Supplementary: Distilling Large Vision-Language Model with Out-of-Distribution Generalizability

	$\mathcal{X}_{\text{ood1}}$	$\mathcal{X}_{\text{ood2}}$
0-shot	25.11	18.01
5-shot on $\mathcal{X}_{\text{ood1}}$	**62.79**	**18.24**

Table 1: Comparison between the student model's zero-shot generalization performance on $\mathcal{X}_{\text{ood2}}$ before and after few-shot finetuning on $\mathcal{X}_{\text{ood1}}$.

1. More Experiments

Zero-shot OOD generalization ability after few-shot learning. In the real world, it is essential for student networks to continuously adapt to new concepts, and we aim to find a student learning strategy that accomplishes such goal. Since we finetune student visual backbones during few-shot learning on $\mathcal{X}_{\text{ood}}$, we would like to know whether finetuned student backbones overfit seen concepts and exhibit weaker zero-shot generalization ability when encountering novel unseen concepts again. We conduct an experiment on Flower102, where we split $\mathcal{Y}_{\text{ood}}$ into two equal sets, and then split $\mathcal{X}_{\text{ood}}$ into $\mathcal{X}_{\text{ood1}}$ and $\mathcal{X}_{\text{ood2}}$ accordingly. We then select the best student model and evaluate it on $\mathcal{X}_{\text{ood2}}$ both before and after few-shot finetuning it on $\mathcal{X}_{\text{ood1}}$. Results are presented in Tab. 1. We observe that the student's zero-shot OOD generalization ability slightly improves after few-shot learning, demonstrating that students can continuously adapt to novel concepts.

2. Application

In this section, we demonstrate that we can adopt our previous findings for improving student's OOD generalization ability towards novel tasks and domains. We augment the PickClutter task from a robot object manipulation skill benchmark ManiSkill2 [4] with language, where given the name of a YCB object [2], a robot needs to detect whether it exists among a pile of objects given the current visual observation captured from a hand camera, and if exists, picks up this object. The task is illustrated in Fig. 1. We randomly sample different configurations of objects, and given observations in each configuration, the student network outputs whether it is feasible to grasp each YCB object.

	$\mathcal{X}_{\text{id}}$	$\mathcal{Y}_{\text{id}}$ on $\mathcal{X}_{\text{ood}}$	$\mathcal{Y}_{\text{ood}}$ on $\mathcal{X}_{\text{ood}}$
Closed-Set	96.4 / 96.5	NA / 86.2	NA / 87.7
$\mathcal{L}_{\text{cls}}$	96.9 / 97.2	79.3 / 85.3	71.7 / 87.3
+ $\mathcal{L}_{\text{im-cst}}$	**99.2 / 99.2**	84.0 / 91.9	76.3 / 88.3
+ Semantic Enrich	98.2 / 99.0	**84.3 / 92.0**	**83.0 / 89.6**

(a) Overall accuracy over all YCB objects

	$\mathcal{X}_{\text{id}}$	$\mathcal{Y}_{\text{id}}$ on $\mathcal{X}_{\text{ood}}$	$\mathcal{Y}_{\text{ood}}$ on $\mathcal{X}_{\text{ood}}$
Closed-Set	91.6 / 91.9	NA / 57.8	NA / 35.9
$\mathcal{L}_{\text{cls}}$	94.0 / 94.4	46.5 / 54.7	23.3 / 32.8
+ $\mathcal{L}_{\text{im-cst}}$	**98.1 / 98.0**	55.3 / **70.8**	**23.7** / 47.3
+ Semantic Enrich	97.2 / 97.4	**55.6 / 70.8**	11.7 / **50.7**

(b) F1-measure over objects that exist in observations.

Table 2: Results on grasp feasibility prediction in the PickClutter task. In each entry, x_1/x_2 denote zero-shot and few-shot evaluation results, respectively. ($\mathcal{L}_{\text{vlprox}}$ is not available for this experiment since multiple objects exists in an observation)

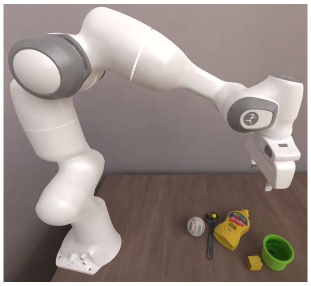

Figure 1: Illustration of the PickClutter task. Given the name of a target object, the robot agent needs to decide whether the corresponding object is graspable based on the current visual observation.

