Few-Shot Recognition via Stage-Wise Retrieval-Augmented Finetuning

Tian Liu[1] Huixin Zhang[1] Shubham Parashar[1] Shu Kong[2]

[1]Texas A&M University [2]University of Macau

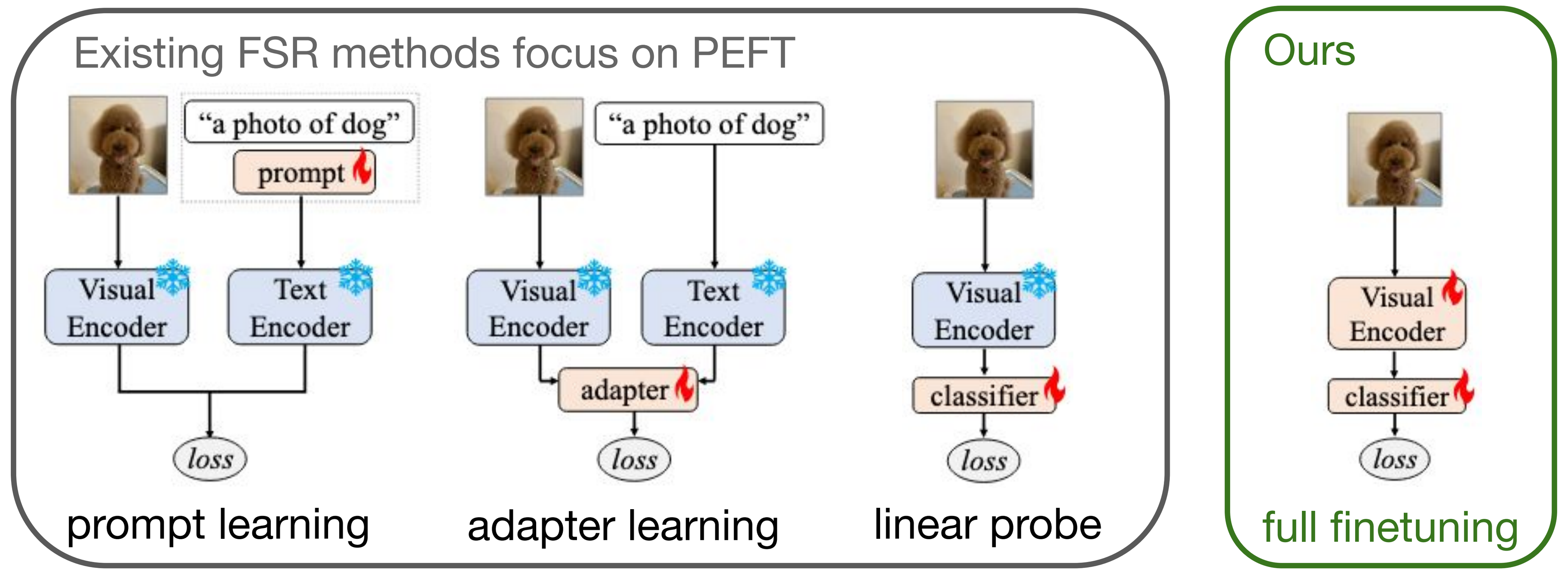
CVPR Nashville JUNE 11-15, 2025

SWAT

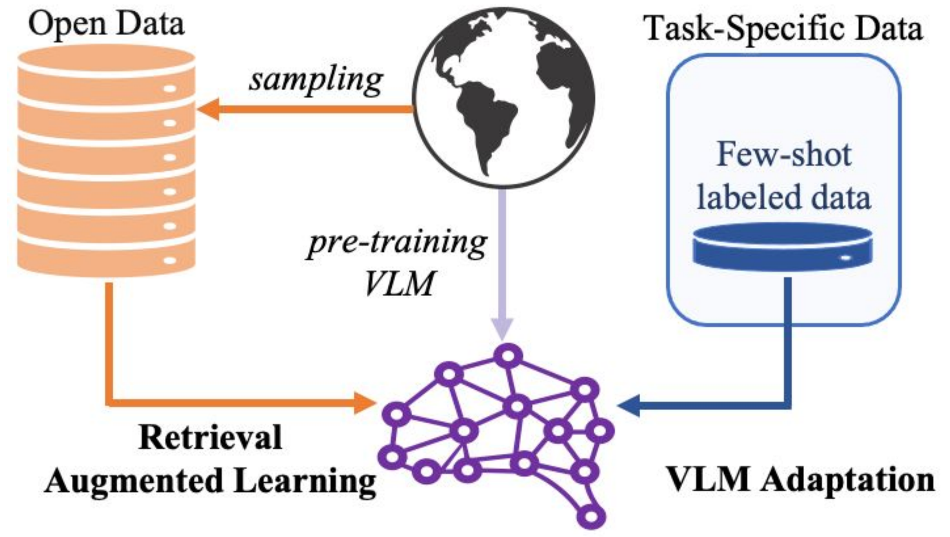
Problem Formulation in the Open World

Few-Shot Recognition (FSR) aims to solve a recognition task by training over only a few labeled task-specific examples per concept concerned by the task.

