Overcoming the Pitfalls of Vision-Language Model Finetuning for OOD Generalization

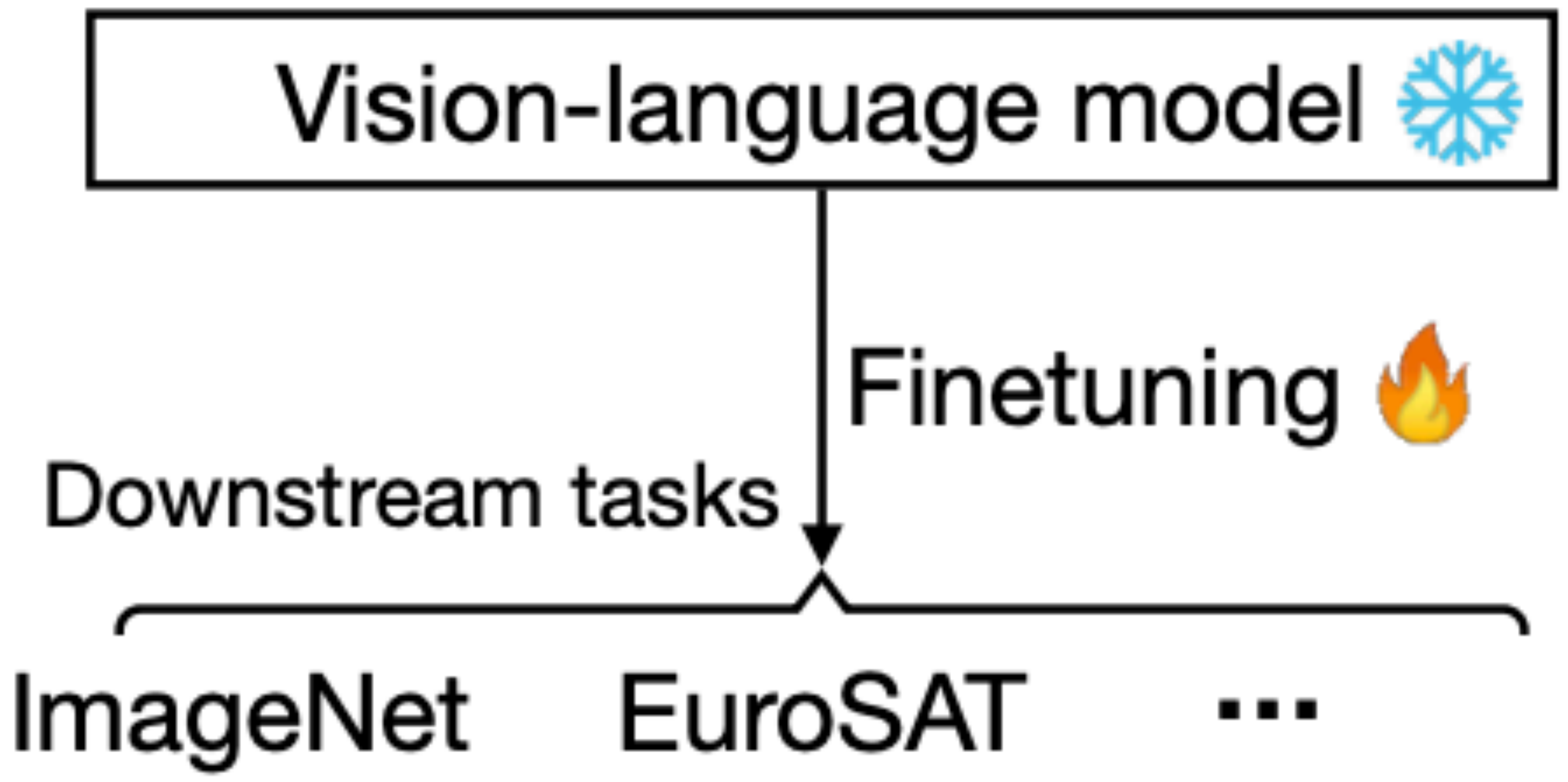
Yuhang Zang[1], Hanlin Goh[2], Josh Susskind[2], Chen Huang[2]

