Open Ad-hoc Categorization with Contextualized Feature Learning* (OAK), which combines these two principles (Fig. 3) through two key techniques: **1)** context-aware visual attention, adapting visual embeddings to support bottom-up image clustering, and **2)** text-guided regularization, aligning visual clusters with top-down text guidance. This design leverages the strengths of both top-down (CLIP) and bottom-up (GCD) approaches.

Background. CLIP [39] trains an image encoder f and a text encoder g to align embeddings in a shared space. Pre-

trained on a large set of image-text pairs, CLIP can perform zero-shot classification across diverse contexts by adapting its vocabulary. We further fine-tune CLIP for category discovery, benefiting from its rich semantic knowledge. We assume the image encoder follows the ViT [10] architecture, which divides an image into patches and applies self-attention [48] to process information across spatial tokens. This attention map serves as a natural saliency map, guiding the image encoder to focus on specific regions.

Clustering the visual embeddings $f(\mathbf{x}_i)$ is a natural way to discover new classes. GCD enhances this with a training objective that makes visual embeddings more informative by grouping visually and semantically similar images. This is achieved through self-supervised ($\ell_{\text{self-con}}$) and supervised ($\ell_{\text{sup-con}}$) contrastive losses [20, 57], applied to unlabeled ($\mathcal{D}_{\mathcal{U}}$) and labeled ($\mathcal{D}_{\mathcal{L}}$) datasets, respectively. The combined GCD loss is given by:

$$\ell_{\text{GCD}}(\mathbf{z}) = (1-\lambda)\cdot\ell_{\text{self-con}}(\mathbf{z};\mathcal{D}_{\mathcal{U}}) + \lambda\cdot\ell_{\text{sup-con}}(\mathbf{z};\mathcal{D}_{\mathcal{L}}) \quad (1)$$

where $\mathbf{z}$ denotes trainable parameters, and $\lambda = \lambda_{\text{GCD}}$ is a hyperparameter balancing the two losses. While we use the original GCD in our framework for simplicity, advanced methods like μGCD [51] can also be applied.

Context-aware visual attention. Our key intuition is that the attention map of the image encoder should adapt to the context. For instance, the model should focus on the foreground to predict actions and on the background to predict locations. To achieve this, we introduce a context token $\mathbf{z}_c$ that makes the ViT attention map context-aware. This token is concatenated with the image patch tokens, orchestrating relationships among patches similarly to a register token [9] but tailored to each context. This allows us to obtain a context-specific image encoder f_c by simply adding a context token to the shared ViT backbone:

$$f_c(\mathbf{x}_i) := f([\mathbf{x}_i, \mathbf{z}_c]). \quad (2)$$

The context token is optimized for each context, while the backbone remains frozen, as in visual prompt tuning [16]. This allows our method to adapt to any ad-hoc categorization rules by modulating only the context token.

Text-guided regularization. The GCD objective clusters visually and semantically similar images but does not specify cluster locations, which can disrupt the semantic knowledge of CLIP during fine-tuning. To address this, we regularize the clusters to align with pretrained text embeddings by freezing the text encoder g and fine-tuning solely the image encoder f using the context token $\mathbf{z}_c$. We apply a $|\mathcal{Y}_{\mathcal{L}} \cup \hat{\mathcal{Y}}_{\mathcal{N}}|$-way classification loss between the image and text embeddings, where $\hat{\mathcal{Y}}_{\mathcal{N}}$ represents potential novel classes. These class names are generated by prompting large language models (LLMs), though methods like textual inversion [55] could also be used. For labeled images in $\mathcal{D}_{\mathcal{L}}$, we use the ground-truth labels, and for unlabeled images in $\mathcal{D}_{\mathcal{U}}$, we generate pseudo-labels based on clusters from the semi-supervised K-means (SS-KMeans) algorithm, computed at the start of each training epoch. Specifically, we apply Hungarian matching [22] between the cluster embeddings (averages of the image embeddings) and the text embeddings of candidate words in $\mathcal{Y}_{\mathcal{L}} \cup \hat{\mathcal{Y}}_{\mathcal{N}}$. Formally, our regularizer is given by:

$$\begin{aligned}\ell_{\text{text-reg}}(\mathbf{z}_c) = &\frac{1}{|\mathcal{D}_{\mathcal{L}}|}\sum_{(\mathbf{x}_i,y_i)\in\mathcal{D}_{\mathcal{L}}} \text{CE}\left(p(y_i|\mathbf{x}_i;\mathbf{z}_c), y_i\right) \\ &+ \frac{1}{|\mathcal{D}_{\mathcal{U}}|}\sum_{\mathbf{x}_i\in\mathcal{D}_{\mathcal{U}}} \text{CE}\left(p(\hat{y}_i|\mathbf{x}_i;\mathbf{z}_c), \hat{y}_i\right)\end{aligned} \quad (3)$$

where $p(y_i|\mathbf{x}_i;\mathbf{z}_i)$ is the i-th value of the softmax probability over the cosine similarities between the image embedding $f([\mathbf{x}_i, \mathbf{z}_c])$ and text embeddings $\{g(y_i) \,|\, y_i \in \mathcal{Y}_{\mathcal{L}}\}$, $\hat{y}_i$ is the pseudo-label for an unlabeled image $\mathbf{x}_i$, and CE denotes cross-entropy. This semi-supervised approach is similar to GPC [62], but it incorporates textual information to obtain pseudo-labels from visual clusters. In summary, the full training objective of OAK is:

$$\ell_{\text{OAK}}(\mathbf{z}_c) = \ell_{\text{GCD}}(\mathbf{z}_c) + \lambda_{\text{text-reg}}\cdot\ell_{\text{text-reg}}(\mathbf{z}_c) \quad (4)$$

where $\lambda_{\text{text-reg}}$ is a hyperparameter for the text-guided regularizer. Text regularization is especially important for contexts less familiar to CLIP, having a stronger impact on location and mood than on action contexts.

Inference. After training, we obtain the visual clusters and their pseudo-labels using the same approach as in our semi-supervised learning, providing predictions for both known and novel classes. Unlike GCD, which discovers clusters in the image encoder alone, text-guided regularization allows OAK to assign text labels to each cluster, offering a clearer interpretation of open categories (Tab. 5).

4. Experiments

4.1. Baselines and evaluation metrics

Baselines. We consider two groups of baselines aligned with our principles: 1) top-down text guidance, such as CLIP [39] with an extended vocabulary, and 2) bottom-up image clustering, such as GCD [49]. For the first group, we evaluate the following baselines:

- **CLIP-ZS**: CLIP zero-shot (ZS) classifier relies solely on known class names, serving as a closed-world baseline unable to discover novel categories.
- **CLIP-ZS + LLM vocab**: Extend CLIP-ZS to predict novel classes by generating potential class names using large language models (LLMs).
- **CLIP-ZS + GT vocab**: Apply CLIP-ZS using ground-truth (GT) class names for novel categories, setting the upper bound for zero-shot methods.

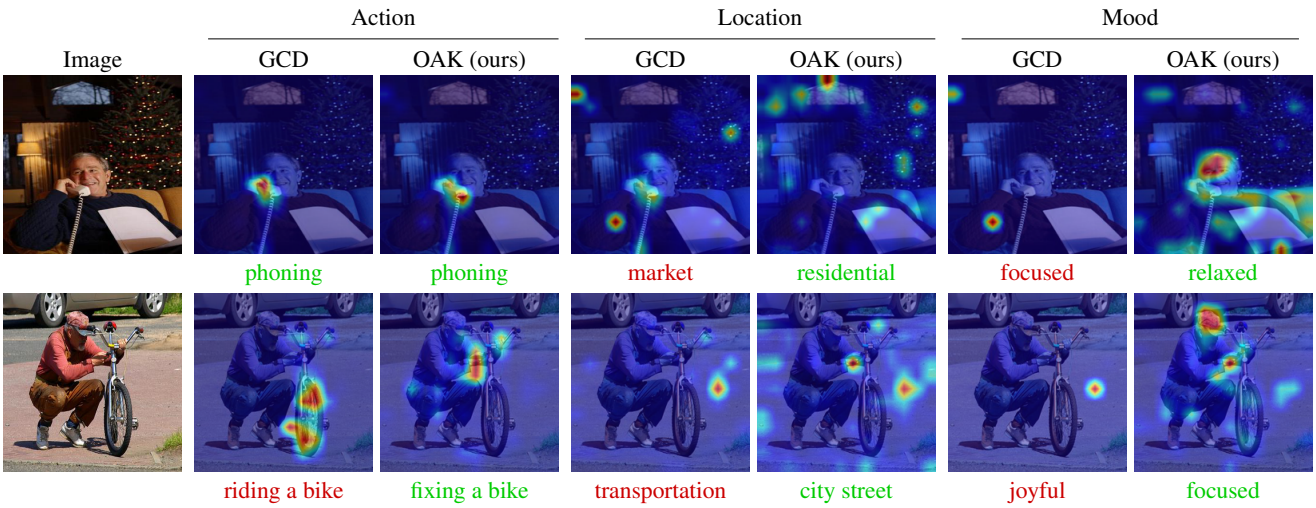


Figure 4. **Saliency maps on the Stanford dataset** show that OAK focuses on relevant regions of images for different contexts, while GCD often distracts to arbitrary regions. We select two samples predicted correctly by OAK across all contexts and visualize the saliency maps of GCD and OAK using the approach of Chefer et al. [5], with predicted classes colored green for correct and red for incorrect predictions. OAK focuses on human behaviors, like *hand movements* for Action, covers the *entire scene* for Location, and highlights a *human face* for Mood, closely aligning with human intuition. GCD produces reasonable saliency maps for Action, as seen in the *phoning* example, but confuses *fixing a bike* with *riding a bike* by focusing on the bike rather than human behavior.

Table 1. **Accuracies on the Stanford dataset** using Action, Location, and Mood contexts show that OAK consistently outperforms open-vocabulary classification (row group 1) and visual clustering (row group 2) baselines, particularly on novel classes and prediction consistency. This advantage is most pronounced in less familiar contexts like Mood. We report known, novel, and overall accuracies for each context, including Omni context, with best results in bold. CLIP-ZS + LLM vocab performs poorly on novel classes, revealing the limitation of using class names alone. GCD addresses this by clustering visual features, but OAK goes further by contextualizing them with CLIP's semantic knowledge, achieving a 50% gain over both CLIP and GCD in Mood. In the Omni context, OAK achieves 70.3% overall accuracy, outperforming all baselines by 2–30% and demonstrating consistency across contexts.