This setup resembles observation-based affordance prediction in works such as SayCan [1]. We select 26 visually-distinctive YCB objects and split them equally into $\mathcal{Y}_{\text{id}}$ and $\mathcal{Y}_{\text{ood}}$. We then construct the datasets such that $\mathcal{X}_{\text{train}}$ and $\mathcal{X}_{\text{id}}$ only contain objects in $\mathcal{Y}_{\text{id}}$, while $\mathcal{X}_{\text{ood}}$ contains objects in both $\mathcal{Y}_{\text{id}}$ and $\mathcal{Y}_{\text{ood}}$. We use 3000 scenes for training and 50 scenes for few shot learning. We adopt EfficientNet [6] as the student network and CLIP ViT-L-14 as the teacher. As we are in a multi-label classification setting, our preliminary experiments show that having positive and negative prompts like [5] and learning these prompts during student training can significantly improve student performance, so we adopt these techniques in our experiments. We calculate two metrics: (1) overall accuracy across all YCB objects in the label set (i.e., $\mathcal{Y}_{\text{id}}$ for $\mathcal{X}_{\text{id}}$, and $\mathcal{Y}_{\text{id}} \cup \mathcal{Y}_{\text{ood}}$ for $\mathcal{X}_{\text{ood}}$); (2) F1-measure calculated over the objects present in an observation, averaged over all observations. Please refer to Appendix Sec. 8 for more implementation and metric details.

We present the results in Tab. 2. Consistent with our findings in the main paper, we observe that improving teacher-student visual space alignments through $\mathcal{L}_{\text{im-cst}}$ significantly enhances the student's generalization ability on OOD objects. Additionally, leveraging language models to enrich the semantic details of object descriptions benefits few-shot OOD generalization. Interestingly, for zero-shot OOD generalization, while this enrichment improves the overall prediction accuracy for novel objects, it adversely affects the recall (and thus the F1-measure) on these objects. Further analysis reveals that students tend to ignore objects unseen during training, resulting in lower recall. However, with just a few examples of novel objects, students achieve significantly better recall on these objects.

3. Dataset Statistics

	CaltechBirds	StanfordCars	Flower102	Food101	SUN397	tiered-ImageNet
$\lvert\mathcal{X}_{\text{train}}\rvert$	4122	2874	3112	35700	38663	314108
$\lvert\mathcal{X}_{\text{id}}\rvert$	1740	1164	1303	15300	16444	134587
$\lvert\mathcal{X}_{\text{ood}}\rvert$	5926	4106	3774	50000	53647	124261
$\lvert\mathcal{Y}_{\text{id}}\rvert$	100	98	51	51	200	351
$\lvert\mathcal{Y}_{\text{ood}}\rvert$	100	98	51	50	197	97

Table 3: Dataset Statistics for our main experiments.

In Tab. 3, we provide dataset split statistics for the experiments in the main paper.

4. Training Details and Hyperparameters for Main Experiments

During agent training on $\mathcal{X}_{\text{train}}$, for CaltechBirds, Stanford Cars, and Flower102, due to their relatively small dataset size, we train agents for 450 epochs to ensure convergence; for Food101, SUN397, and tiered-ImageNet, we train agents for 90 epochs. For ResNet-based students, we adopt an initial learning rate of 0.05 with batch size 128, which is decreased to 0.005 after $1/3$ of training epochs and to 0.0005 after $2/3$ of training epochs. We adopt standard data augmentation (Random cropping image to 224x224, and random horizontal flip). For ViT-based students, we adopt a one-cycle learning rate schedule with a learning-rate peak of 0.0002. We also apply RandAugment [3], a strong data augmentation method, to cope with overfitting. However, even with RandAugment, we still observe that the performance on $\mathcal{X}_{\text{ood}}$ starts to decrease at some point in training, suggesting that it is challenging to train ViT students on small to medium-scale datasets to obtain good out-of-distribution generalizability.