[1]Nanyang Technological University, [2]Apple

ICLR 2024 | Apple | May 7, 2024

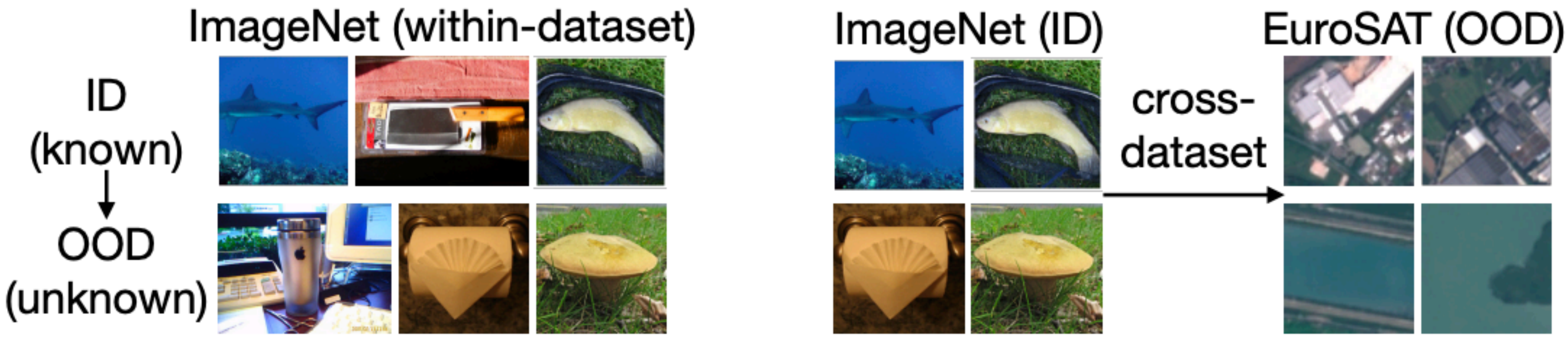
Background

- We study OOD generalization when finetuning vision-language models (e.g., CLIP) on downstream tasks.



