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

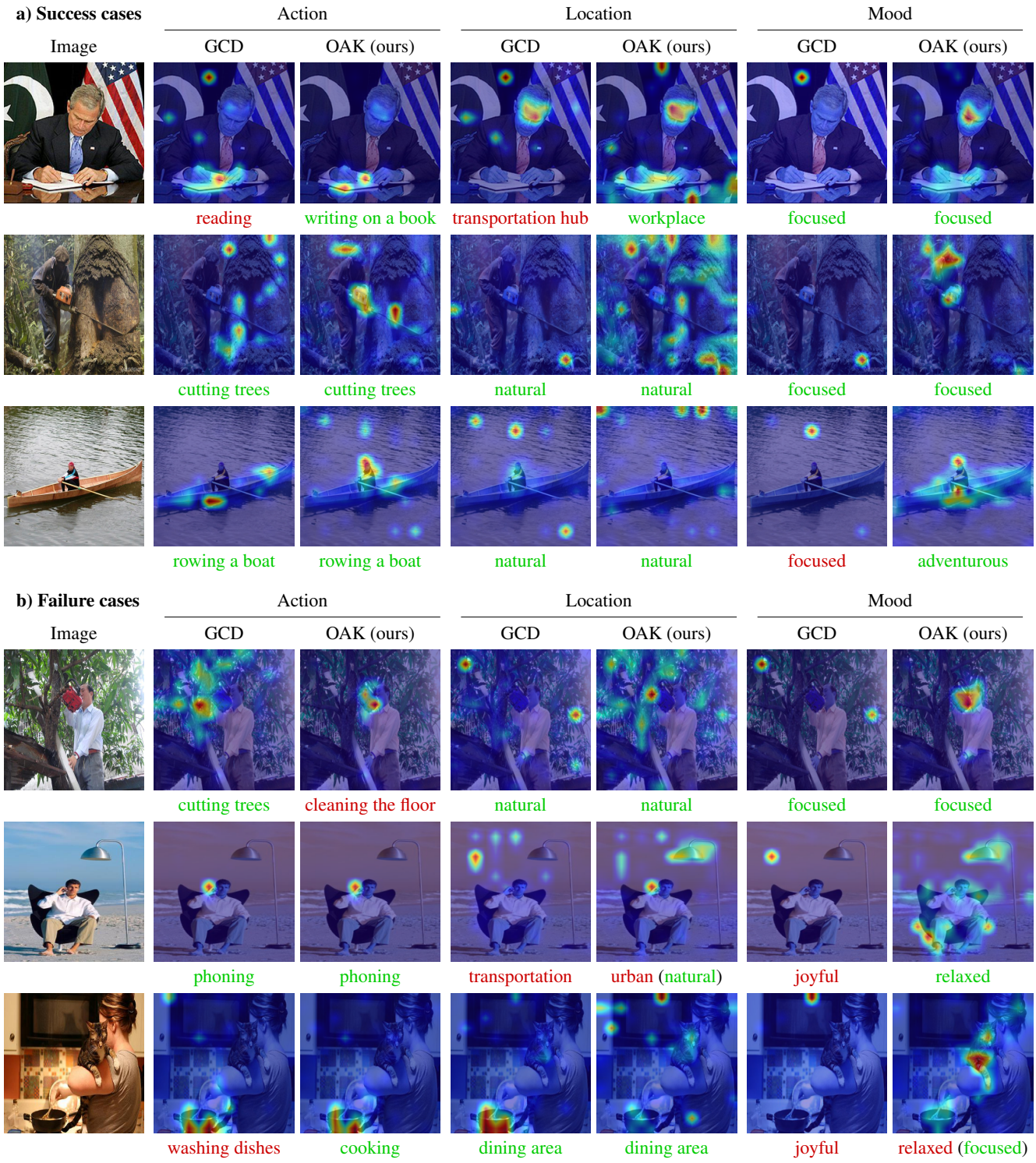


Figure 8. **Additional saliency maps on the Stanford dataset** demonstrate that OAK makes reasonable predictions, focusing on the relevant regions for different contexts. We select three samples correctly predicted by OAK across all contexts and three that fail. In the failure cases, OAK 1) ignores the *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.