	Known				Novel				Overall			
Method	Action	Location	Mood	Omni[a]	Action	Location	Mood	Omni	Action	Location	Mood	Omni
CLIP-ZS	**96.2**	60.4	**87.4**	75.0	-	-	-	-	-	-	-	-
+ LLM vocab	93.5	58.3	75.9	75.0	38.6	34.2	35.4	0.0	65.2	47.5	55.0	43.0
+ GT vocab	93.6	59.6	78.8	75.0	80.3	59.7	65.8	29.4	86.7	59.7	72.1	38.3
SS-KMeans	67.0	56.3	43.2	0.0	53.9	71.1	78.2	35.3	60.2	62.9	61.3	47.7
GCD	89.4	75.4	64.3	75.0	67.8	80.8	40.6	0.0	78.3	77.8	52.1	52.3
OAK (ours)	88.9	**83.9**	68.8	0.0	**85.1**	**88.4**	**87.4**	**47.1**	**86.9**	**85.9**	**78.4**	**70.3**

[a] With only 8 images overlapping across all contexts, text is toned down to light gray due to unreliable results.

For the second group, we evaluate the following baselines:

- **SS-KMeans**: Extract visual embeddings from CLIP without fine-tuning, then use semi-supervised K-means (SS-KMeans) clustering to discover novel classes.
- **GCD**: Fine-tune CLIP with the GCD loss and use SS-KMeans clustering to discover novel classes.

We reimplement all baselines and our method within a consistent experimental setup, using a ViT-B/16 image encoder with the CLIP weights released by OpenAI. Following the GCD training recipe, we adjust only the learning rates and epochs for each dataset. For a fair comparison, the same training procedure is applied across all methods.

Additional baselines. We primarily focus on a few-shot setup, as collecting extensive labels for diverse ad-hoc contexts is impractical. Consequently, our reimplemented baselines may underperform compared to values reported in prior works that use the entire labeled sets. For further validation, we also report comparisons with state-of-the-art methods in the full-shot scenario (Tab. 3), which includes GCD [49], SimGCD [54], and μGCD [51].

Evaluation metrics. Following the standard for generalized category discovery [49], we report accuracy for known ($\mathcal{D}_{\mathcal{U}_\mathcal{L}}$), novel ($\mathcal{D}_{\mathcal{U}_\mathcal{N}}$), and overall ($\mathcal{D}_\mathcal{U}$) classes. Novel accuracy varies by method: CLIP-ZS uses an extended vocabulary, while GCD applies Hungarian matching. Each method is evaluated independently within each context us-

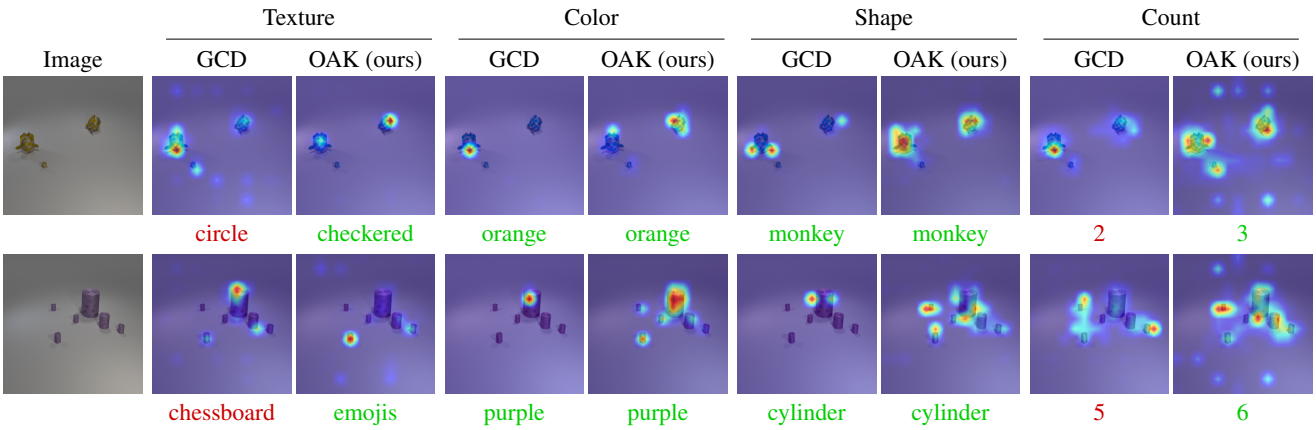


Figure 5. **Saliency maps on the Clevr-4 dataset** show consistent results with those in Fig. 4, following the same setup, demonstrating that OAK effectively adapts its saliency maps to each context. In the Texture and Color contexts, it focuses on *small regions*, sufficient for predicting simple visual concepts. In the Shape context, it attends to *multiple views* of objects, as different 2D views are needed for accurate 3D shape understanding. For the Count context, OAK attends to *all objects* to identify each one accurately. In contrast, GCD often fails to focus on the proper regions, missing objects in the Count context and underestimating the object count.

Table 2. **Accuracies on the Clevr-4 dataset** using Texture, Color, Shape, and Count contexts follow the same setup as in Tab. 1 and show consistent results, with OAK outperforming both CLIP and GCD baselines. CLIP-ZS performs poorly across all contexts due to limited familiarity with synthetic images and abstract concepts, unlike the natural classes in Stanford. GCD does well on Color and Shape, which depend on clear visual cues, but struggles with Count, which requires higher-level reasoning. OAK shows a similar trend but consistently outperforms GCD, achieving reliable gains even when CLIP's semantic knowledge is less effective on synthetic data.

	Known					Novel					Overall				
Method	Texture	Color	Shape	Count	Omni	Texture	Color	Shape	Count	Omni	Texture	Color	Shape	Count	Omni
CLIP-ZS	40.8	94.4	79.7	**46.9**	14.6	-	-	-	-	-	-	-	-	-	-
+ LLM vocab	29.4	81.9	56.5	21.7	3.3	20.7	44.1	49.5	24.4	0.6	25.0	62.8	53.3	23.1	1.7
+ GT vocab	13.3	79.8	60.1	21.7	3.1	26.5	62.7	64.8	24.4	2.4	20.0	71.2	62.3	23.1	2.0
SS-KMeans	12.9	10.4	73.2	24.2	0.5	13.7	13.0	82.8	15.5	0.2	13.4	11.7	77.6	19.8	0.1
GCD	73.4	98.3	99.0	41.9	35.5	43.6	94.9	99.2	42.3	15.7	58.2	96.6	99.1	42.1	22.6
OAK (ours)	**82.3**	**100.0**	**99.9**	45.0	**40.5**	**47.8**	**100.0**	**99.8**	**43.7**	**16.5**	**64.6**	**100.0**	**99.8**	**44.4**	**28.5**

Table 3. **Comparison with state-of-the-art methods** on Clevr-4, using full-shot labels. We report novel accuracies from Table 4 of Vaze et al. [51]. OAK outperforms the baselines, particularly in the challenging Texture context, with an 11% gain over μGCD. Prior work showed that larger models are not always effective for all contexts, training ResNet-18 [13] from scratch. In contrast, our results suggest that OAK effectively addresses all contexts.

Method	Texture	Color	Shape	Count	Avg.
GCD [49]	45.3	90.5	88.5	60.1	71.1
SimGCD [54]	40.2	97.2	95.1	53.9	71.6
μGCD [51]	55.5	92.1	**99.2**	65.2	78.0
OAK (ours)	**66.5**	**99.9**	99.0	**67.6**	**83.3**

ing its respective labeled set. We also report Omni accuracy, which measures correctness across all contexts. Formally, for images $\{\mathbf{x}_i\}_{i=1}^N$ and contexts $c \in \mathcal{C}$, with $\hat{y}^c$ and y^c as predicted and true labels, the Omni accuracy is:

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left(\bigwedge_{c\in\mathcal{C}} \hat{y}_i^c = y_i^c\right). \tag{5}$$

4.2. Stanford results

Dataset. We use the Stanford [23] dataset, originally collected for 40 Action classes [59] and later annotated with 10 Location and 4 Mood classes. For each context, we randomly select half of the classes as known and the rest as novel, sampling 16 labeled images per class. Each context contains only high-confidence images: 10K for Action and 1K each for Location and Mood. These sets overlap but are not identical. To evaluate Omni accuracy, we identified 127 overlapping images and manually annotated them. Among these, 8 belong to known and 17 to novel classes in our split.

Saliency maps. Our key insight is that the label of an image depends on its context, so visual features should be contextualized accordingly. To validate this, we compare saliency maps from OAK and GCD using the method of [5]. Fig. 4 shows saliency maps on the Stanford dataset, suggesting that OAK aligns closely with human intuition. It focuses on hands for Action, the entire scene for Location, and faces for Mood, demonstrating strong interpretability. In contrast, GCD often attends to irrelevant regions, partic-

Table 4. **Ablation study on our method components** shows that both context-aware attention (Eq. (2)) and text-guided regularization (Eq. (3)) contribute to performance. Context-aware attention improves novel accuracy while often reducing known accuracy by adapting the visual encoder to new contexts. In contrast, text-guided regularization boosts known accuracy but may harm novel accuracy by restricting clusters with frozen text embeddings. Combined, our final OAK achieves the best results in novel and overall accuracies, highlighting the importance of both bottom-up image clustering and top-down text guidance for open ad-hoc categorization.

Context-aware attention	Text-guided regularization	Known				Novel				Overall			
		Action	Location	Mood	Omni	Action	Location	Mood	Omni	Action	Location	Mood	Omni
-	-	89.4	75.4	64.3	75.0	67.8	80.8	40.6	0.0	78.3	77.8	52.1	52.3
✓	-	83.1	61.4	44.7	0.0	73.3	75.0	51.8	23.5	78.1	67.5	48.3	11.7
-	✓	**95.7**	**87.6**	66.9	50.0	63.2	72.3	47.4	0.0	79.0	80.0	56.8	43.8
✓	✓	88.9	83.9	**68.8**	0.0	**85.1**	**88.4**	**87.4**	**47.1**	**86.9**	**85.9**	**78.4**	**70.3**