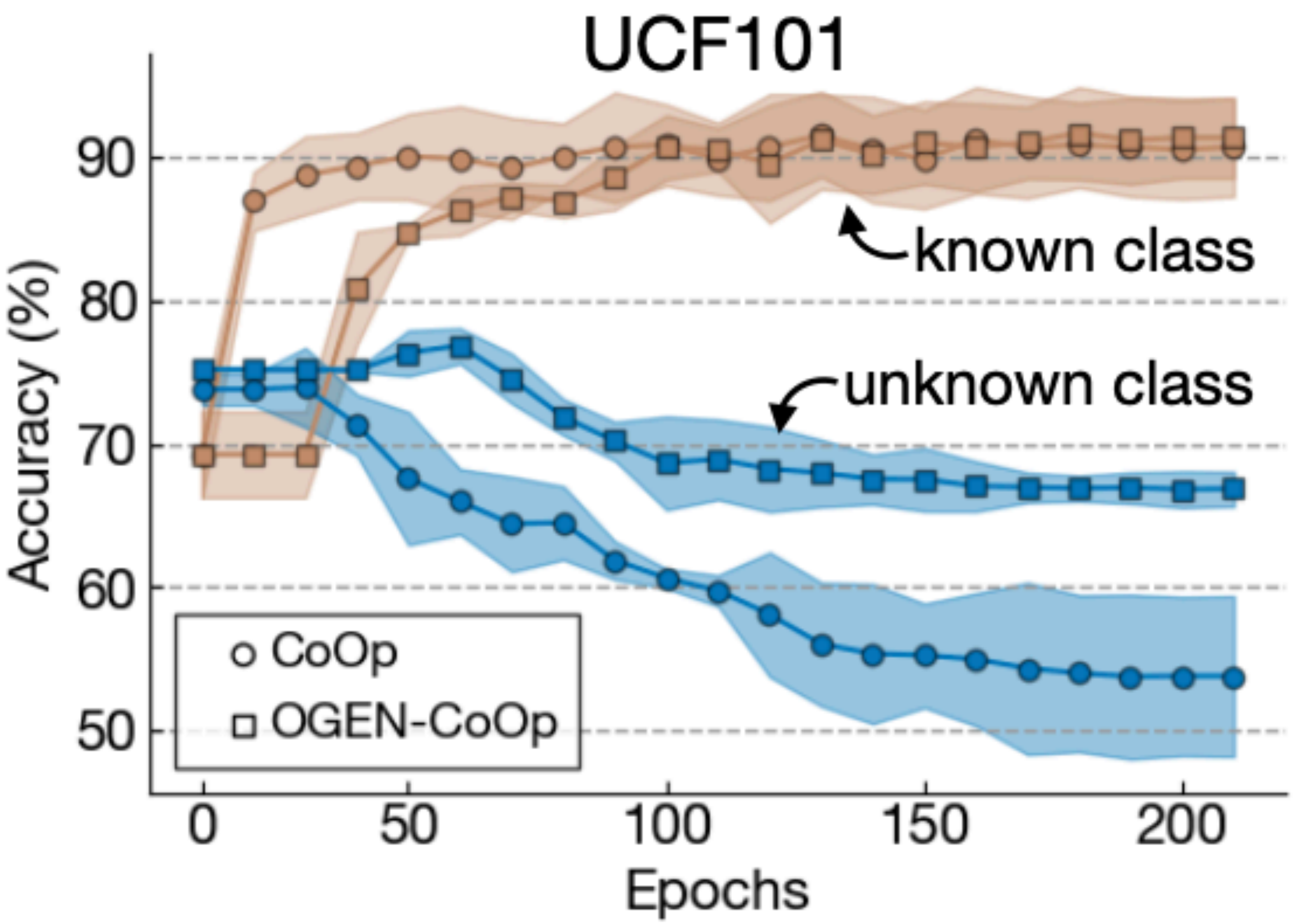
Background

- Two settings for OOD generalization:
  - Within-dataset
  - Cross-dataset