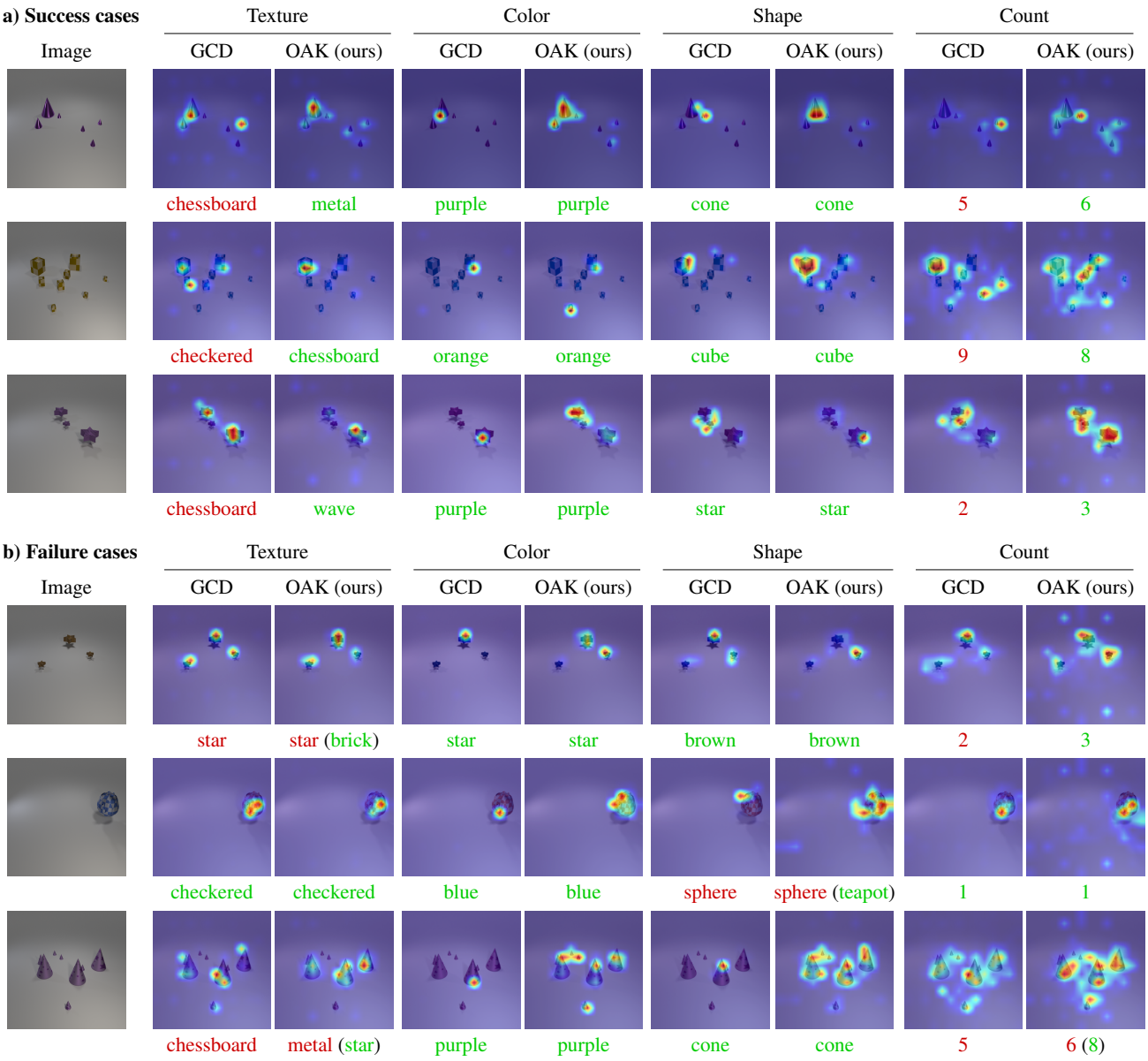


Figure 9. **Additional saliency maps on the Clevr-4 dataset**, showing that OAK makes sensible predictions by focusing on relevant regions across various contexts, using the same setup. In the failure cases, OAK 1) struggles to identify black *brick* patterns on a dark *brown* object, mistaking the *star* 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.