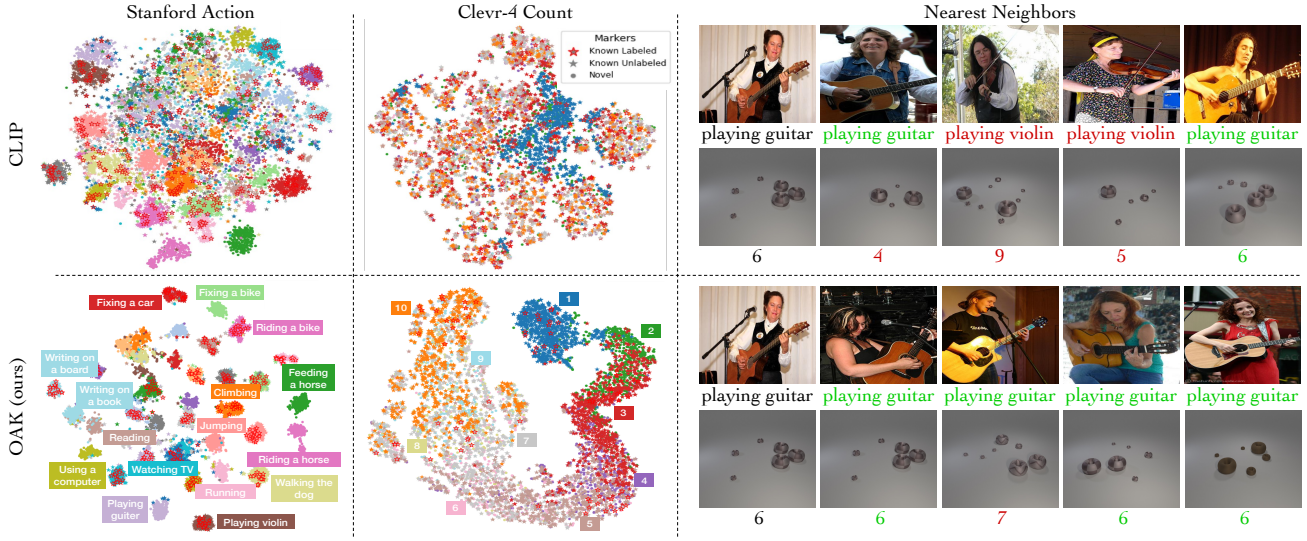


Figure 6. **t-SNE visualization** of CLIP visual embeddings and nearest neighbor examples from CLIP (row 1) and OAK (row 2) on Stanford Action and Clevr-4 Count. CLIP embeddings are not grouped by context, as seen in the Count example where most classes except *1* are clustered together. In contrast, OAK contextualizes the feature space so that both known and novel classes form meaningful groups. In the Action context, OAK produces well-separated clusters and demonstrates compositional generalization. For example, it discovers the novel concept of *fixing a bike* by leveraging the knowledge of *fixing a car* and *feeding a horse*. In the Count context, OAK arranges classes in a smooth progression that reflects the ordinal nature of numbers. For instance, *2* is placed between *1* and *3*. The nearest neighbor examples support this: CLIP retrieves somewhat arbitrary images, while OAK consistently returns similar ones, like *7* for *6*.

ularly in Location and Mood where the domain gap from CLIP is larger, which leads to prediction errors.

Accuracies. We evaluate the effectiveness of OAK by extending the GCD benchmarks. Tab. 1 reports accuracies on the Stanford dataset. OAK outperforms all baselines in both novel and overall accuracy, showing that contextualized features improve predictions. It combines the strengths of CLIP and GCD, surpassing both. While CLIP performs well on known classes, it cannot discover novel ones, which is essential for open categorization. In contrast, OAK performs well on both known and novel classes.

4.3. Clevr-4 results

Dataset. We use Clevr-4 [51], a synthetic dataset generated in the CLEVR [17] environment, with 10 classes per context: Texture, Color, Shape, and Count. For each context, half the classes are randomly selected as known, and the rest as novel, with 16 images per class sampled for the labeled set. All images are annotated with all contexts, using the full dataset of 8.3K images to evaluate Omni accuracy.

Saliency maps. We further verify the capability of OAK using the same procedure as with the Stanford dataset. SM B shows saliency maps of GCD and OAK on the Clevr-4 dataset. OAK adapts its attention based on context complexity, using fewer queries for simple concepts and more for complex ones. For example, it attends to small regions for Texture and Color, considers multiple 2D angles for Shape, and focuses on all objects for Count. In contrast, GCD fails to adjust its attention appropriately, often missing key regions such as objects in the Count context.

Accuracies. Tab. 2 presents GCD benchmark results on the Clevr-4 dataset. As with previous results, OAK out-

Table 5. **Class names associated with novel visual clusters** from our models show that it identifies reasonable words for contexts familiar to CLIP (Action, Location, Color, Shape), but less accurate ones for more unfamiliar contexts (Mood, Texture, Count). We demonstrate true and predicted names, showing two classes per context, with full lists and visual examples in SM C. For familiar contexts, the predicted names align with synonyms, such as *preparing a meal* for *cooking* in Action or *turquoise* for *cyan* in Color. In contrast, for unfamiliar contexts, the names are often unrelated, such as an antonym of *admiring* for *relaxed* in Mood.

a) Stanford	Action		Location		Mood	
GT Label	cooking	reading	restaurant or dining area	urban area or city street	focused	relaxed
Prediction	preparing a meal	reading a book	commercial kitchen	suburban street	explorative	admiring

b) Clevr-4	Texture		Color		Shape		Count	
GT Label	zigzag	circles	pink	cyan	star	diamond	4	9
Prediction	wavy lines	checkerboard	pink	turquoise	star shape	diamond shape	4	17

Figure 7. **Visual examples with true and predicted class names**, showing OAK assigns reasonable labels based on visual cues, such as *jumping* for people appearing *dancing*.

performs all baselines in both novel and overall accuracy, surpassing CLIP and GCD. In this setting, CLIP is less effective due to limited familiarity with synthetic images and abstract contexts, which causes it to fall behind GCD. Nonetheless, textual guidance still benefits OAK, enabling it to consistently outperform GCD. Additional ablation on the role of text in Clevr-4 is provided in SM E.1.

Full-shot results. OAK is effective not only in few-shot ad-hoc categorization but also in broader GCD setups. We evaluate full-shot learning on Clevr-4 using 2K labels per context. Tab. 3 shows that OAK significantly outperforms the state-of-the-art μGCD [51] in novel accuracy, achieving near-optimal results for Color and Shape while improving performance on Texture and Count, including an 11% gain on Texture. Although μGCD argues that no single representation works across all contexts and that pretrained models often fail in abstract domains like Clevr-4, OAK succeeds in all cases through its unified feature contextualization.

4.4. Ablation study and analysis

Ablation study. We study the effect of the components of OAK. Tab. 4 shows that both components contribute to performance, with context-aware attention improving novel accuracy and text-guided regularization enhancing known accuracy. Combining them achieves the best results, confirming the need for both top-down text guidance and bottom-up image clustering to learn categorization rules.

t-SNE visualization. We use t-SNE [47] plots to visualize embeddings from CLIP and OAK. As shown in Fig. 6, OAK contextualizes features, resulting in more discriminative representations. Detailed results for each context are available in SM E.5. In most cases with distinct classes, the clusters are well separated. For the count context in Clevr-4, however, the clusters learned by OAK shift gradually, reflecting smooth transitions between classes.

Naming clusters. OAK can name visual clusters by aligning image and text embeddings. Tab. 5 lists true and predicted class names for novel clusters, with candidate names generated by LLMs. OAK uncovers reasonable names for novel visual concepts, such as *preparing a meal* for *cooking* images. It performs well in contexts familiar to CLIP, like Action, and also in the challenging context of Count, a concept CLIP finds particularly difficult to understand [36], suggesting OAK can learn new concepts from visual clusters. Fig. 7 visualizes predicted class names that differ from true labels, illustrating how naming corresponds to visual patterns, such as images in the *dancing* cluster appearing like *jumping* despite their true label.

Additional results on saliency maps, cluster names, and t-SNE plots, and more analyses are provided in the SM.

Summary. We study open ad-hoc categorization, which aims to predict both known and novel classes in unlabeled data across varying contexts. We propose OAK, a context-token-based feature modulation method for CLIP that integrates visual and semantic information. This task is a key step toward building AI agents capable of seamlessly handling diverse real-world scenarios. While our work focuses on static image recognition, extending ad-hoc categorization to broader settings is an interesting future direction.

Acknowledgments. This project was supported, in part, by Bosch gift funds to S. Yu at UC Berkeley and the University of Michigan.

References

[1] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022. 3

[2] Lawrence W Barsalou. Ad hoc categories. *Memory & cognition*, 11:211–227, 1983. 1

[3] Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. *arXiv preprint arXiv:2102.03526*, 2021. 3

[4] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 20

[5] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *CVPR*, 2021. 5, 6

[6] Sua Choi, Dahyun Kang, and Minsu Cho. Contrastive mean-shift learning for generalized category discovery. In *CVPR*, 2024. 3

[7] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 18

[8] Alessandro Conti, Enrico Fini, Massimiliano Mancini, Paolo Rota, Yiming Wang, and Elisa Ricci. Vocabulary-free image classification. In *NeurIPS*, 2023. 3

[9] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*, 2024. 4

[10] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3, 4

[11] Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *CVPR*, 2023. 3

[12] Kai Han, Xiaohu Huang, Yandong Li, Sagar Vaze, Jie Li, and Xuhui Jia. What's in a name? beyond class indices for image recognition. *arXiv preprint arXiv:2304.02364*, 2023. 3

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[14] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169, 2021. 2

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 3

[16] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022. 3, 4

[17] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 7, 12

[18] Hyunwoo Kang, Sangwoo Mo, and Jinwoo Shin. Oamixer: Object-aware mixing layer for vision transformers. *arXiv preprint arXiv:2212.06595*, 2022. 3

[19] Tsung-Wei Ke, Sangwoo Mo, and X Yu Stella. Learning hierarchical image segmentation for recognition and by recognition. In *ICLR*, 2023. 3

[20] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020. 4

[21] Younghyun Kim, Sangwoo Mo, Minkyu Kim, Kyungmin Lee, Jaeho Lee, and Jinwoo Shin. Discovering and mitigating visual biases through keyword explanation. In *CVPR*, 2024. 3

[22] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 4

[23] Sehyun Kwon, Jaeseung Park, Minkyu Kim, Jaewoong Cho, Ernest K Ryu, and Kangwook Lee. Image clustering conditioned on text criteria. In *ICLR*, 2024. 2, 3, 6, 12

[24] George Lakoff. *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago press, 2008. 1

[25] Mingxuan Liu, Subhankar Roy, Wenjing Li, Zhun Zhong, Nicu Sebe, and Elisa Ricci. Democratizing fine-grained visual recognition with large language models. In *ICLR*, 2024. 3

[26] Mingxuan Liu, Zhun Zhong, Jun Li, Gianni Franchi, Subhankar Roy, and Elisa Ricci. Organizing unstructured image collections using natural language. *arXiv preprint arXiv:2410.05217*, 2024. 3

[27] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, 2019.

[28] Yulin Luo, Ruichuan An, Bocheng Zou, Yiming Tang, Jiaming Liu, and Shanghang Zhang. Llm as dataset analyst: Subpopulation structure discovery with large language model. In *ECCV*, 2024. 3

[29] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *ICLR*, 2023. 3

[30] George A Miller, Galanter Eugene, and Karl H Pribram. Plans and the structure of behaviour. In *Systems Research for Behavioral Science*, pages 369–382. Routledge, 2017. 2

[31] Sangwoo Mo, Hyunwoo Kang, Kihyuk Sohn, Chun-Liang Li, and Jinwoo Shin. Object-aware contrastive learning for debiased scene representation. In *NeurIPS*, 2021. 3

[32] Sangwoo Mo, Minkyu Kim, Kyungmin Lee, and Jinwoo Shin. S-clip: Semi-supervised vision-language learning using few specialist captions. In *NeurIPS*, 2023. 3

[33] Sangwoo Mo, Jong-Chyi Su, Chih-Yao Ma, Mido Assran, Ishan Misra, Licheng Yu, and Sean Bell. Ropaws: Robust semi-supervised representation learning from uncurated data. In *ICLR*, 2023.

