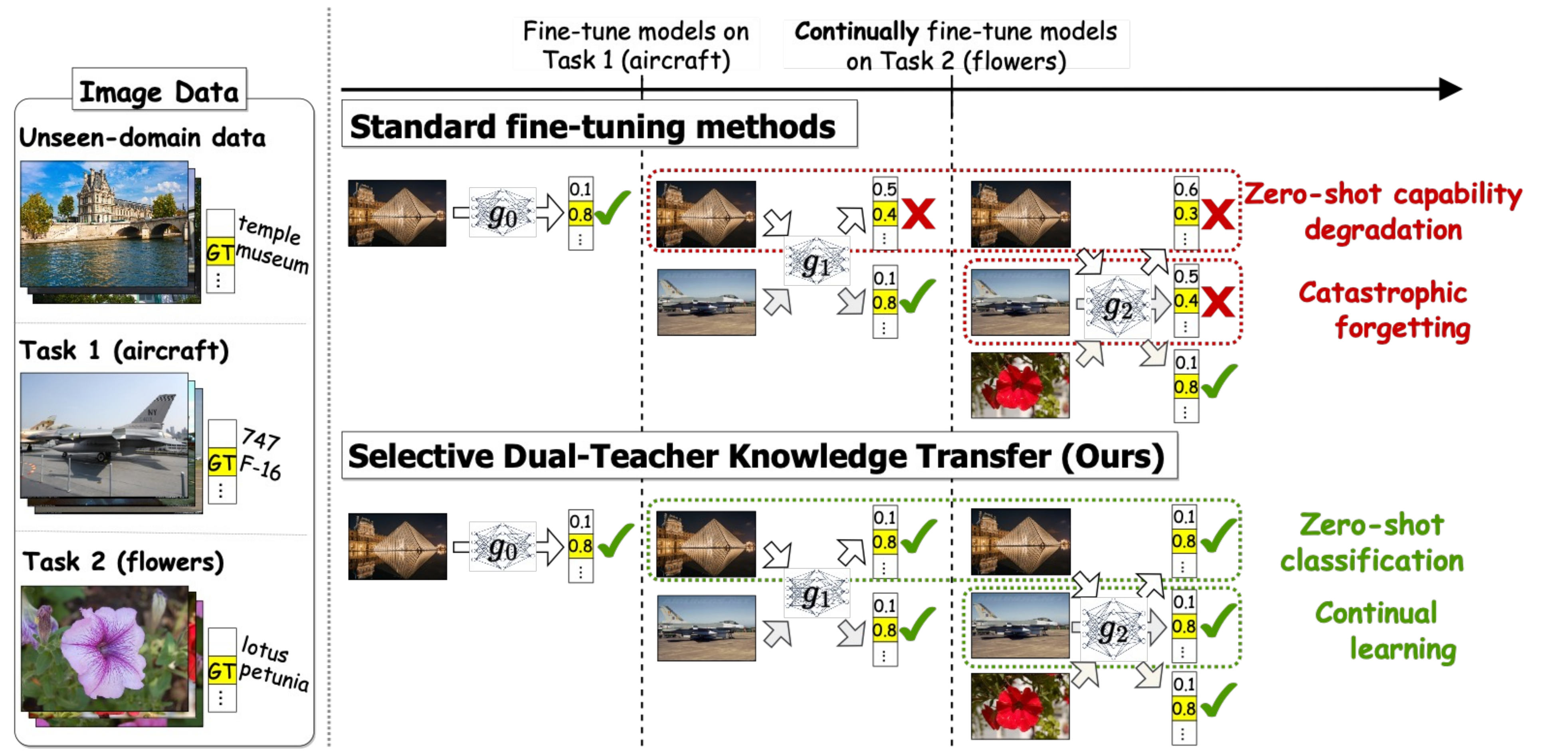
Select and Distill: Selective Dual-Teacher Knowledge Transfer for Continual Learning on Vision-Language Models

Yu-Chu Yu[1] Chi-Pin Huang[1] Jr-Jen Chen[1] Kai-Po Chang[1] Yung-Hsuan Lai[1] Fu-En Yang[2] Yu-Chiang Frank Wang[1,2]

[1] National Taiwan University [2] NVIDIA

TL;DR: We propose Select & Distill, preventing Catastrophic Forgetting and preserving Zero-Shot Transferability for Continual Learning on VLMs.