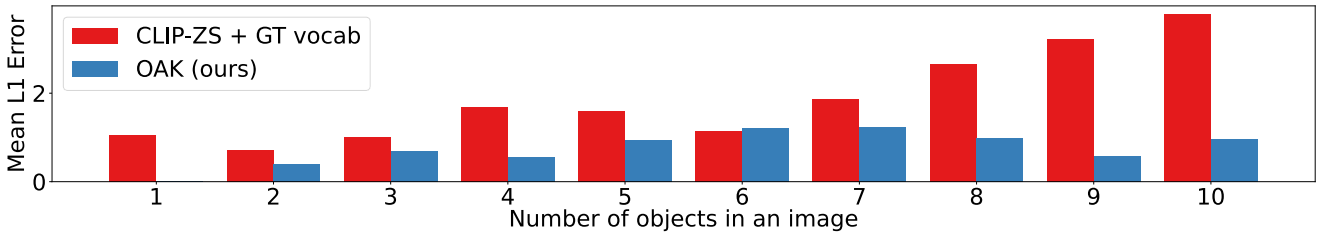
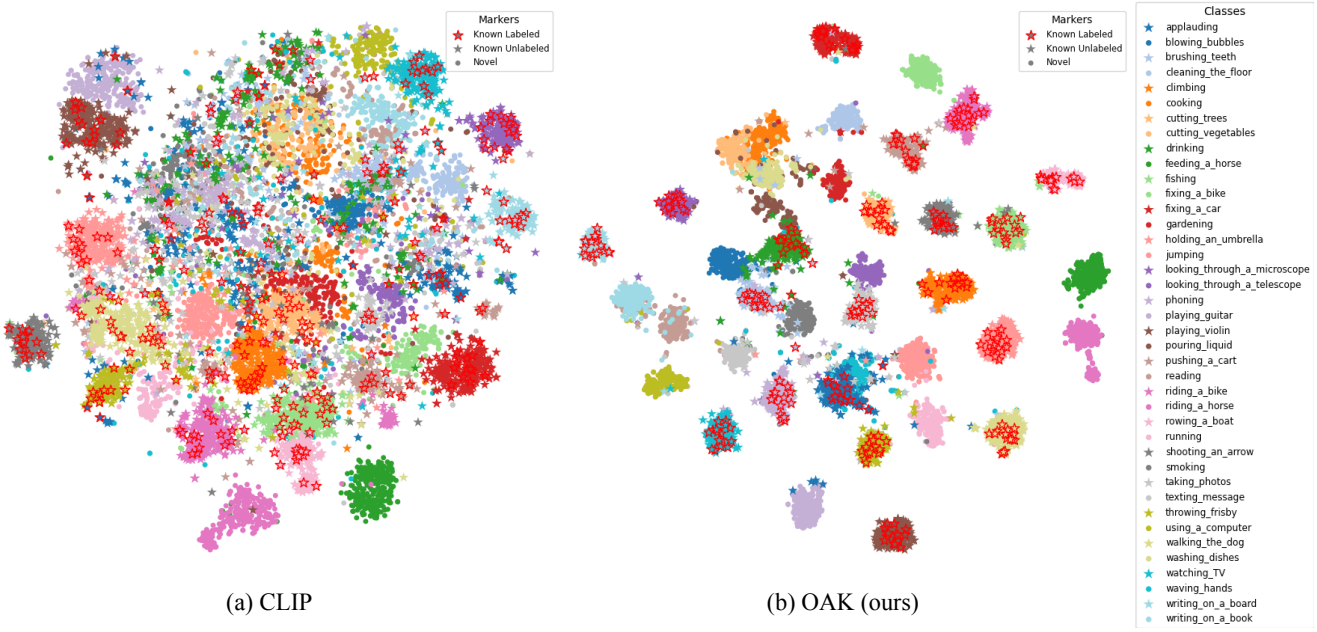


