Enhanced Continual Learning of Vision-Language Models with Model Fusion

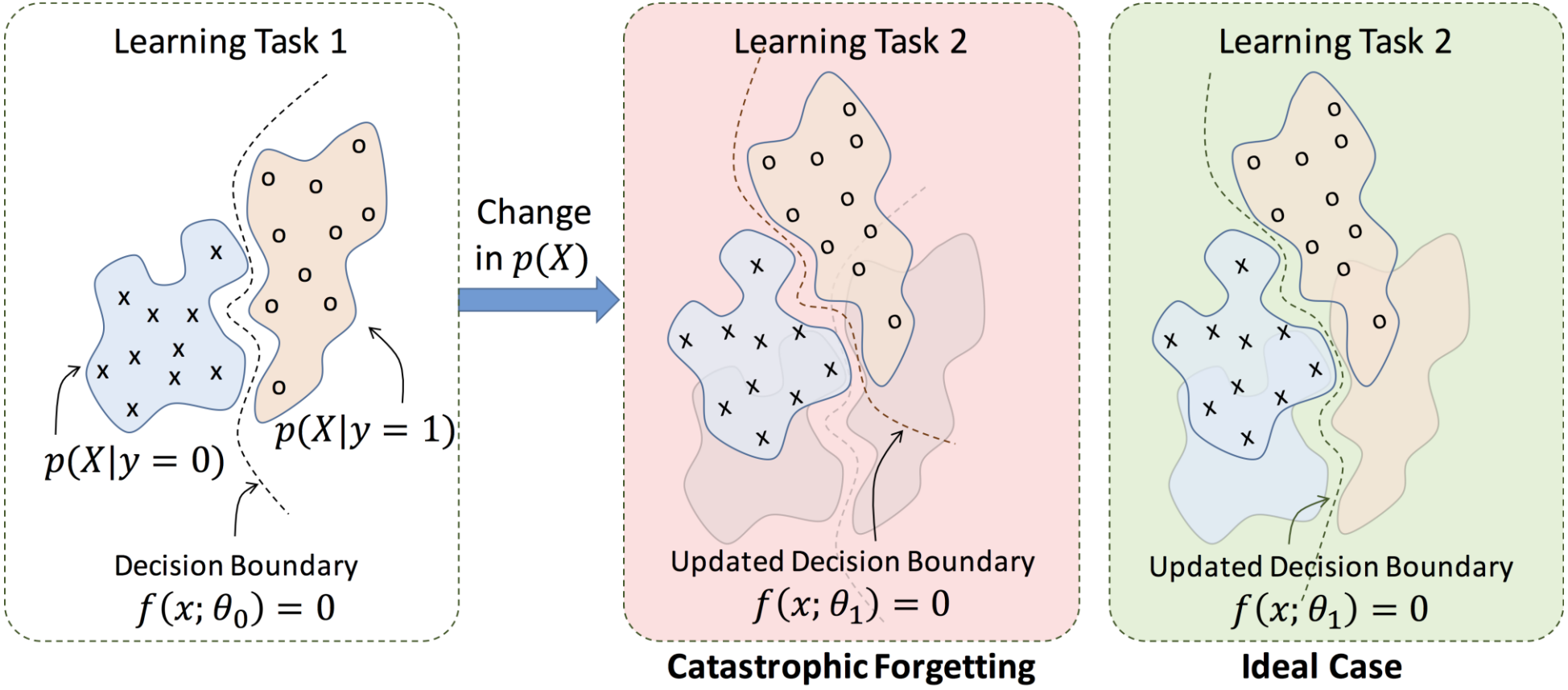
Haoyuan Gao *, Zicong Zhang *, Yuqi Wei , Linglan Zhao ,
Guilin Li , Yexin Li , Bo Wang, Linghe Kong , Weiran Huang †