- Recent FSR methods commonly adopt *parameter-efficient finetuning* (PEFT) with a Vision-Language Model (VLM), which learns a small number of parameters.
- Instead, taking the perspective of *data annotation*[1] that **prioritizes accuracy**, we solve FSR by exploring more methods to adapt a VLM.

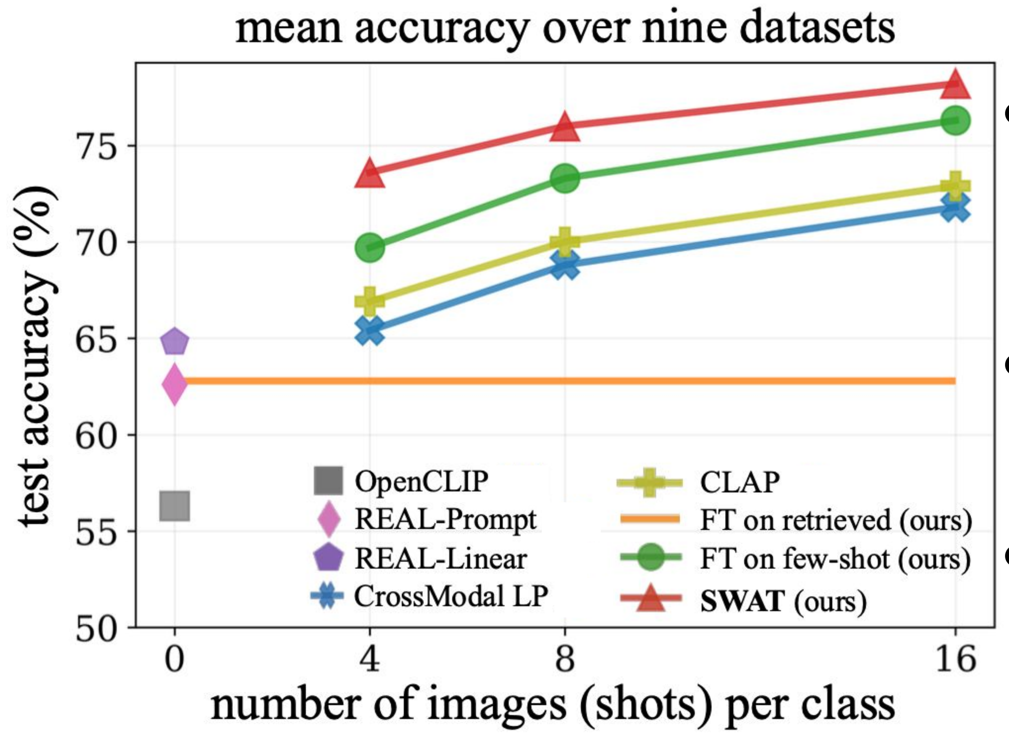


Open-World Pretraining and Open-Data Retrieval



- We exploit the extraordinary zero-shot transfer capability of an **open-world pretrained foundational VLM**.
- We **retrieve open data** (esp. the VLM's publicly-available pretraining dataset[2]) to augment the limited number of labeled task-specific data.