Figure 11. **Mean error of OAK and CLIP versus the number of objects in an image.**

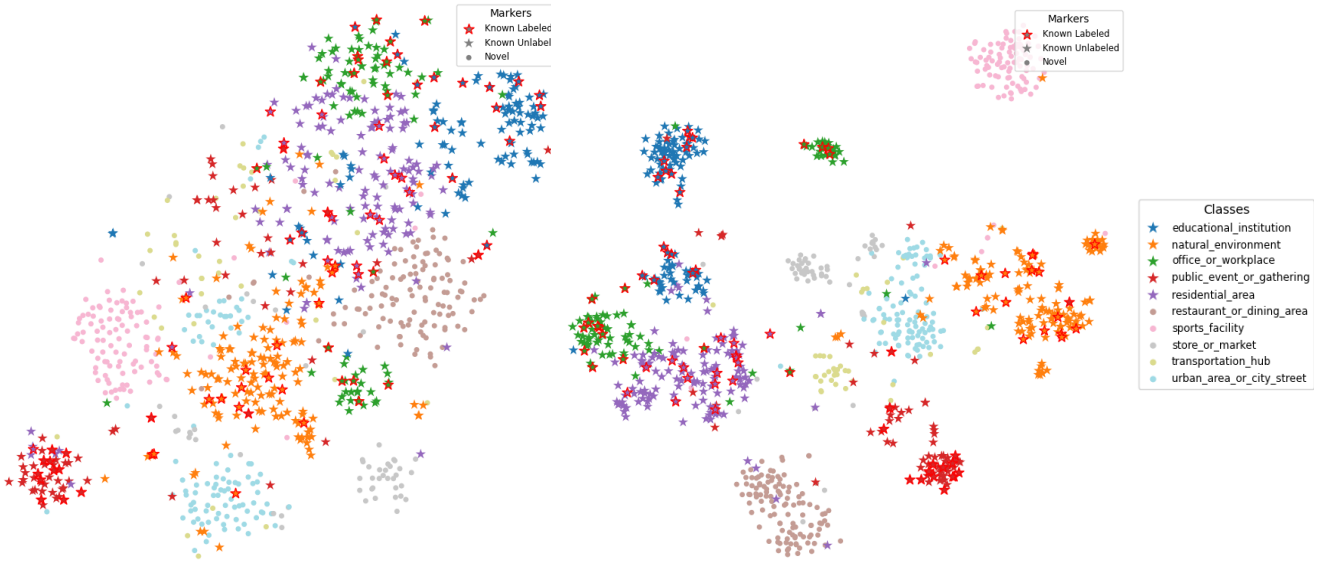
E.5. t-SNE visualizations

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.