Introduction

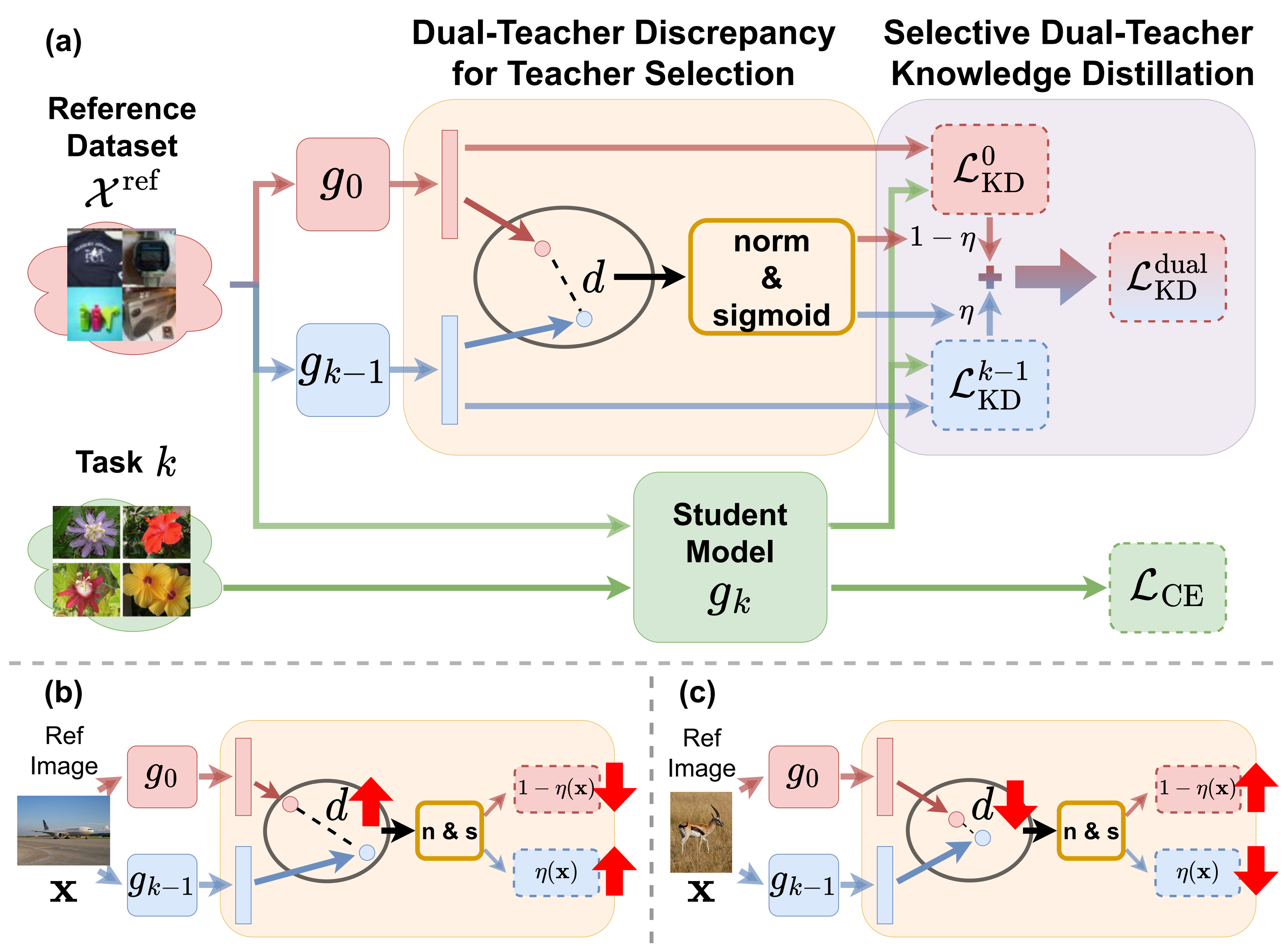


- Continual Learning for VLMs poses 2 challenges: **preventing Catastrophic Forgetting** & **preserving Zero-Shot Transferability**.
- We propose **Select and Distill**, a Selective Dual-Teacher Knowledge Transfer mechanism to address both issues.

Contributions

- Without accessing previous data (neither image nor labels!), we preserve **previous knowledge** & **zero-shot transferability** of VLMs during CL.
- The model is fine-tuned **without requiring any additional memory** to preserve previously learned knowledge & zero-shot knowledge.
- Extensive experiments on **multiple different training orders** demonstrate the state-of-the-art **stability** and **reliability** of our proposed framework.