Speaker : Zicong Zhang

20th March 2026

SJTU MIFA LAB

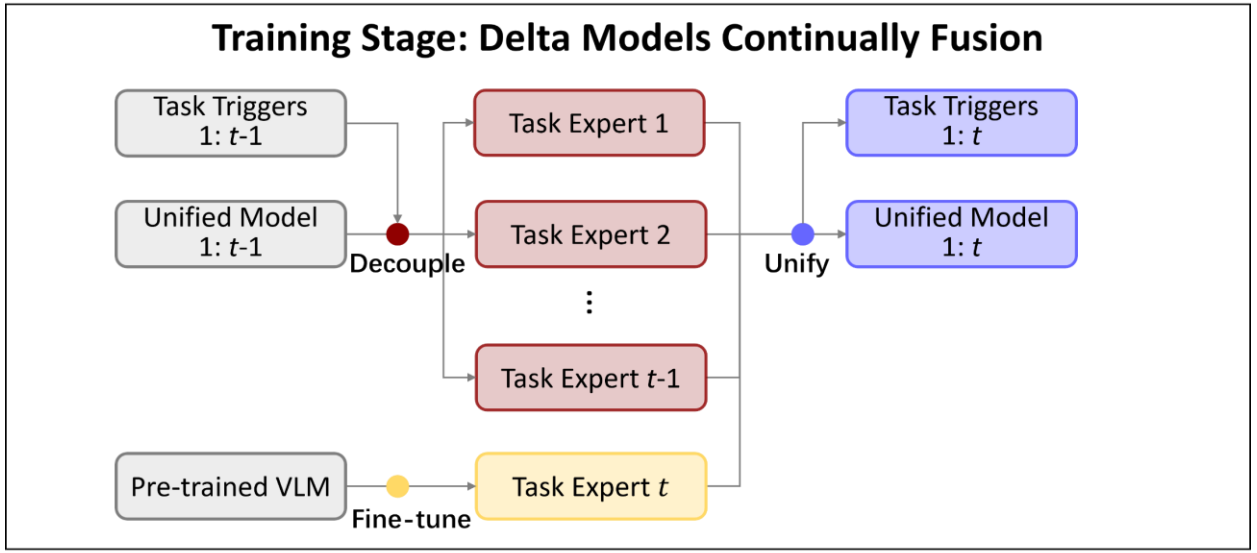
Motivation

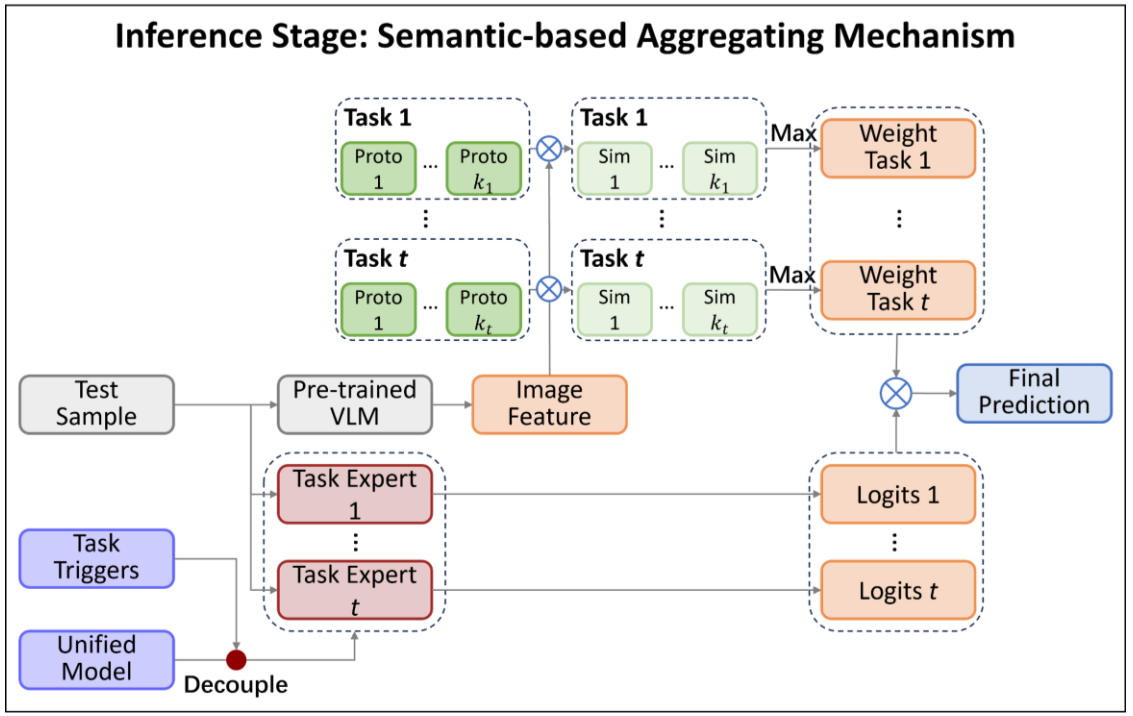


Kolouri, Soheil & Ketz, Nicholas & Zou, Xinyun & Krichmar, Jeff & Pilly, Praveen. (2019). Attention-Based Selective Plasticity.