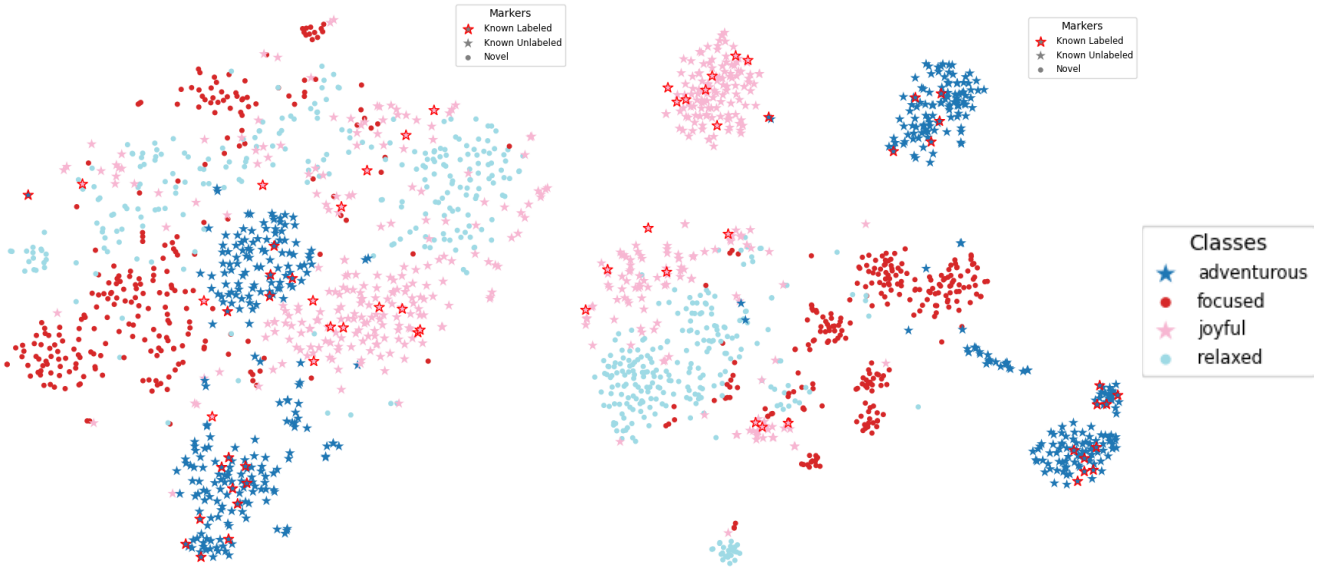
(a) CLIP (b) OAK (ours)

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



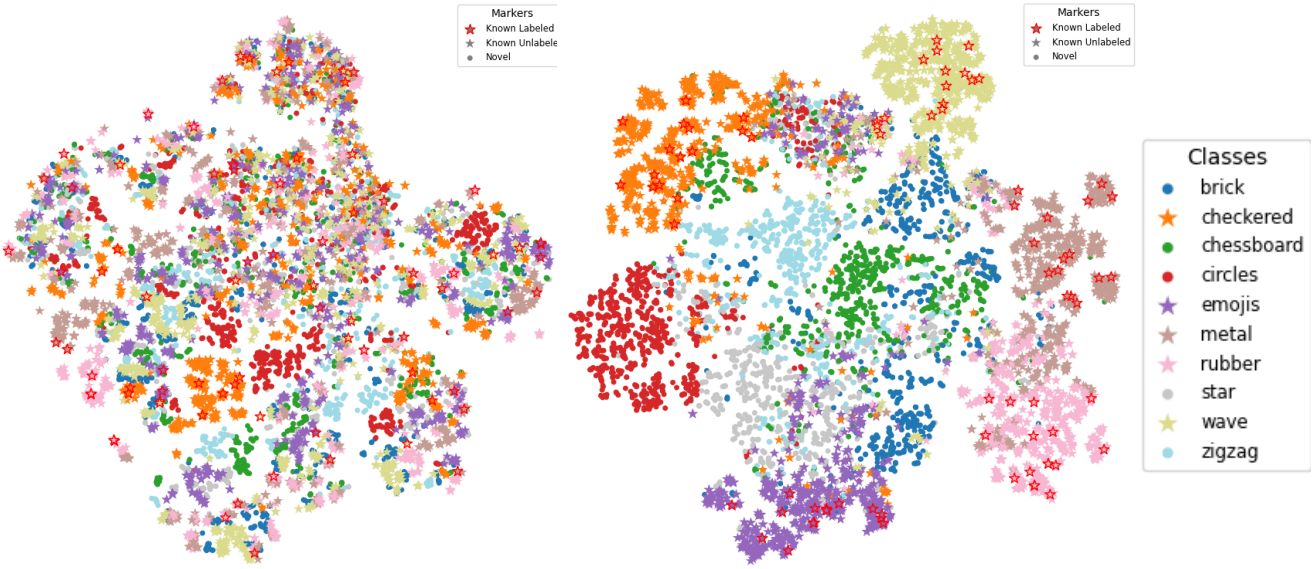
(a) CLIP (b) OAK (ours)