[34] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *NeurIPS*, 2018.

[35] Rabah Ouldnoughi, Chia-Wen Kuo, and Zsolt Kira. Clip-gcd: Simple language guided generalized category discovery. *arXiv preprint arXiv:2305.10420*, 2023. 3

[36] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *ICCV*, 2023. 8

[37] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *ICCV*, 2023. 3

[38] Nan Pu, Zhun Zhong, and Nicu Sebe. Dynamic conceptional contrastive learning for generalized category discovery. In *CVPR*, 2023. 3

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 3, 4

[40] Sarah Rastegar, Mohammadreza Salehi, Yuki M Asano, Hazel Doughty, and Cees GM Snoek. Selex: Self-expertise in fine-grained generalized category discovery. In *ECCV*, 2024. 3

[41] Eleanor Rosch. Principles of categorization. *Cognition and categorization/Erlbaum*, 1978. 1

[42] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *ICML*, 2018.

[43] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 2

[44] Baifeng Shi, Trevor Darrell, and Xin Wang. Top-down visual attention from analysis by synthesis. In *CVPR*, 2023. 3

[45] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017. 2

[46] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. In *NeurIPS*, 2020.

[47] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 2008. 8

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 4

[49] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *CVPR*, 2022. 1, 3, 4, 5, 6, 13

[50] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need? In *ICLR*, 2022.

[51] Sagar Vaze, Andrea Vedaldi, and Andrew Zisserman. No representation rules them all in category discovery. In *NeurIPS*, 2023. 2, 3, 4, 5, 6, 7, 8, 12

[52] Enguang Wang, Zhimao Peng, Zhengyuan Xie, Xialei Liu, and Ming-Ming Cheng. Get: Unlocking the multi-modal potential of clip for generalized category discovery. *arXiv preprint arXiv:2403.09974*, 2024. 3

[53] Hongjun Wang, Sagar Vaze, and Kai Han. Sptnet: An efficient alternative framework for generalized category discovery with spatial prompt tuning. In *ICLR*, 2024. 3

[54] Xin Wen, Bingchen Zhao, and Xiaojuan Qi. Parametric classification for generalized category discovery: A baseline study. In *ICCV*, 2023. 3, 5, 6

[55] Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. In *NeurIPS*, 2024. 4

[56] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *CVPR*, 2022. 3

[57] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 4

[58] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *IJCV*, 2024.

[59] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *ICCV*, 2011. 6, 12

[60] Jiawei Yao, Qi Qian, and Juhua Hu. Multi-modal proxy learning towards personalized visual multiple clustering. In *CVPR*, 2024. 3

[61] Sheng Zhang, Salman Khan, Zhiqiang Shen, Muzammal Naseer, Guangyi Chen, and Fahad Shahbaz Khan. Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. In *CVPR*, 2023. 3

[62] Bingchen Zhao, Xin Wen, and Kai Han. Learning semi-supervised gaussian mixture models for generalized category discovery. In *ICCV*, 2023. 3, 4

[63] Qihao Zhao, Yalun Dai, Hao Li, Wei Hu, Fan Zhang, and Jun Liu. Ltgc: Long-tail recognition via leveraging llms-driven generated content. In *CVPR*, 2024. 3

[64] Haiyang Zheng, Nan Pu, Wenjing Li, Nicu Sebe, and Zhun Zhong. Textual knowledge matters: Cross-modality co-teaching for generalized visual class discovery. *arXiv preprint arXiv:2403.07369*, 2024. 3

[65] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3

Open Ad-hoc Categorization with Contextualized Feature Learning

Supplementary Material

Contents

A. Experimental details

A.1. Dataset details

The Stanford Action dataset [59] is available from its official website, while the Stanford Location and Stanford Mood dataset [23] can be downloaded from its official GitHub page. We generate a text file containing filenames and ground-truth labels for each dataset. In the smaller Stanford Location and Stanford Mood dataset, we retain all filenames present in Stanford Action and use a special symbol to indicate missing images. All available images from these datasets are used. For Clevr-4, the datasets [51] constructed using the CLEVR environment [17] are available on the authors' GitHub page, and we use them without additional preprocessing, utilizing only the training split for our method. We provide the dataset statistics in Tab. 6, complete class names in Tab. 7, evaluation set sizes (overlap across all context) in Tab. 8.

Table 6. **Dataset statistics** used in our experiments. We randomly split the classes, assigning half as known ($\mathcal{Y}_\mathcal{L}$) and the other half as novel ($\mathcal{Y}_\mathcal{N}$), sampling 16 images per class for the labeled set ($\mathcal{D}_\mathcal{L}$) and using the remaining images for the unlabeled set ($\mathcal{D}_\mathcal{U}$) in each context. This setup reflects the practical scenario of ad-hoc categorization, where obtaining extensive labels for diverse contexts is challenging. Please note that our results are not directly comparable to prior work, which often uses thousands of labeled samples.

a) Stanford	Action	Location	Mood
Examples	`drinking`, `phoning`	`market`, `residential`	`focused`, `relaxed`
$\lvert\mathcal{Y}_\mathcal{L}\rvert$	20	5	2
$\lvert\mathcal{Y}_\mathcal{N}\rvert$	20	5	2
$\lvert\mathcal{D}_\mathcal{L}\rvert$	320	80	32
$\lvert\mathcal{D}_\mathcal{U}\rvert$	9.2K	920	968

b) Clevr-4	Texture	Color	Shape	Count
Examples	`metal`, `rubber`	`red`, `blue`	`torus`, `cube`	`1`, `2`
$\lvert\mathcal{Y}_\mathcal{L}\rvert$	5	5	5	5
$\lvert\mathcal{Y}_\mathcal{N}\rvert$	5	5	5	5
$\lvert\mathcal{D}_\mathcal{L}\rvert$	80	80	80	80
$\lvert\mathcal{D}_\mathcal{U}\rvert$	8.3K	8.3K	8.3K	8.3K

Table 7. **Class names for each dataset.** Classes in **bold** represent the known classes for the respective datasets.

Dataset	Class Names
Stanford Action	**applauding, brushing teeth, climbing, cutting trees, drinking, fishing, fixing a car, holding an umbrella, looking through a microscope, phoning, playing violin, pushing a cart, riding a bike, rowing a boat, shooting an arrow, taking photos, throwing frisby, walking the dog, watching TV, writing on a board**, blowing bubbles, cleaning the floor, cooking, cutting vegetables, feeding a horse, fixing a bike, gardening, jumping, looking through a telescope, playing guitar, pouring liquid, reading, riding a horse, running, smoking, texting message, using a computer, washing dishes, waving hands, writing on a book
Stanford Location	**educational institution, natural environment, office or workplace, public event or gathering, residential area**, restaurant or dining area, sports facility, store or market, transportation hub, urban area or city street
Stanford Mood	**adventurous, joyful**, focused, relaxed
Clevr-4 Texture	**rubber, metal, checkered, emojis, wave**, brick, star, circles, zigzag, chessboard
Clevr-4 Color	**gray, red, blue, green, brown**, purple, cyan, yellow, pink, orange
Clevr-4 Shape	**cube, sphere, monkey, cone, torus**, star, teapot, diamond, gear, cylinder
Clevr-4 Count	**7, 10, 1, 3, 5**, 2, 4, 6, 8, 9

Table 8. **Omni accuracy evaluation set sizes for each dataset.** To compute the Omni accuracy, we gather all images labeled across all contexts for the evaluation set. For instance, this table shows that only 8 images overlap between the known image sets of the Action, Location, and Mood contexts in the Stanford dataset.

	Known	Novel	Overall
Stanford	8	17	128
Clevr-4	583	496	8,424

A.2. Training of GCD and OAK

We begin each experiment using the exact training recipe from GCD [49]. However, we find the default hyperparameters lead to ineffective and unstable training due to the reduced number of labeled examples, the overall dataset size (Stanford Location and Mood), and out-of-distribution settings (Clevr-4). To address this, we perform hyperparameter tuning directly on the unlabeled images in the training set for each dataset, based on the training loss curves and clustering quality based on the silhouette score. The silhouette score evaluates the quality of clustering by measuring how similar data points are within the same cluster compared to points in other clusters, which is an effective estimator of how well our model understands current context and discovery open categories. A separate validation set is also suboptimal for this task, as category discovery relies on the grouping of similar images, making dataset size critical. The hyperparameters used are detailed in Tab. 9.

Table 9. **Hyperparameters for training OAK. on Stanford and Clevr-4.** We start from GCD training recipe and perform unsupervised hyperparameter tuning based on training loss curves and clustering quality. CFJ = [RandomCrop, RandomHorizontalFlip, ColorJitter].

Hyperparameter	Stanford			Clevr-4			
	Action	Location	Mood	Texture	Color	Shape	Count
batch_size	128	128	128	128	128	128	128
total_epochs	50	50	50	50	50	50	50
learning_rate	0.1	0.01	0.1	0.01	0.1	0.1	0.01
learning_rate_scheduler	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine
min_learning_rate_multiplier	1e-3	1e-3	1e-3	1e-3	1e-3	1e-3	1e-3
optimizer	SGD	SGD	SGD	SGD	SGD	SGD	SGD
momentum	0.9	0.9	0.9	0.9	0.9	0.9	0.9
weight_decay	5e-5	5e-5	5e-5	5e-5	5e-5	5e-5	5e-5
context_tokens_length	50	50	50	50	50	50	50
$\ell_{\text{self-con}}$: temperature	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$\ell_{\text{self-con}}$: n_views	2	2	2	2	2	2	2
$\ell_{\text{self-con}}$: augmentation	CFJ	CFJ	CFJ	CFJ	CFJ	CFJ	CFJ
$\ell_{\text{sup-con}}$: λ (loss weight)	0.35	0.35	0.35	0.35	0.35	0.35	0.35
$\ell_{\text{text-reg}}$: $\lambda_{\text{text-reg}}$ (labeled, unlabeled)	(0.1, 0.01)	(1.0,1.0)	(1.0, 0.1)	(1.0,1.0)	(0.1, 0.1)	(0.1, 1.0)	(1.0, 1.0)
SS-KMeans: n_init	10	10	10	10	10	10	10
SS-KMeans: tolerance	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4
SS-KMeans: max_iterations	200	200	200	200	200	200	200

Assumptions beyond GCD. We remark that OAK makes no additional distributional assumptions beyond GCD. Both methods assume the number of novel classes is known, though they are capable of estimating it. The only difference is access to a pool of class names, which can be generated by an LLM at minimal cost.