Motivation

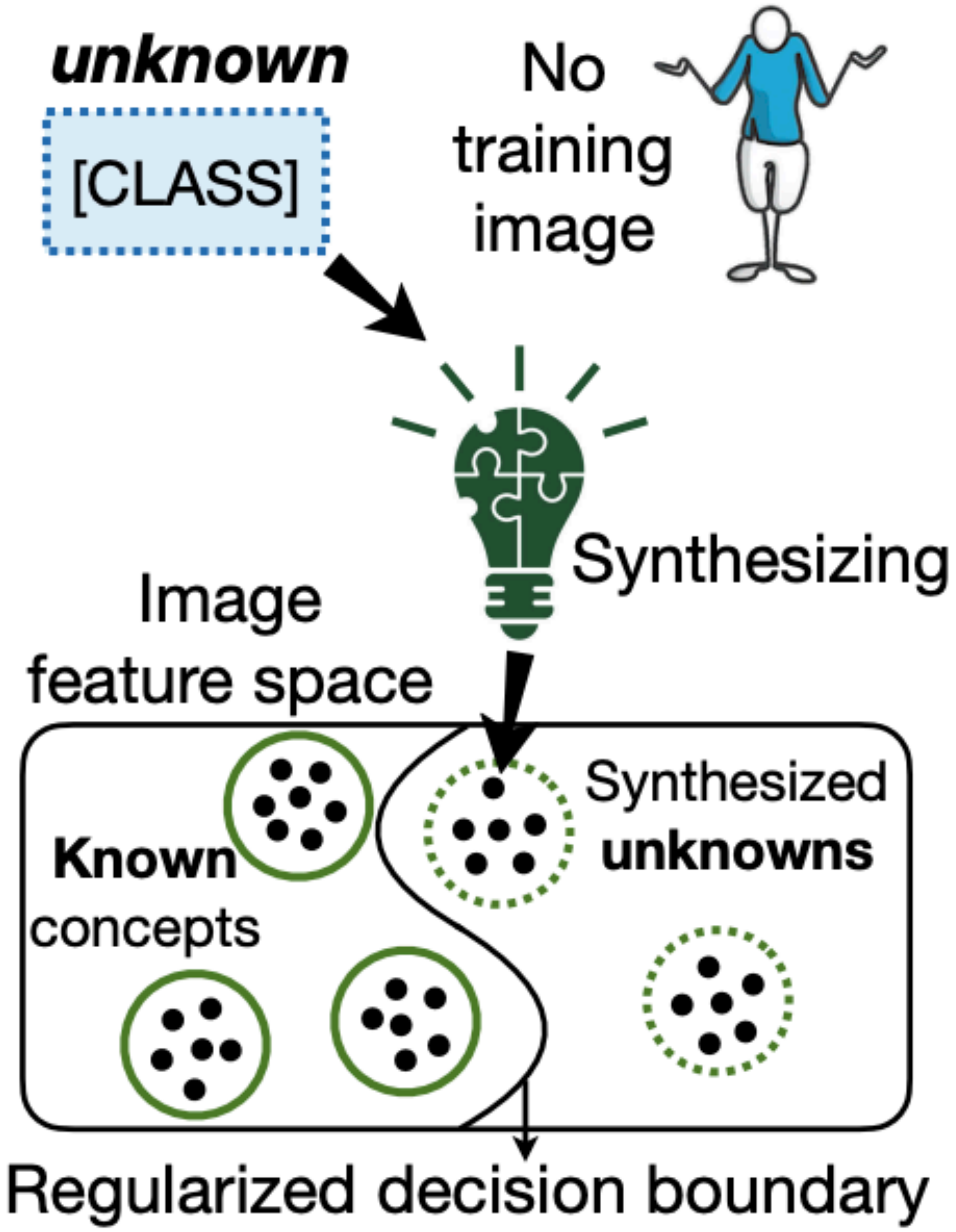
- Recent finetuning methods for vision-language models often lead to overfitting