During few-shot learning on $\mathcal{X}_{\text{ood}}$, we adopt a balanced training batch, where at most half of samples come from few-shot samples on $\mathcal{X}_{\text{ood}}$ and the rest of samples come from $\mathcal{X}_{\text{train}}$. For CaltechBirds, Stanford Cars, and Flower102, we train agents for 100 epochs. For Food101, SUN397, and tiered-ImageNet, we train agents for 20 epochs. We adopt a one-cycle learning rate schedule for all student networks. For ResNet-based students, the learning-rate peak is 0.003. For ViT-based students, the learning-rate peak is 0.0001. We adopt the same data augment strategy as before.

All experiment results are obtained from the average performance of the last 5 epochs of training.

5. More Metric Evaluations

In this section, we evaluate $\mathcal{M}_{\text{rel}}$, $\mathcal{M}_{\text{neigh}}$, and $\mathcal{M}_{\text{vlalign}}$ on more datasets to complement our results on teacher-student visual space and vision-language alignment. $\mathcal{M}_{\text{rel}}$ results are presented in Tab. 4. $\mathcal{M}_{\text{neigh}}$ results are presented in Tab. 5. $\mathcal{M}_{\text{vlalign}}$ results are presented in Tab. 6.

6. Example ChatGPT Label Description Generations from Different Prompts

In the main paper, we investigated leveraging different prompts to control the level of semantic details ChatGPT generates for the description of each label, and how these different generated label descriptions impact student performance. In this section, we provide a list of example label descriptions generated using various prompts.

- **Original prompt**: "Use a single sentence to describe the appearance and shape of {cls}. Only describe the shape and appearance."
  - The Egyptian Cat, or Egyptian Mau, is a medium-sized feline with a lithe and muscular body, a short, spotted coat in colors such as silver or bronze, and large, almond-shaped green eyes.
  - The space bar is a rectangular key on a computer keyboard that is longer horizontally than other keys and has a slightly concave shape for the thumb.
  - Hair spray typically comes in a cylindrical container with a nozzle on top for spraying, and may vary in size and color depending on the brand and type.

$\mathcal{M}_{\text{rel}} \uparrow$		CaltechBirds	StanfordCars	Flower102	Food101	SUN397	tiered-ImageNet
$\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{mse}}$	$\mathcal{X}_{\text{train}}$	0.035	0.059	0.030	0.005	0.009	0.013
	$\mathcal{X}_{\text{ood}}$	0.014	0.017	0.004	0.002	0.003	0.007
$\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{mse}} + \mathcal{L}_{\text{im-cst}}$	$\mathcal{X}_{\text{train}}$	0.592	0.669	0.305	0.108	0.062	0.088
	$\mathcal{X}_{\text{ood}}$	0.081	0.106	0.022	0.028	0.019	0.041

Table 4: We evaluate $\mathcal{M}_{\text{rel}}$ (higher the better) on different datasets to measure how students preserve the relative feature relationships of the teacher's visual representation space.

$\mathcal{M}_{\text{neigh}} \uparrow$		CaltechBirds	StanfordCars	Flower102	Food101	SUN397	tiered-ImageNet
$\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{mse}}$	$k = 3$	0.11 / 0.03	0.15 / 0.04	0.13 / 0.06	0.01 / 0.00	0.03 / 0.01	0.03 / 0.01
	$k = 5$	0.15 / 0.04	0.21 / 0.05	0.18 / 0.07	0.02 / 0.01	0.04 / 0.01	0.05 / 0.01
	$k = 10$	0.24 / 0.06	0.30 / 0.07	0.27 / 0.08	0.03 / 0.01	0.06 / 0.02	0.07 / 0.02
$\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{mse}} + \mathcal{L}_{\text{im-cst}}$	$k = 3$	0.22 / 0.05	0.32 / 0.08	0.20 / 0.10	0.04 / 0.01	0.06 / 0.02	0.07 / 0.02
	$k = 5$	0.28 / 0.06	0.36 / 0.09	0.25 / 0.11	0.05 / 0.01	0.07 / 0.02	0.09 / 0.03
	$k = 10$	0.35 / 0.09	0.43 / 0.11	0.34 / 0.13	0.08 / 0.02	0.11 / 0.03	0.13 / 0.03

Table 5: We evaluate $\mathcal{M}_{\text{neigh}}$ (higher the better) on different datasets to measure how students preserve the local structure of the teacher's visual representation space. x_1/x_2 in each entry denote $\mathcal{M}_{\text{neigh}}(\mathcal{X}_{\text{train}})$ and $\mathcal{M}_{\text{neigh}}(\mathcal{X}_{\text{ood}})$, respectively.