Class names for GCD. Class names are assigned via Hungarian matching between predicted cluster IDs and true labels across all images, based on one-hot label distance for both GCD and Ours. This is only for visualization (Figs. 4 and 5), as true labels are unavailable in practice. Instead, our cluster names (Tab. 5) are inferred by matching cluster centers of image embeddings with text embeddings of class names via CLIP using cosine similarity.

A.3. LLM prompt for CLIP-ZS and OAK

To adapt CLIP zero-shot methods for predicting novel classes, we generate potential novel class names using the publicly available ChatGPT. We provide the known class names, the number of novel classes required, and a specific prompt to ChatGPT, then use the generated responses as the discovered novel class names for zero-shot classification. OAK's text regularization algorithm and naming clustering algorithm for the unlabeled images follows a similar pipeline, with the key difference being that we request a significantly (up to 4 times) larger vocabulary from ChatGPT to construct our constrained vocabulary set. Our prompt used is detailed below:

> *I have a dataset of images from the following classes:* ***[KNOWN_CLASSES]****. What are the most possible classes that will also be included in this dataset? Give me* ***[NUMBER_OF_NOVEL_CLASSES]*** *class names, only return class names separated by commas. Include quotation marks for each one.*

B. Additional saliency maps

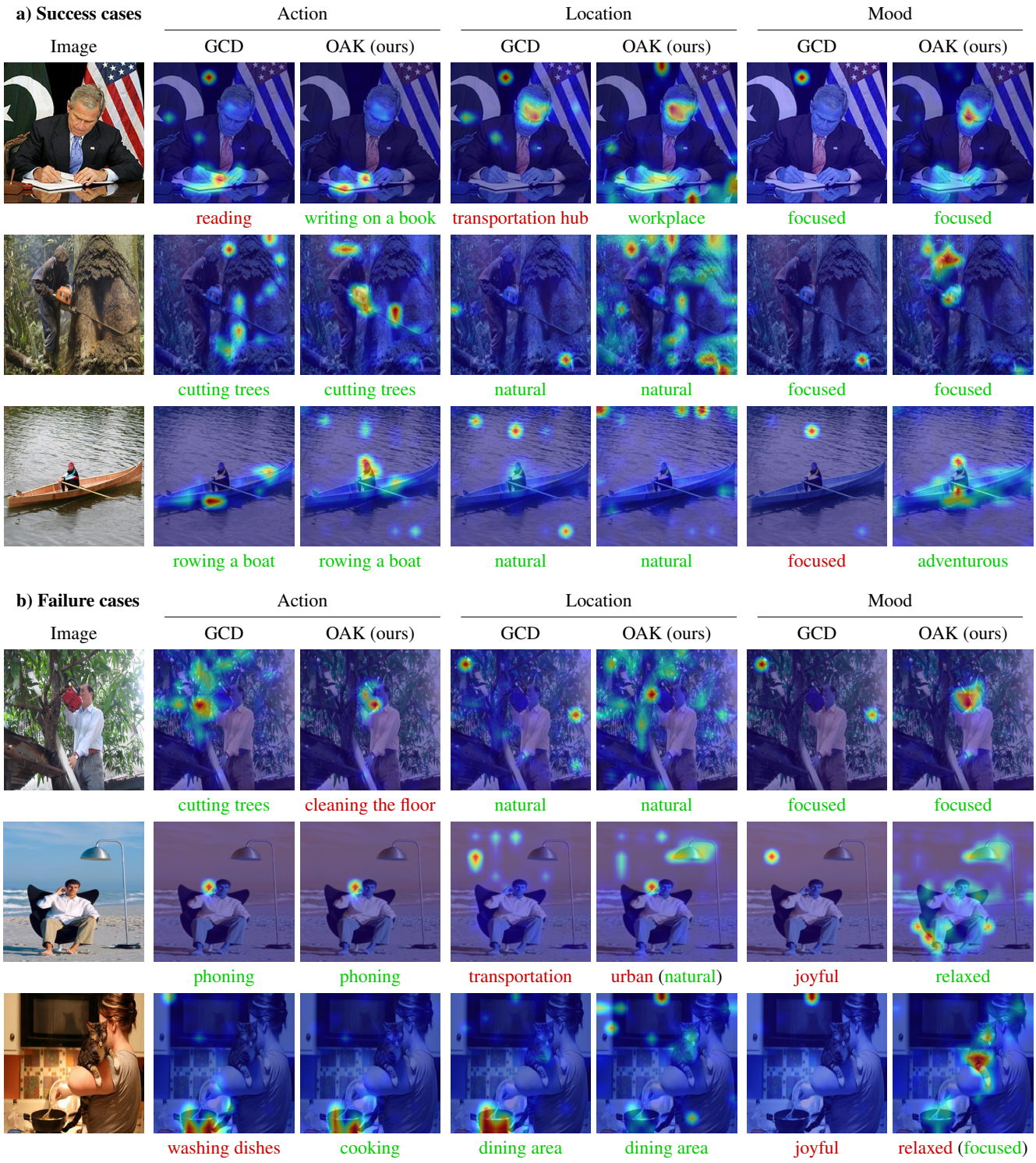


Figure 8. **Additional saliency maps on the Stanford dataset** demonstrate that OAK makes reasonable predictions, focusing on the relevant regions for different contexts. We select three samples correctly predicted by OAK across all contexts and three that fail. In the failure cases, OAK 1) ignores the *trees* with indirect interaction, mistaking the red saw for a *cleaning* tool; 2) focuses on a lamp and a phone in a *natural* beach scene, mistaking it for *urban*; and 3) focuses on the *relaxed* cat held by a *focused* person closer to the camera.

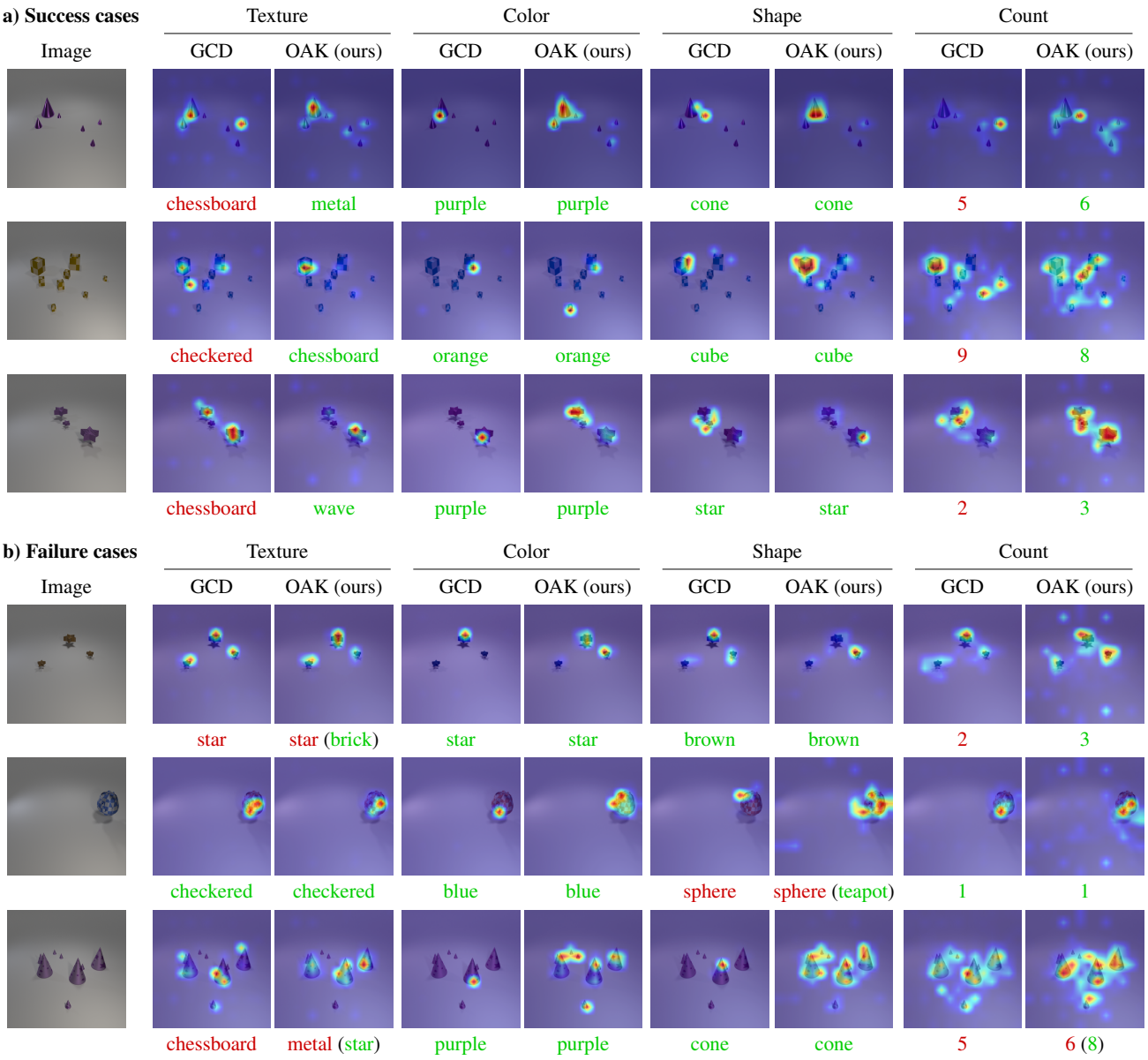


Figure 9. **Additional saliency maps on the Clevr-4 dataset**, showing that OAK makes sensible predictions by focusing on relevant regions across various contexts, using the same setup. In the failure cases, OAK 1) struggles to identify black *brick* patterns on a dark *brown* object, mistaking the *star* shape for a *star* texture; 2) fails to recognize a *teapot* at a challenging angle, mistaking it for a *sphere*; and 3) has difficulty with smaller objects, leading to undercounting. Best viewed zoomed in.

We present additional saliency maps for both the success cases and failure cases on the Stanford datasets and Clevr-4 datasets in Fig. 8 and Fig. 9 respectively. OAK effectively switches between contexts, appropriately focusing on different aspects (regions) of the same image based on the context. Even in failure cases, the errors are easily interpretable and often arise from inherent ambiguities within the image, such as focusing on *manufactured objects* like a lamp and a phone, leading to mispredicting *nature* as *urban*, as illustrated in the second failure example in the Stanford results.

C. Full list of cluster names

Following Tab. 5, we present the class names associated with every known or novel visual clusters for Stanford Action, Stanford Location, Stanford Mood, Clevr-4 Texture, Clevr-4 Color, Clevr-4 Shape and Clevr-4 Count, in Tabs. 10 to 16. OAK identifies novel clusters accurately, as shown by predicted names like *blowing bubbles* in Stanford Action. Failure cases are also fairly reasonable, such as predicting *waving hands* as *clapping*.

Table 10. **Class names associated with every visual cluster from Stanford Action.** OAK's predictions largely align with the ground-truth labels provided by humans, often differing only in synonymous terms, with a few exceptions, such as the *texting message* cluster being predicted as *shaking hands*. Known classes are marked in **bold**.