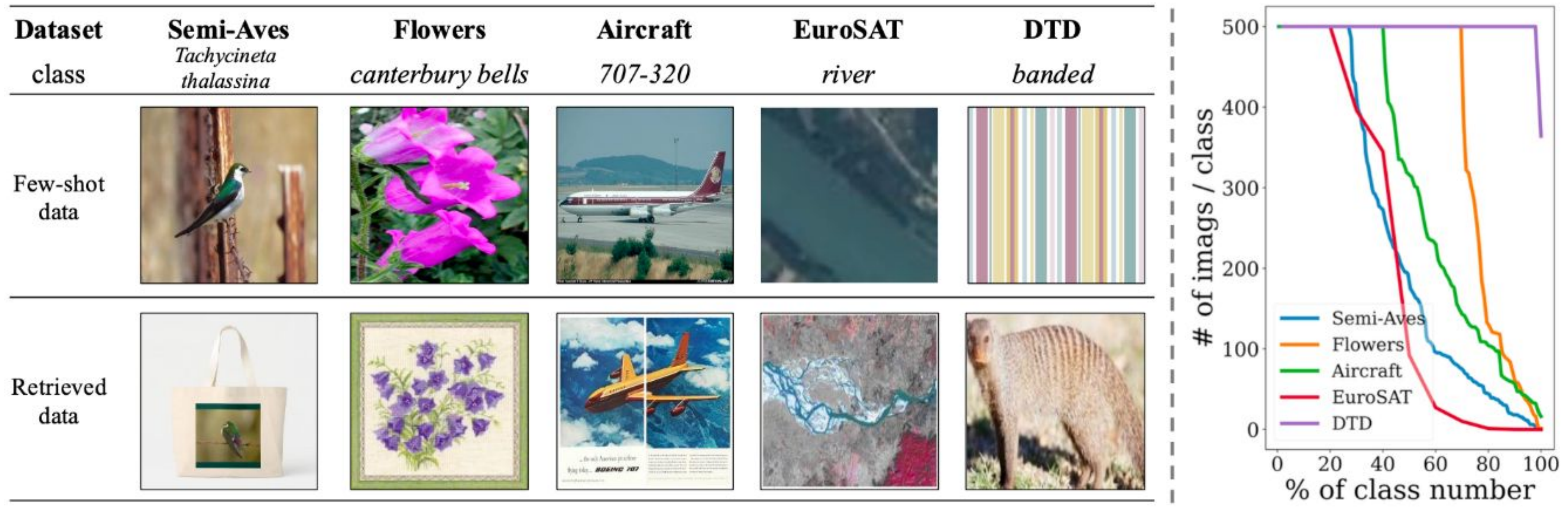
Performance Overview



- Finetuning on a large amount of retrieved data barely surpass SOTA zero-shot methods due to *domain gaps* and *imbalance distributions*.
- Simply finetuning a VLM on few-shot examples alone outperforms existing FSR methods by 3 in accuracy.
- Our method **SWAT** outperforms SOTA FSR by >6 in accuracy.