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



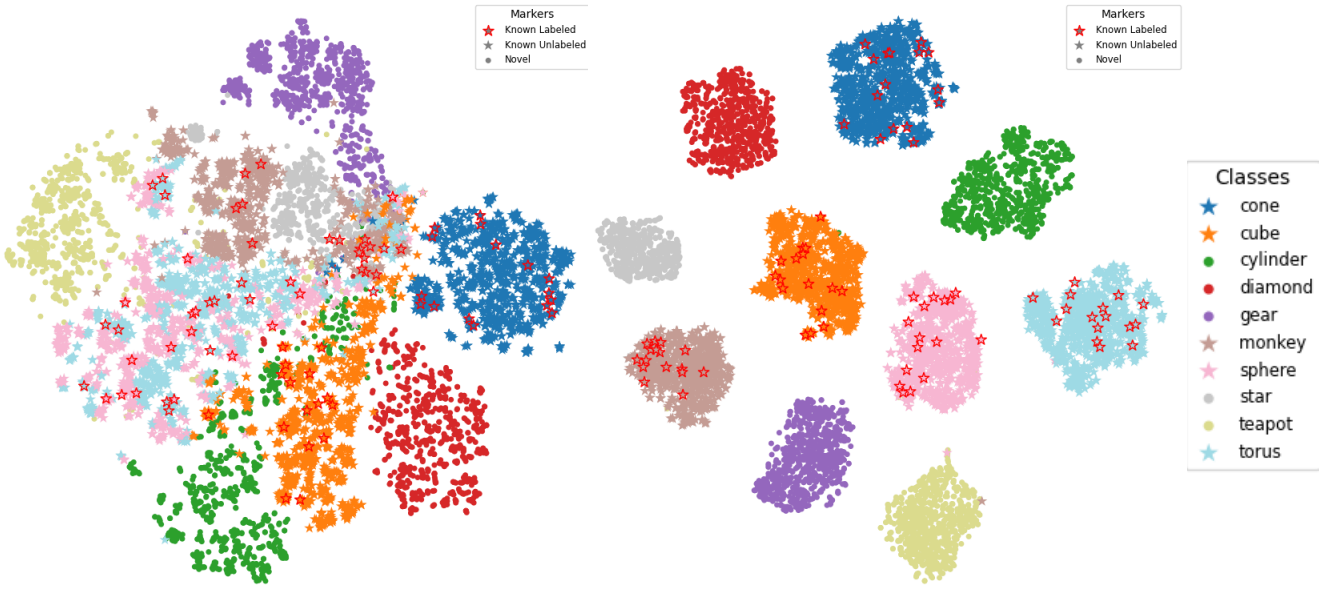
(a) CLIP (b) OAK (ours)