GT Label	Prediction
applauding	**applauding**
brushing teeth	**brushing teeth**
climbing	**rock climbing**
cutting trees	**cutting trees**
drinking	**drinking**
fishing	**catching a fish**
fixing a car	**fixing a car**
holding an umbrella	**holding an umbrella**
looking through a microscope	**looking in a microscope**
phoning	**talking on a phone**
playing violin	**playing violin**
pushing a cart	**pushing a cart**
riding a bike	**riding a bike**
rowing a boat	**rowing a boat**
shooting an arrow	**practicing archery**
taking photos	**taking photos**
throwing frisby	**fishing**
walking the dog	**walking the dog**
watching TV	**watching TV**
writing on a board	**writing on a board**
blowing bubbles	blowing bubbles
cleaning the floor	mopping the floor
cooking	preparing a meal
cutting vegetables	climbing
feeding a horse	petting a horse
fixing a bike	fixing a bike
gardening	weeding a garden
jumping	dancing
looking through a telescope	looking through a microscope
playing guitar	strumming a guitar
pouring liquid	carrying a box
reading	reading a book
riding a horse	running
running	jogging
smoking	smoking
texting message	shaking hands
using a computer	texting
washing dishes	washing dishes
waving hands	clapping
writing on a book	writing a letter

Table 11. **Class names associated with every visual cluster from Stanford Location.** OAK's predictions often surpass the ground-truth labels in precision, capturing finer semantic meanings with greater granularity. We verify the correctness of these finer predictions through manual visual inspection. For example, many *educational institutions* in our dataset are specifically *science labs*, and many *sports facilities* are *rock climbing walls*. Known classes are marked in **bold**.

GT Label	Prediction
educational institution	**science lab**
natural environment	**natural environment**
office or workplace	**office or workplace**
public event or gathering	**public event or gathering**
residential area	**residential area**
restaurant or dining area	commercial kitchen
sports facility	rock climbing wall
store or market	road or highway
transportation hub	language school
urban area or city street	suburban street

Table 12. **Class names associated with every visual cluster from Stanford Mood.** Known classes are marked in **bold**.

GT Label	Prediction
adventurous	**adventurous**
joyful	**exhilarated**
focused	explorative
relaxed	admiring

Table 13. **Class names associated with every visual cluster from Clevr-4 Texture.** Known classes are marked in **bold**.

GT Label	Prediction
checkered	**checkered**
emojis	**emojis**
metal	**metal**
rubber	**rubber**
wave	**wave**
brick	abstract wave
chessboard	chrome
circles	checkerboard
star	pixelated
zigzag	wavy lines

Table 14. **Class names associated with every visual cluster from Clevr-4 Color.** Known classes are marked in **bold**.

GT Label	Prediction
blue	**indigo blue**
brown	**warm brown**
gray	**gray**
green	**kelly green**
red	**scarlet red**
cyan	turquoise
orange	orange
pink	pink
purple	lilac purple
yellow	mustard yellow

Table 15. **Class names associated with every visual cluster from Clevr-4 Shape.** Known classes are marked in **bold**.

GT Label	Prediction
cone	**cone**
cube	**cube**
monkey	**monkey**
sphere	**sphere**
torus	**torus**
star	star shape
cylinder	cylinder
diamond	diamond shape
gear	gear
teapot	teapot

Table 16. **Class names associated with every visual cluster from Clevr-4 Count.** Known classes are marked in **bold**.

GT Label	Prediction
1	**23**
3	**3**
5	**5**
7	**7**
10	**10**
2	24
4	4
6	6
8	19
9	17

D. Additional results

D.1. Results on standard benchmarks

OAK also enhances GCD on standard single-context benchmarks by leveraging CLIP's semantic knowledge and context-aware attention. Tab. 17 shows full-shot results on CUB-200 and Stanford Cars using the CLIP ViT-B/16 backbone, demonstrating OAK's superiority in novel class discovery. Moreover, OAK is compatible with state-of-the-art GCD methods and can be further improved by integrating them.

Table 17. **Results on standard GCD benchmarks.**

	CUB-200			Stanford Cars		
	Old	New	All	Old	New	All
CLIP-ZS	**69.4**	-	-	**81.4**	-	-
CLIP-ZS + LLM vocab	46.4	44.0	44.8	54.6	47.4	49.7
CLIP-ZS + GT vocab	55.6	56.1	55.9	70.0	61.1	64.0
SS-KMeans	46.2	46.6	46.5	51.1	43.5	46.0
GCD	60.4	60.8	60.7	75.4	56.6	62.7
OAK (ours)	59.6	**62.4**	**61.5**	71.0	**63.4**	**65.9**

D.2. Results on abstract textures

We conduct experiments on the DTD [7] dataset, which contains images of abstract textures. Its 47 texture classes are split evenly into known and novel classes, using 20 labeled images per class. Tab. 18 shows that OAK outperforms the baselines, and Fig. 10 shows that OAK successfully discovers abstract classes such as *bubbly*.

Table 18. **Results on abstract textures.**

	Old	New	All
CLIP-ZS	53.3	-	-
CLIP-ZS + LLM vocab	34.0	43.7	40.4
GCD	55.4	61.7	59.6
OAK (ours)	**56.7**	**65.0**	**62.1**

Known: *bumpy* Novel: *bubbly*

Figure 10. **Example of known and novel classes in the DTD dataset.**

E. Additional analyses

E.1. Ablation study on Clevr-4

Following Tab. 4, we present ablation study on the method components on the Clevr-4 datasets in Tab. 19. Consistent with the proper observation, both context-aware attention and text-guided regularization enhance performance. While CLIP-ZS did not provide much benefit for synthetic images with abstract contexts, leveraging text semantics improved the overall accuracy of the baseline GCD, particularly for higher-level contexts like Texture and Count.

Table 19. **Ablation study on Clevr-4** shows consistent results as those on the Stanford datasets, as shown in Tab. 4.

Context-aware attention	Text-guided regularization	Known					Novel					Overall				
		Texture	Color	Shape	Count	Omni	Texture	Color	Shape	Count	Omni	Texture	Color	Shape	Count	Omni
-	-	73.4	98.3	99.0	41.9	35.5	43.6	94.9	99.2	42.3	15.7	58.2	96.6	99.1	42.1	22.6
✓	-	35.0	99.5	98.9	39.2	8.2	22.9	90.0	98.4	34.5	9.3	28.8	94.7	98.7	36.9	7.6
-	✓	74.8	98.1	99.4	**52.3**	**53.9**	46.0	90.2	99.4	34.4	10.9	60.1	94.1	99.4	43.3	27.9
✓	✓	**82.3**	**100.0**	**99.9**	45.0	40.5	**47.8**	**100.0**	**99.8**	**43.7**	**16.5**	**64.6**	**100.0**	**99.8**	**44.4**	**28.5**

E.2. Multi-seed results

We test the sensitivity of the 16 labeled images used for our final performance on the Stanford and CLEVR-4 datasets, applying five different random seeds for image selection in Tab. 20 and Tab. 21, respectively. OAK consistently outperforms the baselines with statistical significance, achieving substantial margins beyond the standard deviations.

Table 20. **Sensitivity analysis on the selection of 16 labeled images in the Stanford datasets.** We use five different random seeds for image selection, train GCD and OAK accordingly, and report the mean and standard deviation across the five runs.

Method	Known				Novel				Overall			
	Action	Location	Mood	Omni	Action	Location	Mood	Omni	Action	Location	Mood	Omni
SS-KMeans	63.0 ±4.2	62.8 ±7.1	25.9 ±0.6	12.5 ±0.0	57.8 ±3.5	67.9 ±4.7	**78.3** ±0.2	23.5 ±4.2	60.3 ±1.5	65.1 ±3.8	52.9 ±0.4	22.6 ±4.1
GCD	87.8 ±6.7	78.7 ±5.4	46.2 ±20.5	27.5 ±28.5	62.1 ±7.8	78.4 ±1.8	46.1 ±12.6	17.6 ±20.8	74.6 ±6.9	78.6 ±2.9	46.1 ±6.5	45.3 ±10.1
OAK (ours)	**89.8** ±0.4	**84.2** ±1.5	**59.6** ±12.3	5.0 ±6.8	**79.0** ±0.3	**80.3** ±1.5	77.4 ±12.7	**37.6** ±14.2	**84.2** ±1.7	**82.4** ±1.1	**68.6** ±11.0	**49.7** ±14.9

Table 21. **Sensitivity analysis on the selection of 16 labeled images in the Clevr-4 datasets**, following the same settings in Tab. 20.

Method	Known					Novel					Overall				
	Texture	Color	Shape	Count	Omni	Texture	Color	Shape	Count	Omni	Texture	Color	Shape	Count	Omni
SS-KMeans	13.0 ±0.1	11.3 ±0.8	79.3 ±8.2	24.2 ±0.4	0.2 ±0.1	13.6 ±0.3	12.1 ±0.8	78.7 ±6.3	15.3 ±0.5	0.2 ±0.3	13.3 ±0.1	11.7 ±0.0	79.0 ±1.9	19.7 ±0.1	0.1 ±0.02
GCD	47.4 ±27.4	76.3 ±25.2	98.0 ±3.3	43.0 ±8.2	32.0 ±11.4	37.1 ±9.2	64.9 ±32.9	99.1 ±1.1	33.5 ±6.0	10.0 ±4.8	42.1 ±17.9	70.5 ±27.4	98.5 ±1.7	38.2 ±6.5	18.4 ±6.7
OAK (ours)	**78.8** ±4.0	**99.5** ±1.0	**100.0** ±0.0	**45.0** ±3.7	**44.5** ±4.1	**47.0** ±2.4	**99.8** ±0.3	**99.8** ±0.03	**39.2** ±1.6	**14.5** ±1.9	**62.6** ±2.5	**99.6** ±0.5	**99.9** ±0.03	**42.1** ±1.2	**26.7** ±1.9

E.3. Class names from large datasets

Ad-hoc category discovery is an open-ended problem covering diverse custom contexts, making LLMs a natural choice since large datasets for these contexts are generally unavailable. Nevertheless, we compare our class names with those from the Kinetics [4] dataset, which contains 700 action classes. Tab. 22 shows that both produce similar novel class names when the candidate set is sufficiently large, such as *sweeping floor* vs. *cleaning the floor*.