$\mathcal{M}_{\text{vlalign}} \downarrow$		CaltechBirds	StanfordCars	Flower102	Food101	SUN397	tiered-ImageNet
$\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{mse}}$	$k = 2$	0.20 / 0.39	0.20 / 0.37	0.20 / 0.50	0.09 / 0.38	0.21 / 0.38	0.21 / 0.42
	$k = 3$	0.72 / 1.17	0.57 / 1.15	0.68 / 1.45	0.51 / 1.16	0.72 / 1.17	0.75 / 1.30
	$k = 5$	2.47 / 3.91	2.06 / 3.70	2.67 / 4.73	2.33 / 3.98	2.78 / 3.96	2.81 / 4.35
$\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{mse}} + \mathcal{L}_{\text{im-cst}}$	$k = 2$	0.21 / 0.37	0.19 / 0.33	0.18 / 0.43	0.10 / 0.29	0.20 / 0.35	0.21 / 0.39
	$k = 3$	0.71 / 1.15	0.54 / 1.07	0.62 / 1.30	0.49 / 0.93	0.67 / 1.08	0.72 / 1.22
	$k = 5$	2.30 / 3.74	1.89 / 3.53	2.52 / 4.24	2.19 / 3.42	2.55 / 3.68	2.72 / 4.11
$\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{mse}} + \mathcal{L}_{\text{im-cst}} + \mathcal{L}_{\text{vlprox}}$	$k = 2$	0.19 / 0.35	0.17 / 0.37	0.17 / 0.39	0.09 / 0.29	0.19 / 0.34	0.19 / 0.36
	$k = 3$	0.66 / 1.06	0.46 / 1.11	0.59 / 1.17	0.45 / 0.93	0.65 / 1.06	0.67 / 1.12
	$k = 5$	2.18 / 3.52	1.70 / 3.61	2.17 / 4.20	2.06 / 3.35	2.44 / 3.60	2.54 / 3.82

Table 6: We evaluate $\mathcal{M}_{\text{vlalign}}$ (lower the better) on different datasets to measure the proximity between student and teacher vision-language alignment structures. x_1/x_2 in each entry denote $\mathcal{M}_{\text{vlalign}}(\mathcal{X}_{\text{train}})$ and $\mathcal{M}_{\text{vlalign}}(\mathcal{X}_{\text{ood}})$, respectively.

- **"More Succinct" prompt**: "Use a single sentence to broadly describe the appearance and shape of {cls}. Don't give too much details. Only describe the shape and appearance."
  - The Egyptian cat, also known as the Egyptian Mau, has a slender, athletic body with a spotted coat and large, expressive eyes.
  - The space bar is a rectangular-shaped key located at the bottom-center of a standard computer keyboard.
  - Hair spray is typically contained in an aerosol can with a long, narrow nozzle used for spraying the product onto hair.
- **"More Detailed" prompt**: "Use a single sentence and short, simple, descriptive phrases to describe the detailed appearance and detailed shape of {cls}."
  - The Egyptian Cat has a slender, muscular build, a distinctively slender face with large pointy ears, almond-shaped green or gold eyes, and a short, smooth coat in colors such as black, white, or tortoiseshell.
  - The space bar is a horizontal, elongated key located at the bottom center of a computer keyboard, typically featuring a smooth, concave surface, and comes in various sizes and colors for different models.
  - The hair spray is a pressurized aerosol can or pump bottle with a plastic or metal container, typically transparent or colored, and comes in various strengths and formulas, used for holding and styling the hair in place.
- **"More Distinct" prompt**: "Use a single sentence to describe the unique, distinctive appearance and shape

of {cls}. Only describe the unique, distinctive shape and appearance.”

- The Egyptian Cat, also known as the Egyptian Mau, is characterized by its medium size, muscular build, distinctive spots on its fur, a black stripe running from its eyes to the neck, and green or gold eyes.
- The space bar has a long, rectangular shape with a concave top and a convex bottom, and is wider than most other keys on the keyboard.
- Hair spray usually comes in a cylindrical container with a long nozzle that sprays a fine mist, and may have a cap or cover on top to protect the nozzle.