- We propose **OGEN**: our approach to improve **O**OD **GEN**eralization

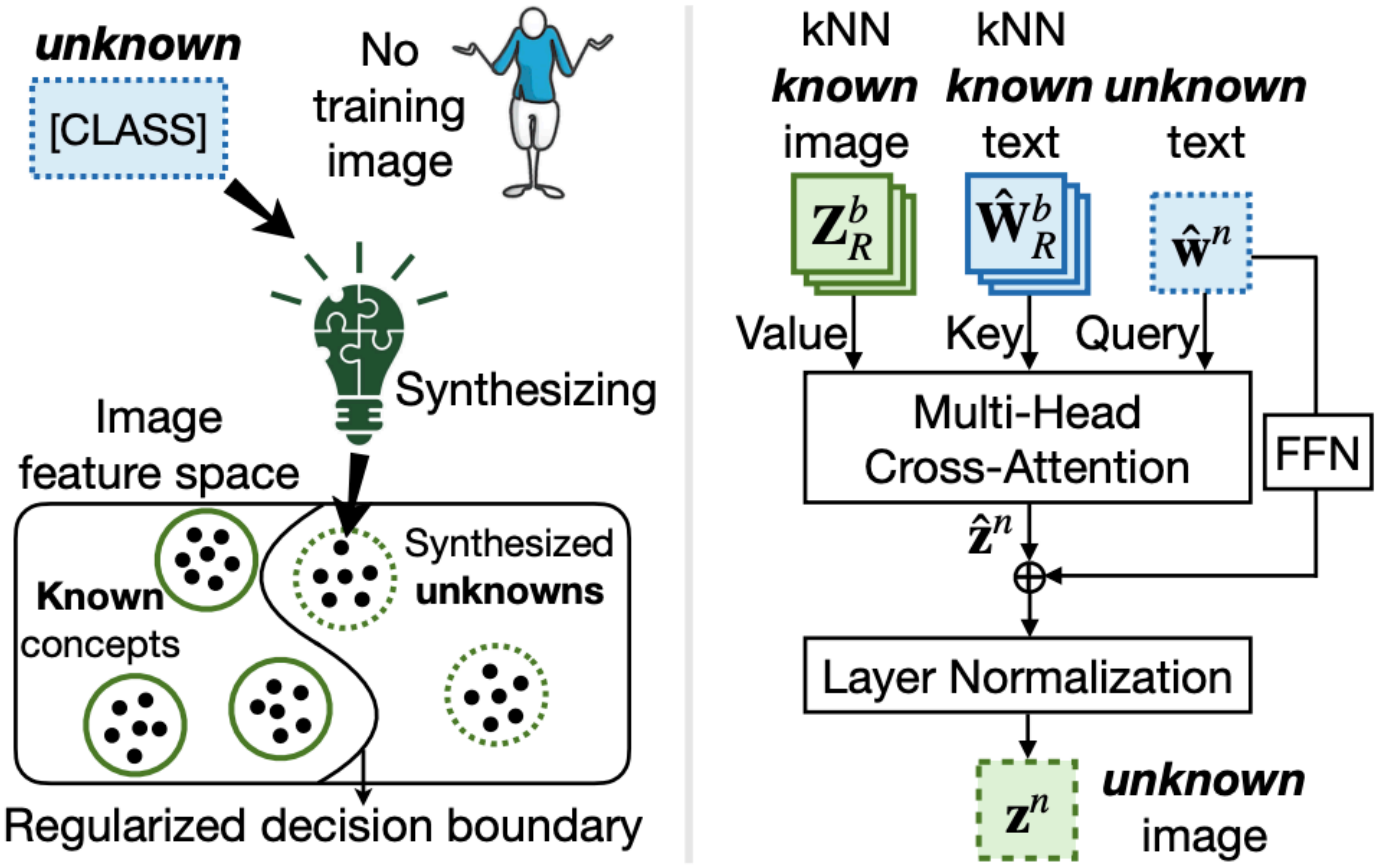
Method

Main contribution: jointly trained class-conditional feature generator



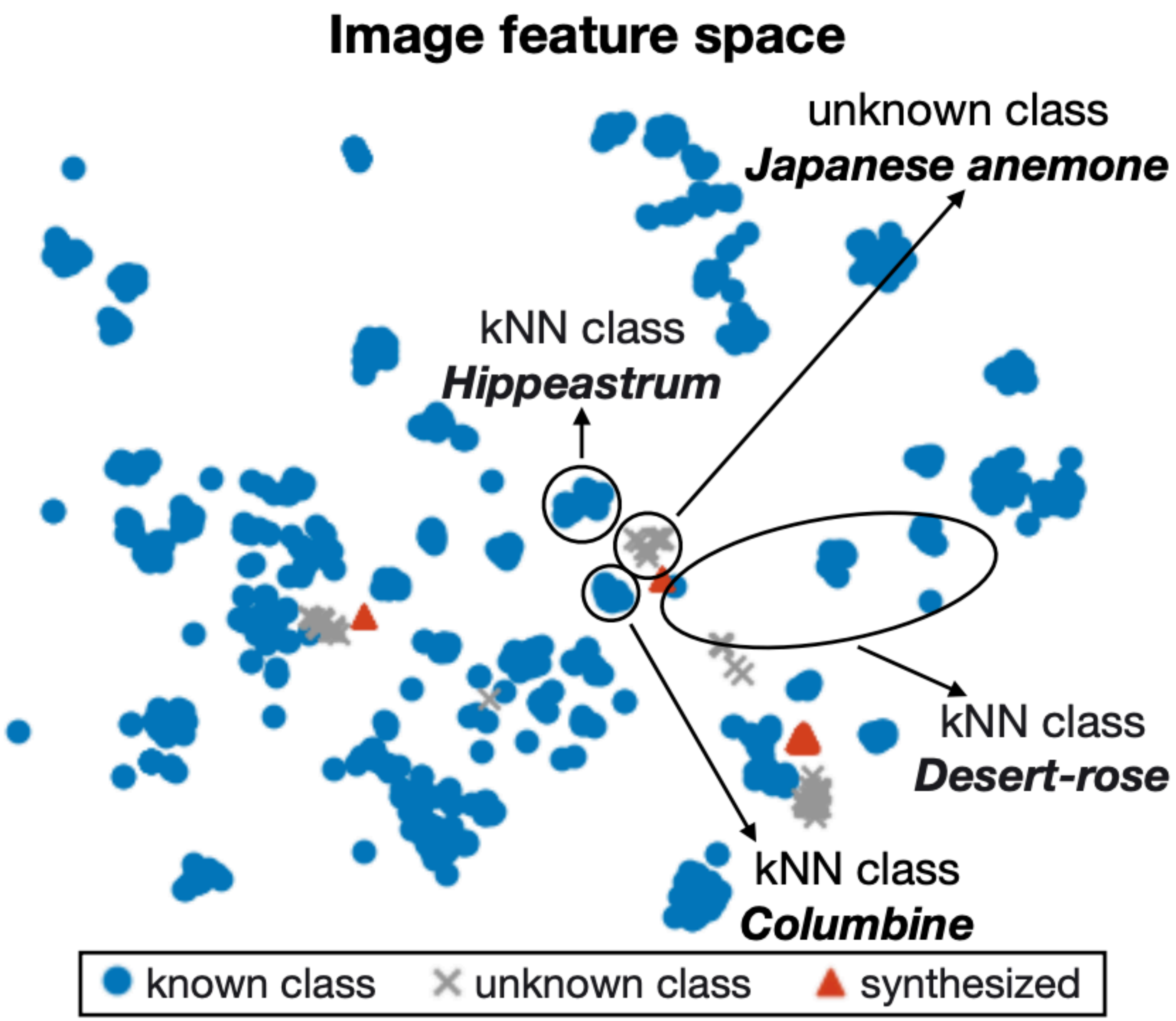
Method

Implementation of the class-conditional feature generator



Qualitative Results

Visualization: unknown image feature synthesis via extrapolation



Optimization

- Joint optimization of ***known*** and synthetic ***unknown*** class data
- Adaptive self-distillation on the ***unknown*** feature generator to further reduce overfitting
  - Mean Teacher model with adaptive window size

Mean Teacher $$\mathbf{MT}_{[1,t]}:\ \theta_i^T = \alpha\theta_{i-1}^T + (1-\alpha)\theta_i, \quad for \quad i = \{1, \ldots, t\},$$

Adaptive window $$\mathbf{ALMT}_t:\ \mathbf{MT}_{[t-m_t,t]}, \quad m_t = \left\lfloor \left(1 + \cos\left(\frac{t_{\max}+t}{t_{\max}}\pi\right)\right) \cdot \frac{1}{2}(m_{\max} - m_{\min}) + m_{\min} \right\rfloor,$$

Main Results

- Within-dataset generalization (base-to-new class)

	+OGEN	CoOp ✗	CoOp ✓	CoCoOp ✗	CoCoOp ✓	VPT ✗	VPT ✓	SHIP ✗	SHIP ✓	KgCoOp ✗	KgCoOp ✓	MaPLe ✗	MaPLe ✓	PromptSRC ✗	PromptSRC ✓