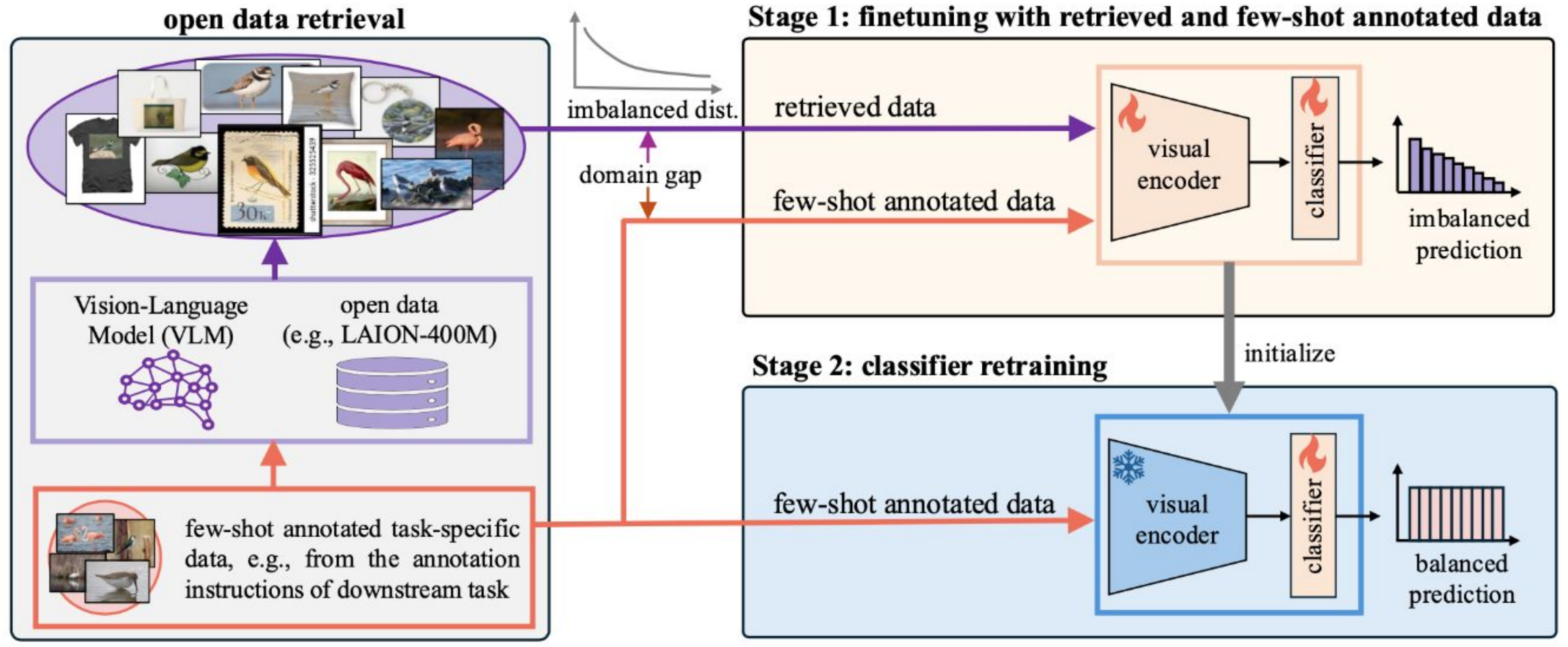
Insights and Methods

Domain gaps and imbalanced distributions exhibited by the retrieved data

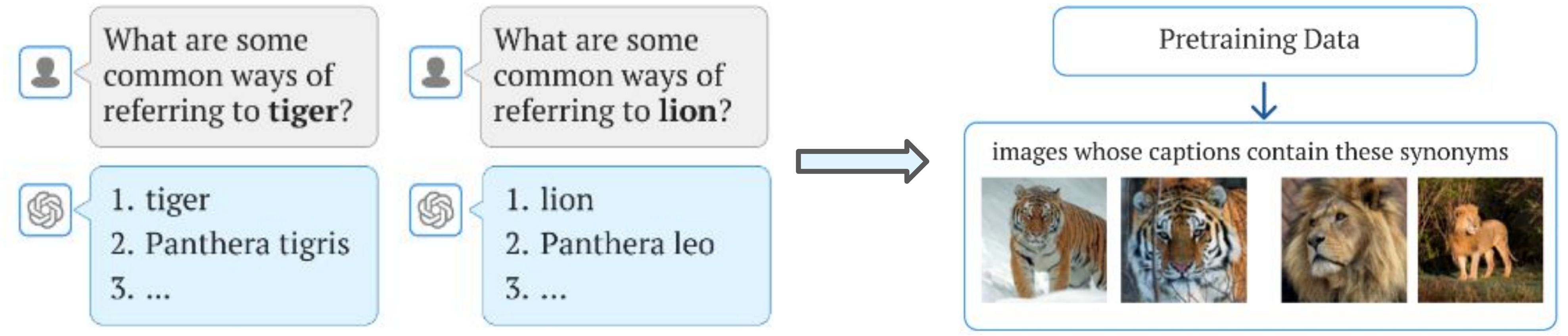


Addressing the above issues by SWAT (Stage-Wise retrieval-Augmented fineTuning)

- Decouple representation learning and classifier learning to mitigate imbalanced training[3].
- Retraining the classifier practices transfer learning to mitigate domain gaps.



"String-matching" based retrieval improves efficiency and diversity.[4]