Methods

- We introduce model fusion to VLMs and propose a novel **Decoupling-Unifying framework** compatible with PEFT and full-finetune paradigms.





Methods

Delta Models Continually Fusion at Training Stage :

1. Tuning Individually：

finetune pre-trained VLM on *Current Dataset t* to get θ^t

subtracting θ^t from pre-trained model θ^0 to obtain delta model $\delta^t = \theta^t - \theta^0$

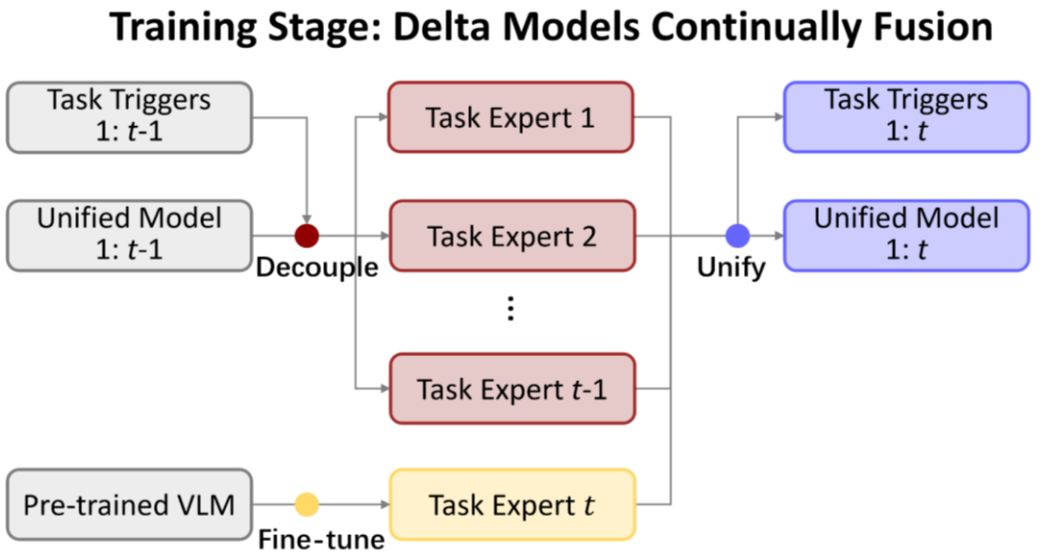
2. Decoupling Unified Model :

apply *Task Triggers* on *Unified Model* to reconstruct models

$\tilde{\delta}^i = \lambda^i M^i \odot \delta^{1:t}$ $\quad\quad$ $\tilde{\theta}^i = \tilde{\delta}^i + \theta^0$

3. Unifying Models :

combine reconstructed models $\tilde{\delta}^i$and δ^t to get unified delta model $\delta^{1:t} = \text{unify}(\tilde{\delta}^1, \tilde{\delta}^2 \cdots \delta^t)$



Methods

Semantic-Based Aggregating Mechanism at Inference Stage:

1. Computing Prototypes :

for each category in each task, save its prototype

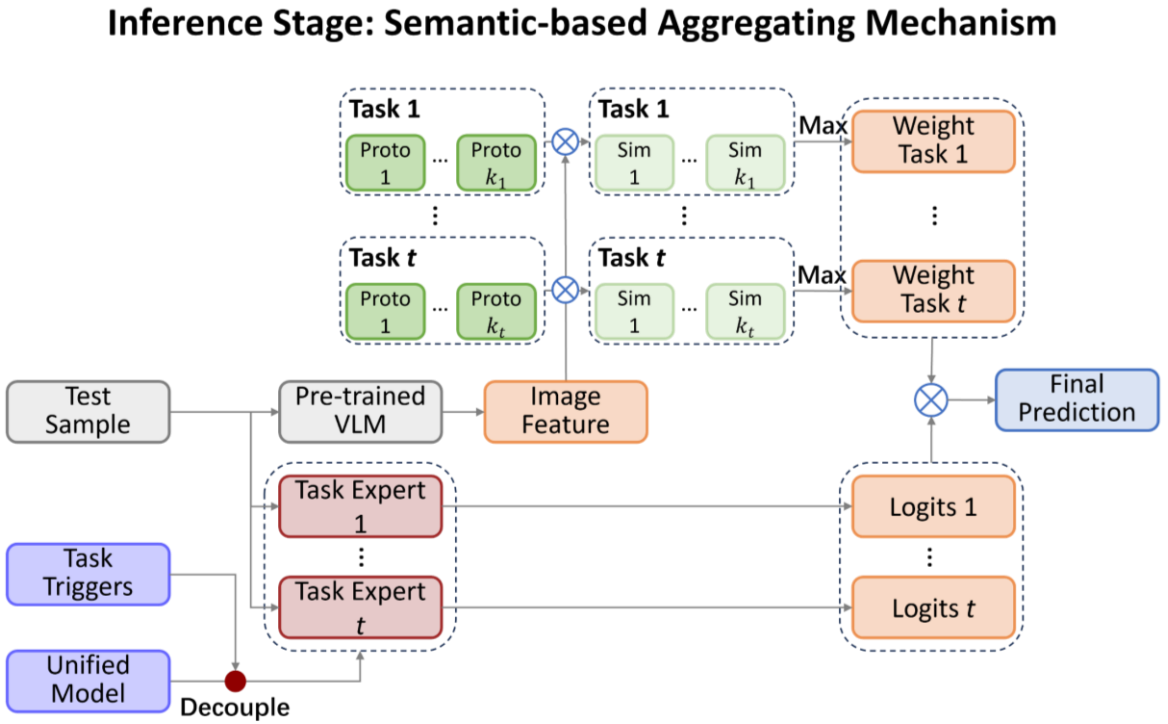
2. Aggregating Predictions :

- for a test image with task-id

  directly use the corresponding reconstructed model

- for a test image without task-id or from unseen tasks

  weighted fuse the predictions of corresponding K models based on feature cosine similarity



Methods

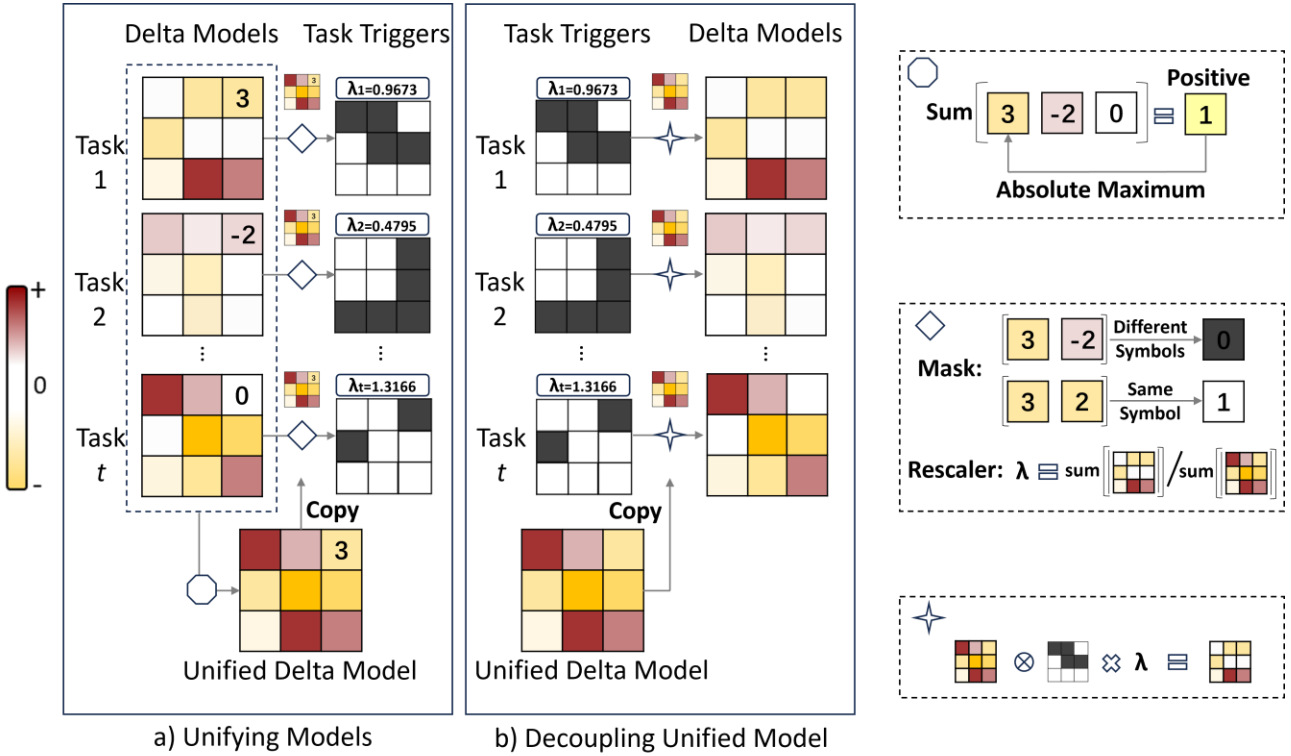
Unifying and Decoupling:

1. Election-based Unifying:

select the value with the largest magnitude that aligns with the majority sign

2. Decoupling vis Task-Triggers:

apply task-triggers to unified delta model to reconstruct task-expert with high fidelity



Experiments : MTIL

We evaluate our method on Multi-domain Task Incremental Learning (MTIL) benchmark

	Method	Aircraft	Caltech101	CIFAR100	DTD	EuroSAT	Flowers	Food	MNIST	OxfordPet	Cars	SUN397	*Average*
	Zero-shot	24.3	88.4	68.2	44.6	54.9	71.0	88.5	59.4	89.0	64.7	65.2	65.3
	Individual FT	62.0	95.1	89.6	79.5	98.9	97.5	92.7	99.6	94.7	89.6	81.8	89.2
Transfer	ZSCL	-	86.0	67.4	45.4	50.4	69.1	87.6	61.8	86.8	60.1	66.8	68.1
	Dual-RAIL	-	88.4	68.2	44.6	54.9	71.0	88.5	59.6	89.0	64.7	65.2	69.4
	DPeCLIP	-	88.2	67.2	44.7	54.0	70.6	88.2	59.5	89.0	64.7	64.8	69.1
	MulKI	-	87.8	69.0	46.7	51.8	71.3	88.3	64.7	89.7	63.4	68.1	70.1
	ConDU (LoRA)	-	88.1	68.9	45.7	57.0	71.3	88.8	61.2	89.3	65.1	67.8	70.3
	ConDU (FT)	-	88.1	68.9	46.4	57.1	71.4	88.7	65.5	89.3	65.0	67.8	**70.8**
Average	ZSCL	45.1	92.0	80.1	64.3	79.5	81.6	89.6	75.2	88.9	64.7	68.0	75.4
	Dual-RAIL	52.5	96.0	80.6	70.4	81.3	86.3	89.1	73.9	90.2	68.5	66.5	77.8
	DPeCLIP	49.9	94.9	82.4	69.4	82.2	84.3	90.0	74.0	90.4	68.3	66.3	77.5
	MulKI	52.5	93.6	79.4	67.0	79.8	83.9	89.6	77.1	91.2	67.1	69.1	77.3
	ConDU (LoRA)	51.9	94.9	84.4	69.8	81.1	84.4	90.0	77.3	89.5	69.0	69.3	78.3
	ConDU (FT)	59.6	93.4	83.7	68.1	83.4	83.7	90.1	76.7	90.6	68.6	68.6	**78.8**
Last	ZSCL	40.6	92.2	81.3	70.5	94.8	90.5	91.9	98.7	93.9	85.3	80.2	83.6
	Dual-RAIL	52.5	96.8	83.3	80.1	96.4	99.0	89.9	98.8	93.5	85.5	79.2	86.8
	DPeCLIP	49.9	95.6	85.8	78.6	98.4	95.8	92.1	99.4	94.0	84.5	81.7	86.9
	MulKI	49.7	93.0	82.8	73.7	96.2	92.3	90.4	99.0	94.8	85.2	78.9	85.1
	ConDU (LoRA)	48.9	95.2	87.8	78.5	96.3	95.2	91.7	97.6	93.0	85.3	78.8	86.2
	ConDU (FT)	58.6	93.7	86.6	76.1	98.2	93.4	91.9	99.6	94.8	84.9	80.5	**87.1**