7. More details on Comparing Different Few Shot Learning Strategies

In this section, we provide more implementation details on comparing finetuning student visual backbone vs. training-free retrieval for few shot learning on $\mathcal{X}_{\text{ood}}$. Finetuning student visual backbone follows the same settings in Appendix Sec. 4. For training-free retrieval, we adopt an approach similar to Tip-Adapter [7], except that we use the student image model S as the visual encoder and teacher's language model T_{img} as the textual encoder. We set $\alpha = 1.0$ and $\beta = 5.5$.

8. Additional details on Appendix Sec. 2 Application

For our experiments in Appendix Sec. 2, we are in a multi-label classification setting, and the student network needs to output whether it is feasible to grasp each YCB object given the current observation. During student training, we find learning a common positive prompt and a common negative prompt for all labels to be very helpful, and we adopt DualCoOp [5] to learn these prompts. To encourage the global student visual feature be region-aware, we adopt the region aggregation method from DualCoOp for all of our experiments. We additionally calibrate the positive and negative probabilities (for each label, positive probability + negative probability = 1) by learning a common probability bias on the fly. An object y is then predicted as positive if its positive probability is greater than $0.5 + \text{bias}$.

The two evaluation metrics we calculated in Appendix Sec. 2 can be formally defined as follows:

- Overall accuracy over all YCB objects in the label set $\mathcal{Y}$ ($\mathcal{Y} = \mathcal{Y}_{\text{id}}$ on $\mathcal{X}_{\text{id}}$ and $\mathcal{Y} = \mathcal{Y}_{\text{id}} \cup \mathcal{Y}_{\text{ood}}$ on $\mathcal{X}_{\text{ood}}$):

$$\mathcal{M}_1 = \frac{\sum_{i=1}^{|\mathcal{Y}|} \sum_{j=1}^{|\mathcal{X}|} \text{correct}(x_j, y_i)}{|\mathcal{X}||\mathcal{Y}|} \tag{1}$$

Here $\text{correct}(x_j, y_i)$ outputs 1 if the existence of object y_i is predicted correctly in the observation x_j, and 0 otherwise.

- F1-measure over objects that exist in an observation, averaged over all observations:

$$\begin{aligned}
\mathcal{M}_{2,\text{precision}}(y) &= \frac{\sum_{i=1}^{|\mathcal{X}|} \mathbf{1}_{y \in \text{obj}(x_i)} * \text{correct}(x_i, y)}{\sum_{i=1}^{|\mathcal{X}|} \min(\mathbf{1}_{y \in \text{obj}(x_i)} + \text{pred}(x_i, y), 1)} \\
\mathcal{M}_{2,\text{precision}} &= \frac{\sum_{y \in \mathcal{Y}} \mathcal{M}_{2,\text{precision}}(y)}{|\mathcal{Y}|} \\
\mathcal{M}_{2,\text{recall}}(y) &= \frac{\sum_{i=1}^{|\mathcal{X}|} \mathbf{1}_{y \in \text{obj}(x_i)} * \text{correct}(x_i, y)}{\sum_{i=1}^{|\mathcal{X}|} \mathbf{1}_{y \in \text{obj}(x_i)}} \\
\mathcal{M}_{2,\text{recall}} &= \frac{\sum_{y \in \mathcal{Y}} \mathcal{M}_{2,\text{recall}}(y)}{|\mathcal{Y}|} \\
\mathcal{M}_{2,\text{F1}} &= \frac{2}{\frac{1}{\mathcal{M}_{2,\text{precision}}} + \frac{1}{\mathcal{M}_{2,\text{recall}}}}
\end{aligned} \tag{2}$$

Here $\text{pred}(x_i, y)$ equals 1 if the student network predicts that the object y exists in observation x_i, and 0 otherwise.

References

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022. 1

[2] Berk Calli, Aaron Walsman, Arjun Singh, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. Benchmarking in manipulation research: The ycb object and model set and benchmarking protocols. *arXiv preprint arXiv:1502.03143*, 2015. 1

[3] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 2

[4] Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, Xiaodi Yuan, Pengwei Xie, Zhiao Huang, Rui Chen, and Hao Su. Maniskill2: A unified benchmark for generalizable manipulation skills. In *International Conference on Learning Representations*, 2023. 1

[5] Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. *arXiv preprint arXiv:2206.09541*, 2022. 2, 4

[6] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 1

[7] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. 4