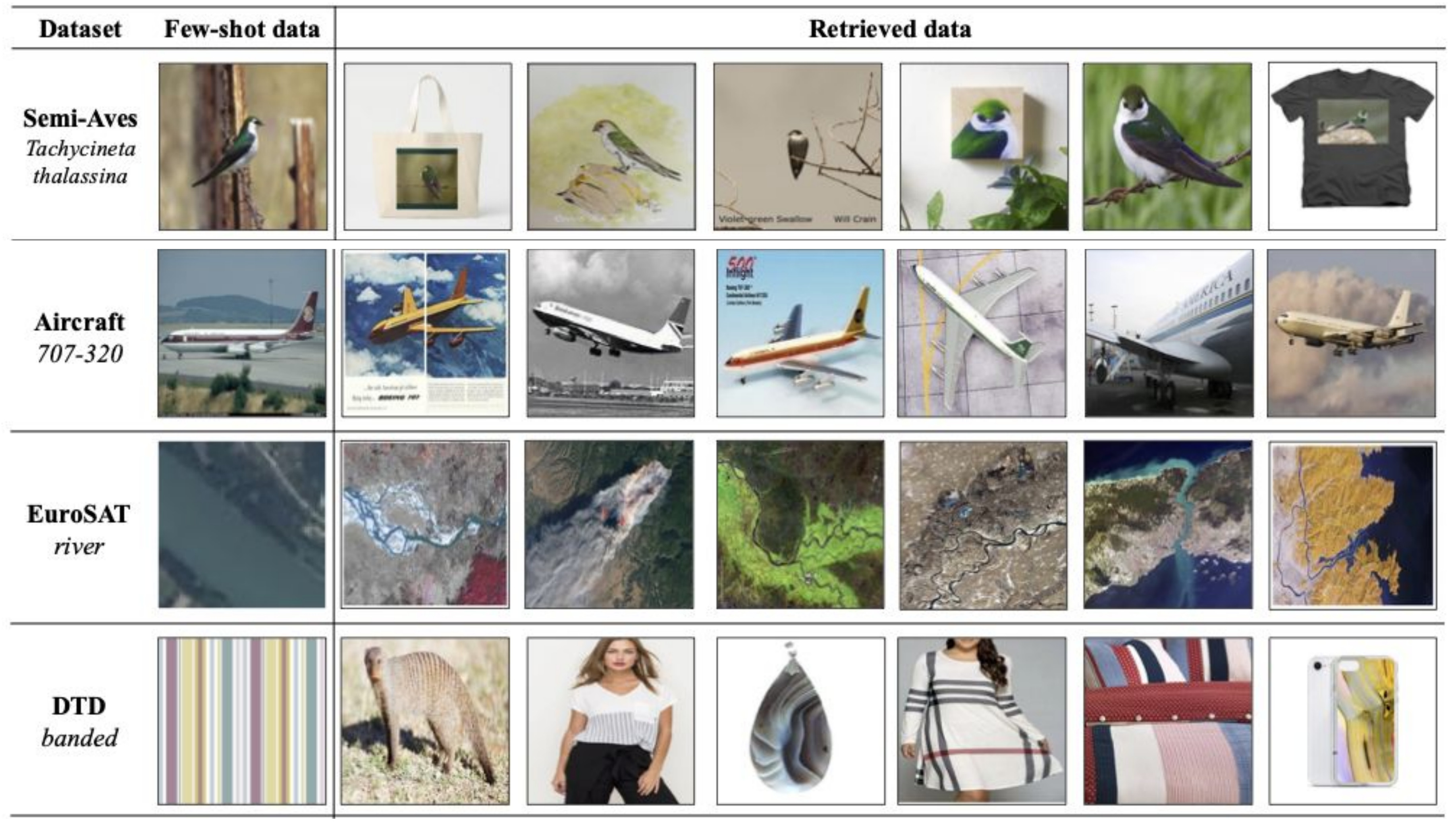
Results

	strategy	method	venue & year	mean accuracy of nine datasets		
zero-shot methods	prompting-based	OpenCLIP	CVPR 2023	56.3		
		REAL-Prompt	CVPR 2024	62.6		
	retrieval-augmented	REAL-Linear	CVPR 2024	64.8		
				4-shot	*8-shot*	*16-shot*
few-shot methods	prompt-learning	CoOp	IJCV 2022	61.0	64.6	68.4
		PLOT	ICLR 2023	62.9	65.7	68.7
	adapter-based	CLIP-Adapter	IJCV 2023	59.6	64.5	68.1
		TIP-Adapter	ECCV 2022	56.6	57.8	59.5
		TIP-Adapter(f)	ECCV 2022	60.8	63.5	67.1
		TaskRes(e)	ECCV 2022	63.5	67.1	69.9
		CrossModal-LP	CVPR 2023	65.4	68.8	71.8
		CLAP	CVPR 2024	66.9	70.0	72.9
	finetuning-based	few-shot finetuning	ours	69.7 (+2.8)	73.3 (+3.3)	76.3 (+3.4)
		SWAT	ours	**73.5** (+6.6)	**76.0** (+6.0)	**78.2** (+5.3)

method	venue & yr	mem.	time	mean acc.
CrossModal LP	CVPR'23	2 GB	2 mins	71.8
CLAP	CVPR'24	2 GB	2 mins	72.9
few-shot finetuning	ours	5 GB	20 mins	**76.3** (+3.4)
SWAT retrieval	ours	2 GB	1 hr	**78.2** (+5.3)
SWAT training	ours	5 GB	2.5 hrs	

- Our simple few-shot finetuning surpasses SOTA FSR by 3 in accuracy, without overfitting issue.
- Our SWAT outperforms SOTA FSR by 6 in accuracy, with only small overhead.

Demonstrations of Retrieved Data



References

[1] "Revisiting few-shot object detection with vision-language models". NeurIPS, 2024

[2] "LAION-400m: Open dataset of clip-filtered 400 million image-text pairs". arXiv preprint arXiv:2111.02114, 2021.

[3] "Decoupling representation and classifier for long-tailed recognition". ICLR, 2020.

[4] "The Neglected Tails in Vision Language Model". CVPR, 2024.