Analysis : Visualization

We perform t-SNE visualization of features extracted from training data of 10 categories from Task1 (AirCraft)

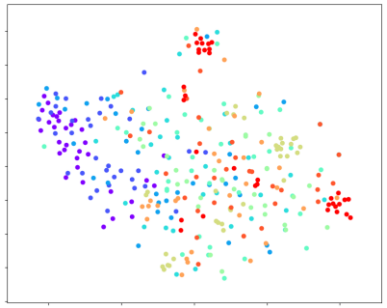


Fig 1:Pre-trained Model

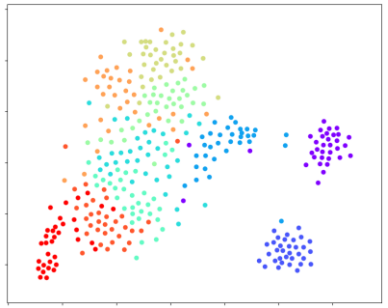
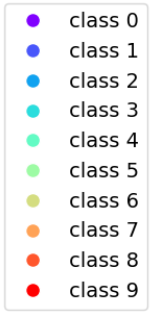


Fig 2:Session 1



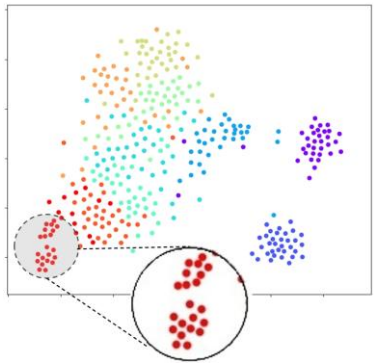


Fig 3:Session 1

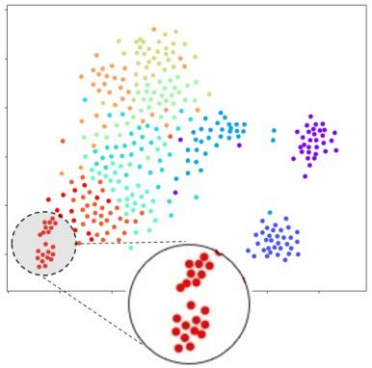


Fig 4:Session 6

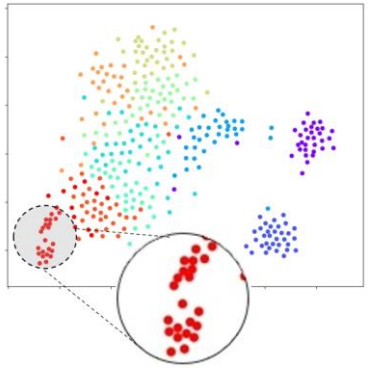
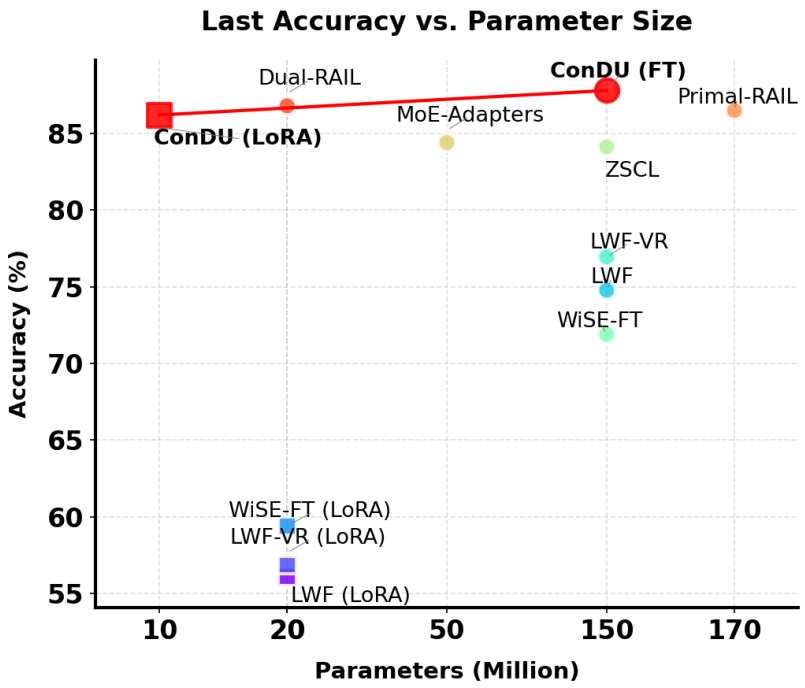