Method

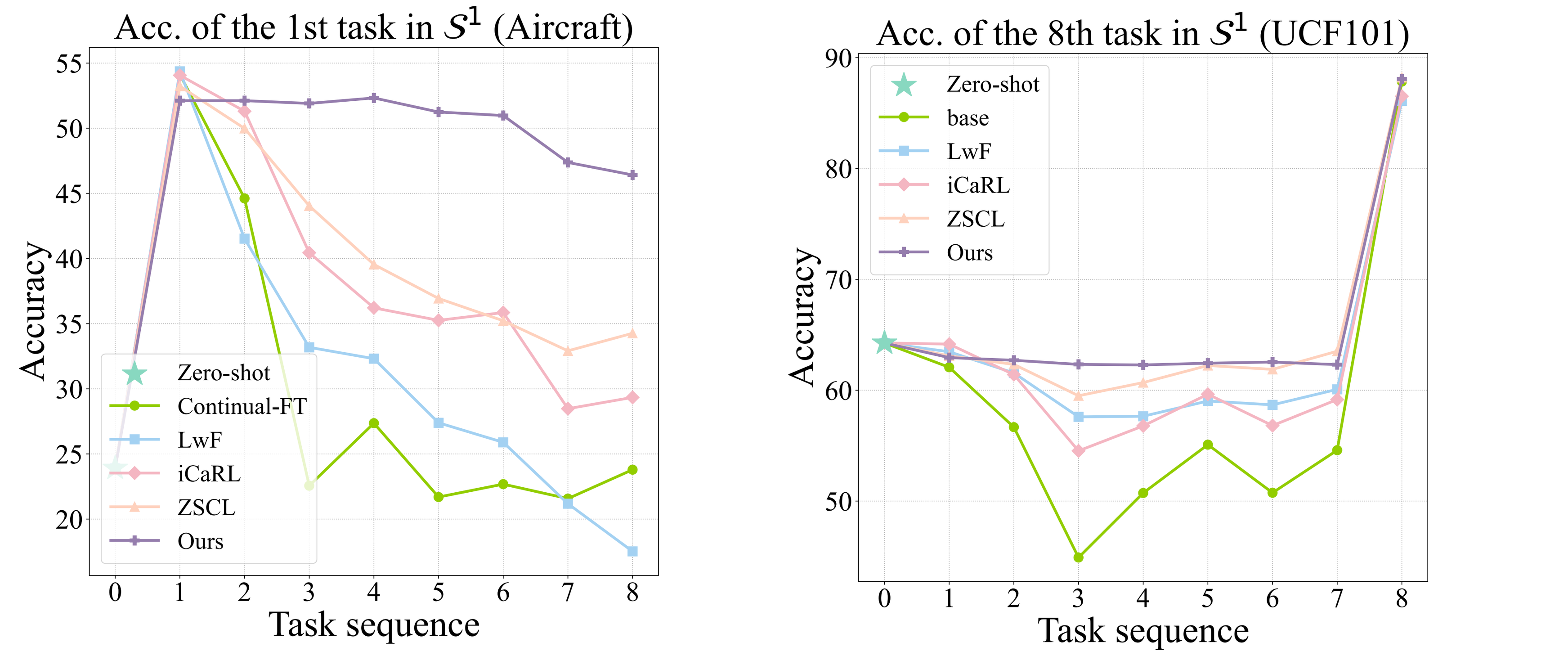


- For data that has been learned before, the feature distance between g_0 & g_{k-1} tends to be large, and we should distill knowledge from g_{k-1}.
- Conversely, for unseen data, the feature distance is typically small, and we should instead distill knowledge from g_0.

Ablation Study

Method	Forgetting (↓)	Degradation (↓)	Avg. Accuracy (↑)
Distill from g_0	5.26	2.51	81.35
Distill from g_{k-1}	2.63	3.36	83.61
Ours	**1.70**	**1.55**	**84.48**

Experiments



- After 8 rounds of Continual Learning, the performance drop of 1st task remains under 5%.
- The zero-shot performance for 8th task can also be properly maintained.

Method / Sequence	$\mathcal{S}^1$	$\mathcal{S}^2$	$\mathcal{S}^3$	$\mathcal{S}^4$	$\mathcal{S}^5$	$\mathcal{S}^6$	$\mathcal{S}^7$	$\mathcal{S}^8$	**Mean**
Average accuracy (↑)									
Continual FT	76.16	76.24	78.03	68.69	76.64	75.44	72.71	77.45	75.17
LwF	76.78	80.45	80.65	77.52	79.64	79.45	77.31	78.70	78.81
iCaRL	77.99	79.77	79.93	76.66	79.26	79.08	77.06	78.61	78.55
ZSCL	81.89	83.98	84.30	83.49	83.41	82.38	81.92	81.97	82.92
MoE-Adapters	82.71	80.74	81.15	83.97	83.68	83.68	82.73	79.68	82.29
Ours	**84.48**	**84.92**	**84.97**	**84.89**	**85.50**	**85.07**	**85.02**	**84.52**	**84.92**

- Stable performance across different training order sequences.

Further Information

Check our project page for detailed explanation!

Feel free to contact me through my personal page!