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



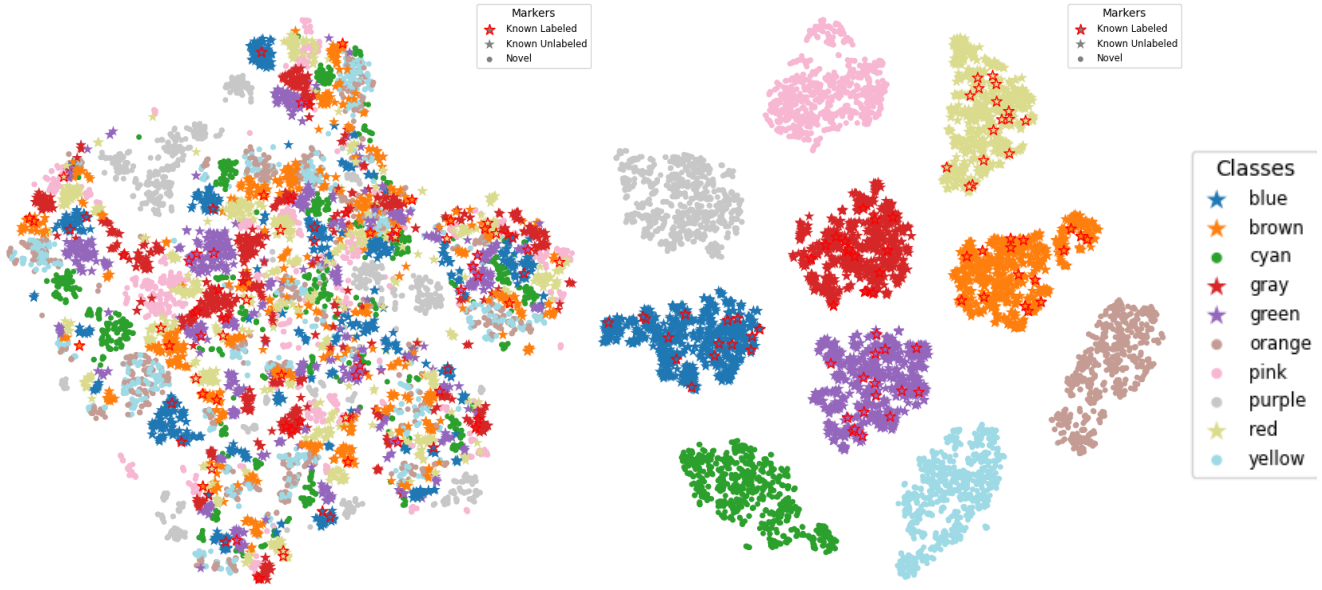
(a) CLIP (b) OAK (ours)

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



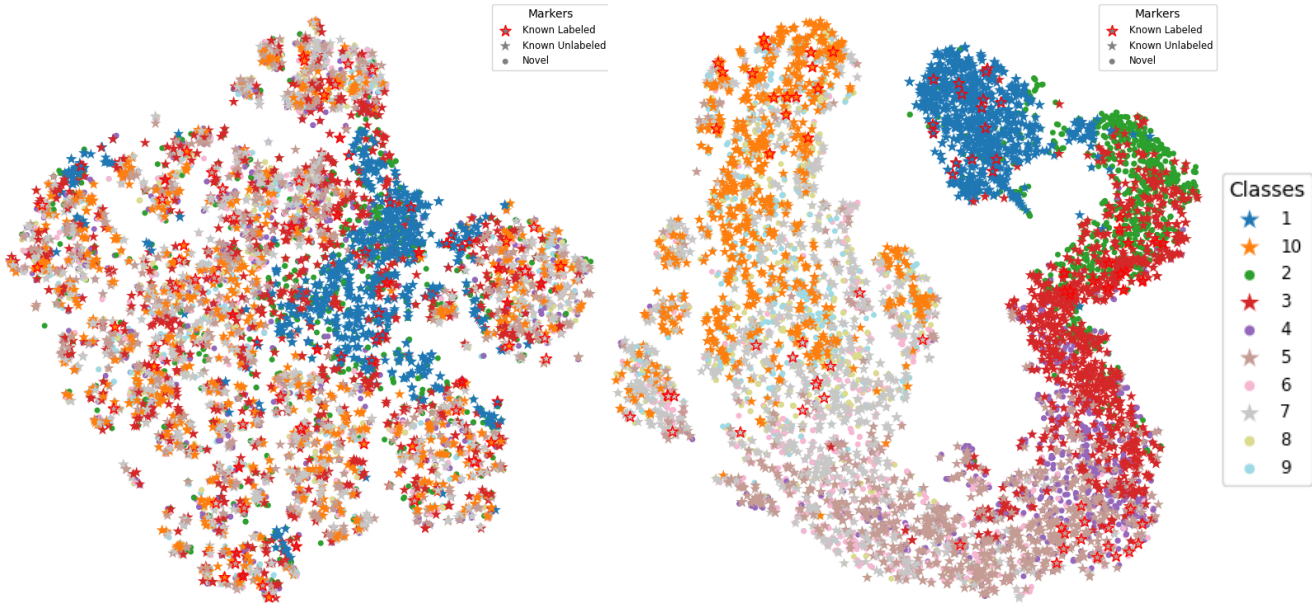
(a) CLIP (b) OAK (ours)

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.**