Table 22. **Comparison of predicted class names using candidate sets generated by GPT and those retrieved from Kinetics-700.**

GT Label	From ChatGPT-4o	From Kinetics-700
blowing bubbles	blowing bubbles	blowing bubble gum
cleaning the floor	mopping the floor	sweeping floor
cooking	preparing a meal	cooking egg
cutting vegetables	climbing	cutting apple
feeding a horse	petting a horse	petting horse
fixing a bike	fixing a bike	fixing bicycle
gardening	weeding a garden	digging
jumping	dancing	high jump
looking through a telescope	looking through a microscope	using a microscope
playing guitar	strumming a guitar	playing guitar
pouring liquid	carrying a box	pouring milk
reading	reading a book	reading book
riding a horse	running	riding or walking with horse
running	jogging	jogging
smoking	smoking	smoking
texting message	shaking hands	texting
using a computer	texting	assembling computer
washing dishes	washing dishes	washing dishes
waving hands	clapping	waving hand
writing on a book	writing a letter	reading book

E.4. Additional analysis on Count

We plot the mean error of OAK and CLIP against the number of objects in an image from the Clevr-4 dataset. For CLIP, we use the true names of novel classes, while OAK predicts them by matching cluster embeddings. Fig. 11 shows that CLIP struggles as object count increases, whereas OAK maintains stable performance. This highlights OAK 's ability to infer object counts through visual clustering, which is difficult to learn purely from semantics. Nonetheless, specialized object-counting models may still be needed for higher object counts (>10) beyond those in Clevr-4.

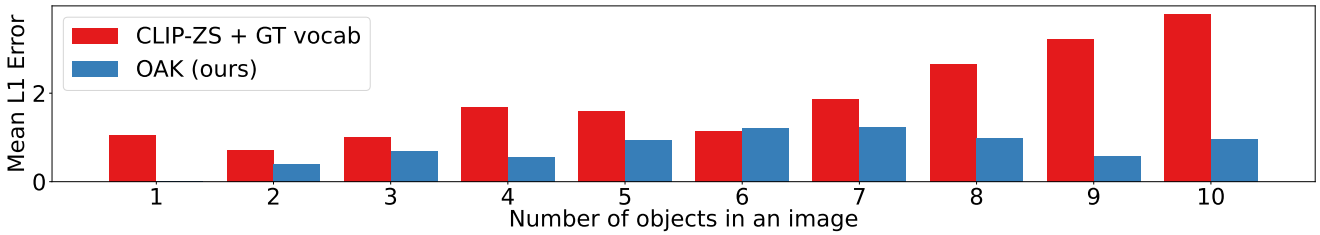
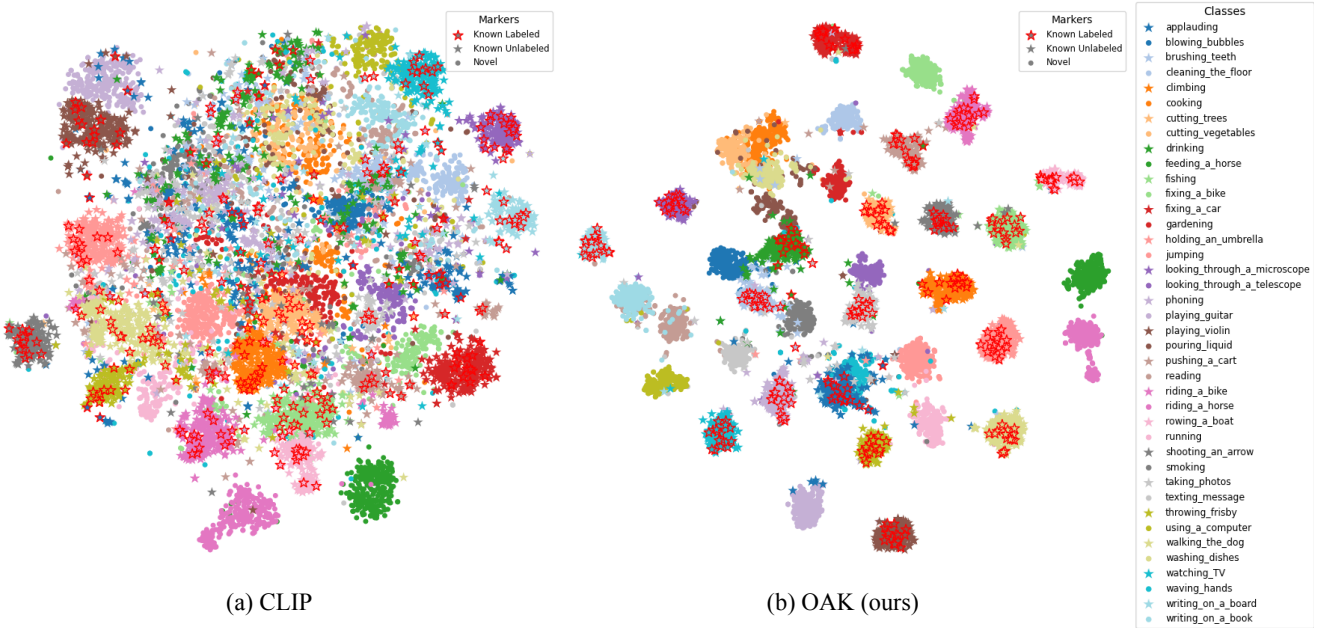


Figure 11. **Mean error of OAK and CLIP versus the number of objects in an image.**

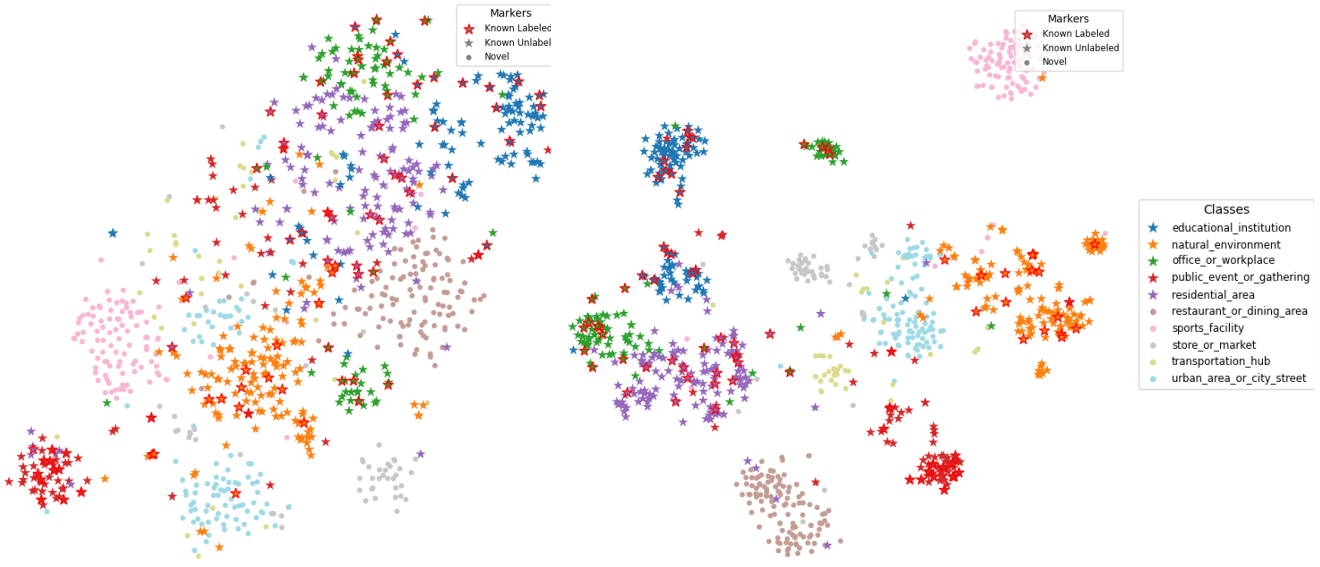
E.5. t-SNE visualizations

We present t-SNE plots of the feature spaces of CLIP and OAK on Stanford Action, Stanford Location, Stanford Mood, Clevr-4 Texture, Clevr-4 Color, Clevr-4 Shape, and Clevr-4 Count in the following figures. The results show that OAK refines CLIP features into well-clustered representations aligned with each context. Notably, OAK performs well in contexts CLIP does not inherently capture, such as Stanford Location. For out-of-distribution (OOD) images like Clevr-4 Shape and Clevr-4 Color, OAK achieves near-perfect clustering. Even in cases that are both OOD and outside CLIP's primary focus, such as Clevr-4 Texture and Clevr-4 Count, OAK forms reasonably coherent clusters, demonstrating its effectiveness.



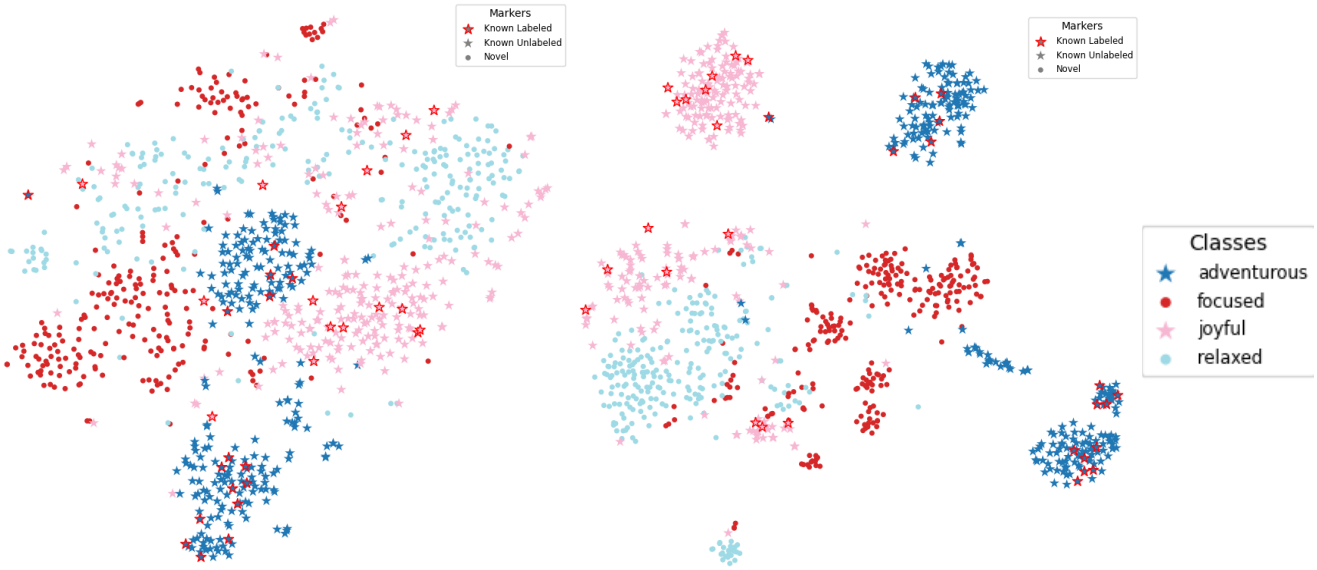
(a) CLIP (b) OAK (ours)