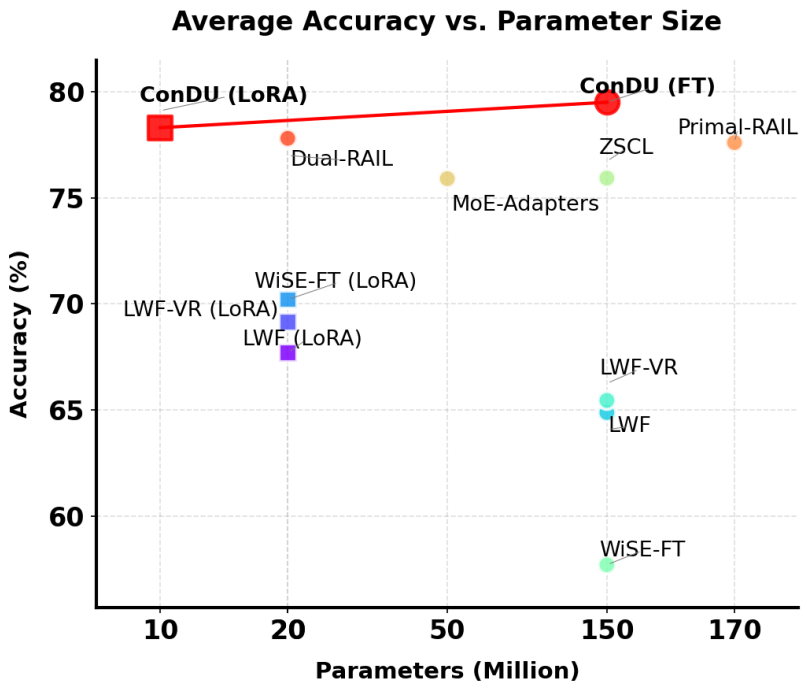
Fig 5:Session 11

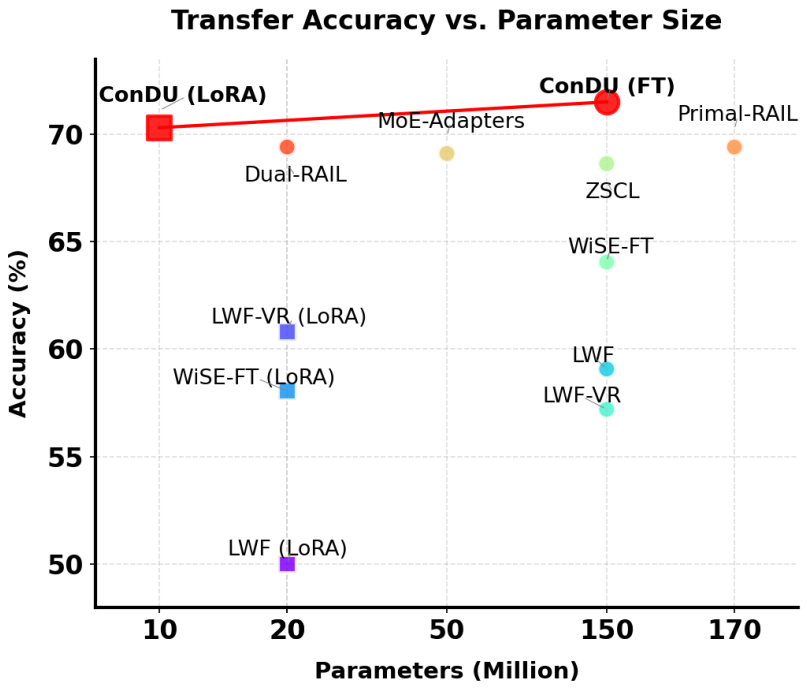
- From Fig1 and Fig2, the fine-tuned task-specific model 1 shows significantly better data discrimination onTask1 compared to the pre-trained VLM
- Fig3 to Fig5 indicates that the task-specific model reconstructed by ConDU closely matches the representation ability of the model obtained through initial fine-tuning.

Analysis : Visualization

We compare the parameter-accuracy trade-off of ConDU with other previous SOTA methods.








arXiv

Thanks!