Avg across 11 datasets	Base	82.69	**83.47**	**80.47**	79.86	82.51	**82.52**	80.03	**80.79**	80.73	**81.34**	82.28	**82.40**	**84.26**	84.17
	New	63.22	**69.54**	71.69	**73.35**	69.01	**70.61**	73.69	**76.14**	73.60	**75.68**	75.14	**76.37**	76.10	**76.86**
	Δ		**+6.32**		**+1.66**		**+1.60**		**+2.45**		**+2.08**		**+1.23**		**+0.76**
	H	71.66	**75.87**	75.83	**76.47**	75.16	**76.10**	76.73	**78.40**	77.00	**78.40**	78.55	**79.27**	79.97	**80.34**

- Cross-dataset generalization

	Source	Target										
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAir	SUN397	DTD	EuroSAT	UCF101	*Average*
CoOp	71.51	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
OGEN-CoOp	**71.52**	94.60	90.73	65.07	70.55	87.26	19.84	65.77	44.90	49.53	69.36	**65.76**
CoCoOp	71.02	94.43	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21	65.74
OGEN-CoCoOp	**71.28**	95.12	91.37	66.04	72.90	86.54	22.95	68.42	46.38	45.82	69.74	**66.53**

Conclusions

- Study and improve OOD generalization of CLIP finetuning
- Class-conditional feature generator helps regularize the unknowns
- Adaptive self-distillation scheme to further reduce overfitting
- Superior generalization capability under different OOD settings

Code link

TM and © 2024 Apple Inc. All rights reserved.