Figure 12. **t-SNE plot of CLIP and OAK's feature space on Stanford Action.**



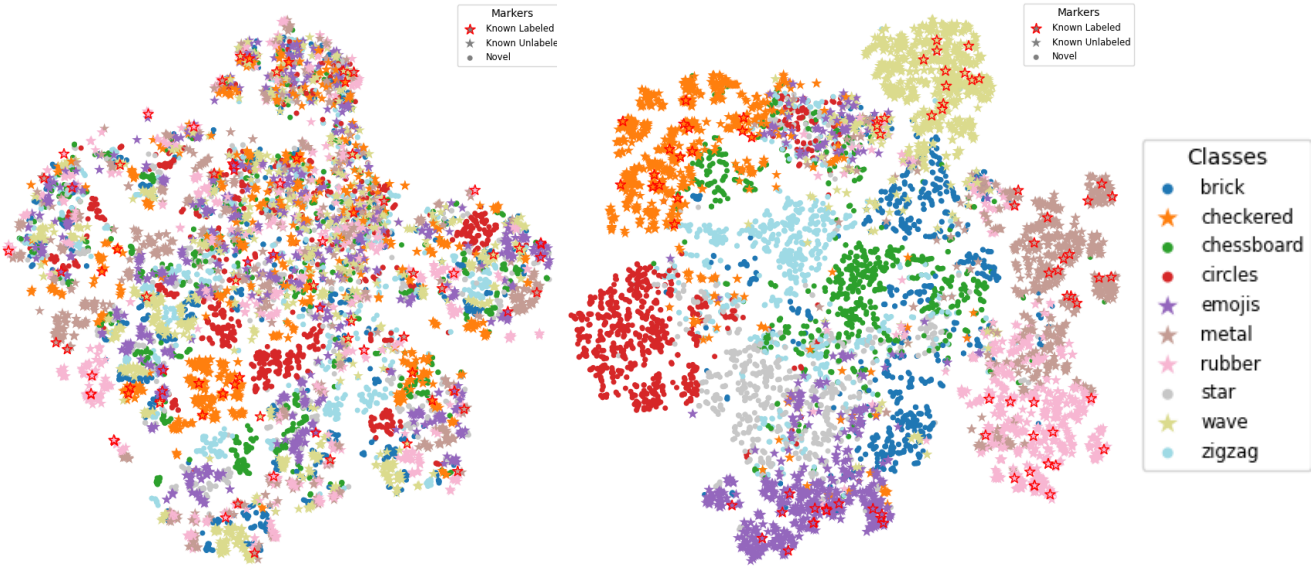
(a) CLIP (b) OAK (ours)

Figure 13. **t-SNE plot of CLIP and OAK's feature space on Stanford Location.**



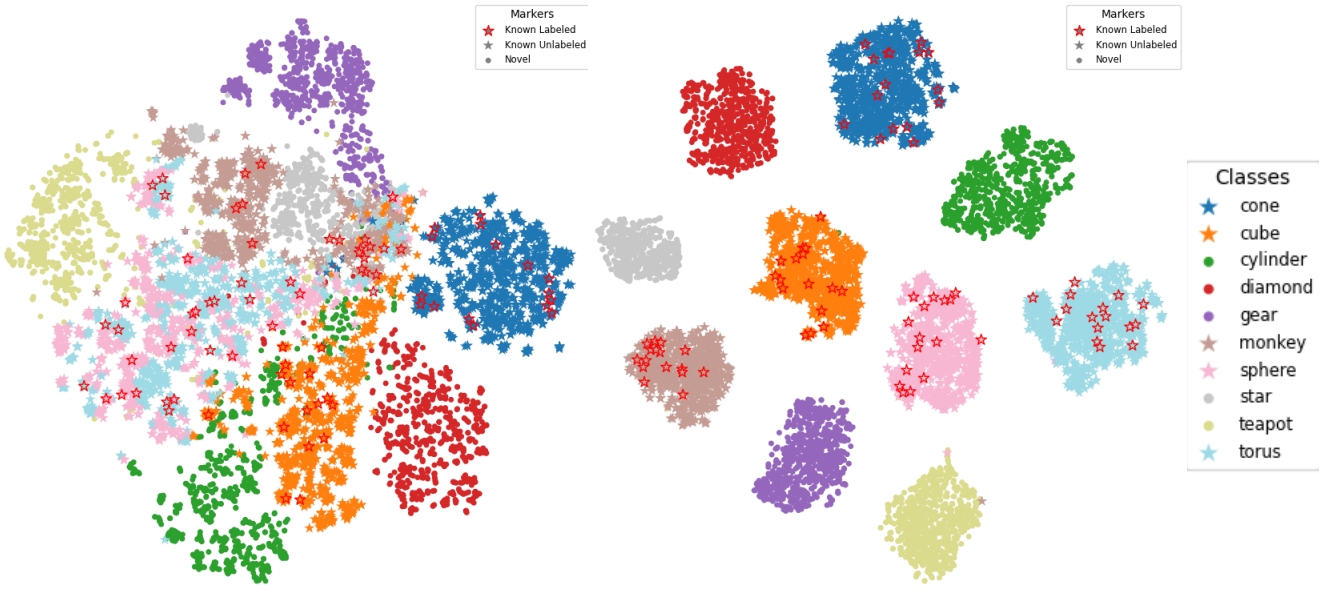
(a) CLIP (b) OAK (ours)

Figure 14. **t-SNE plot of CLIP and OAK's feature space on Stanford Mood.**



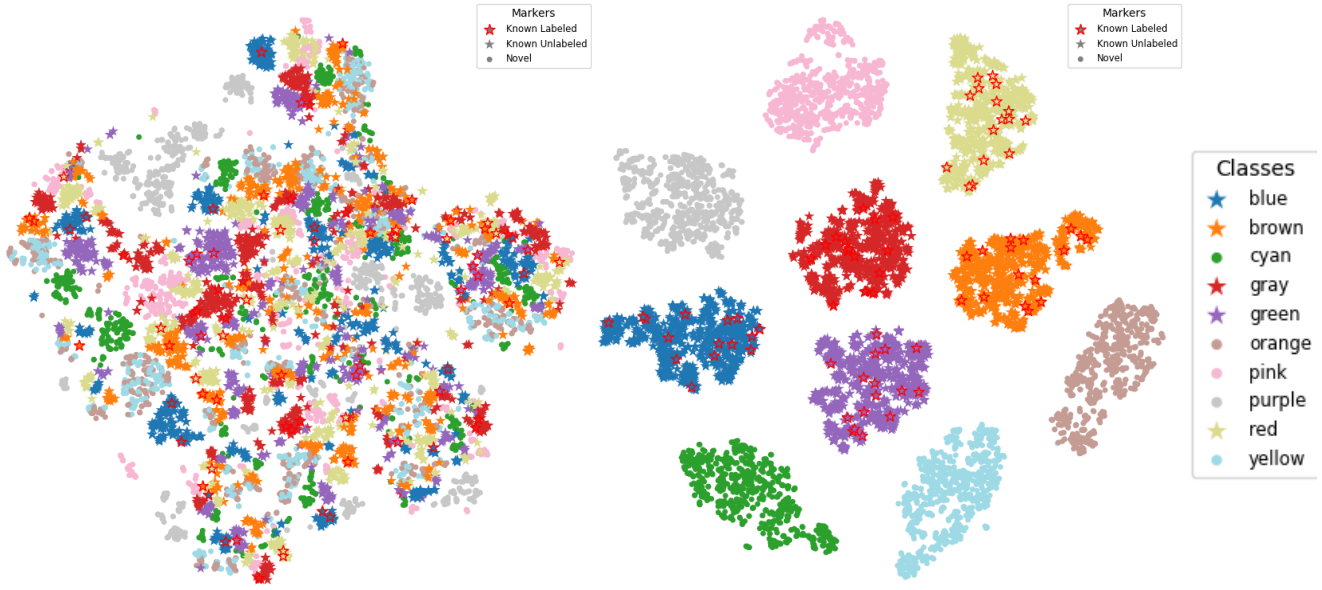
(a) CLIP (b) OAK (ours)

Figure 15. **t-SNE plot of CLIP and OAK's feature space on CLEVR4 Texture.**



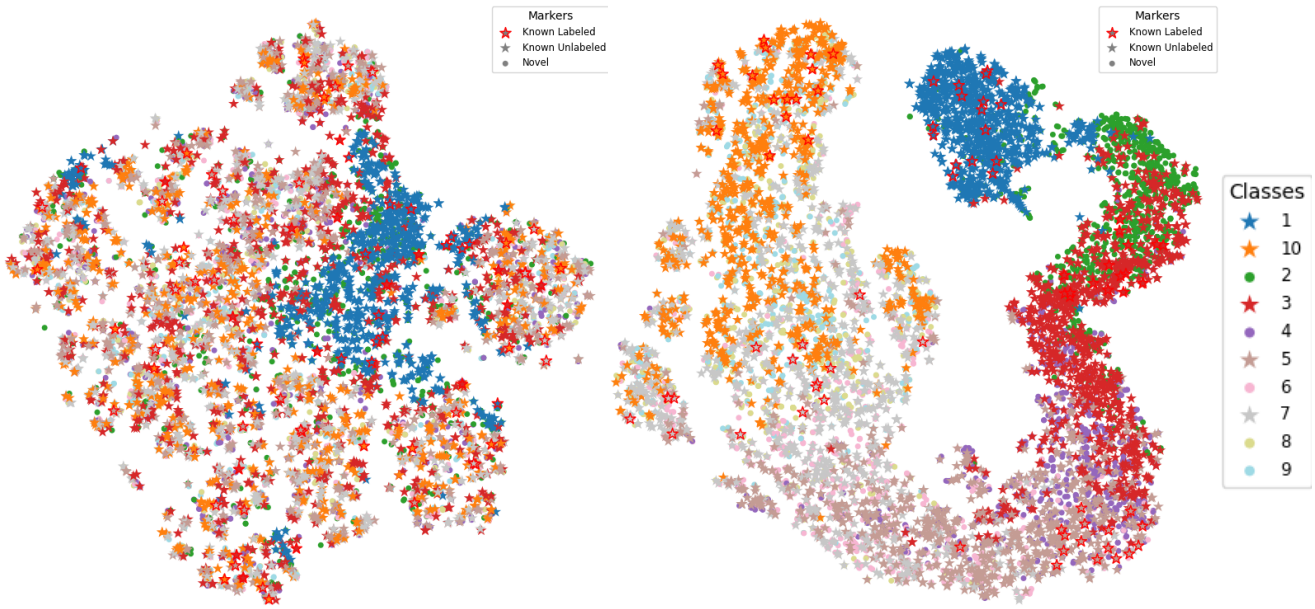
(a) CLIP (b) OAK (ours)

Figure 16. **t-SNE plot of CLIP and OAK's feature space on CLEVR4 Shape.**



(a) CLIP (b) OAK (ours)

Figure 17. **t-SNE plot of CLIP and OAK's feature space on CLEVR4 Color.**



(a) CLIP (b) OAK (ours)

Figure 18. **t-SNE plot of CLIP and OAK's feature space on CLEVR4 Count.**