InfoCLIP: Bridging Vision-Language Pretraining and Open-Vocabulary Semantic Segmentation via Information-Theoretic Alignment Transfer

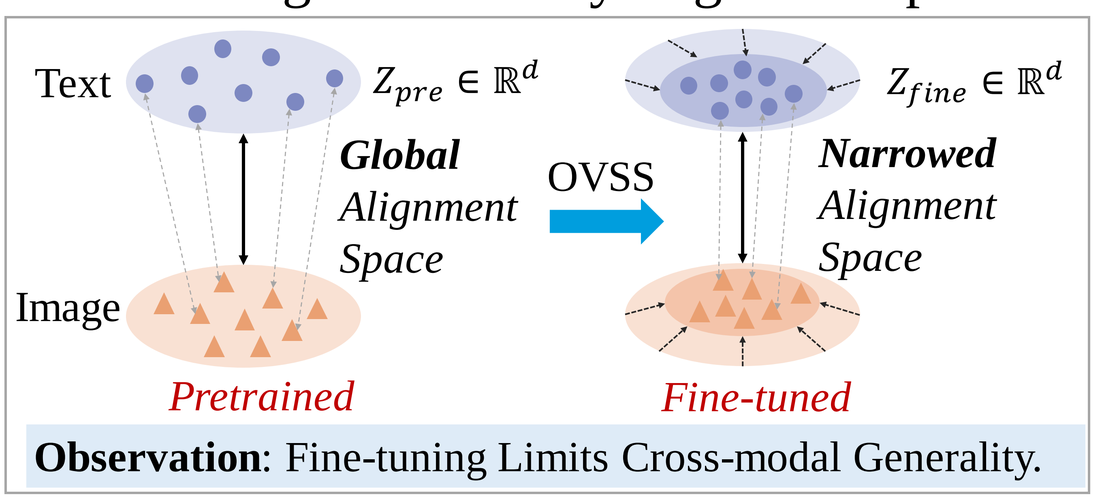
Project Page

Muyao Yuan[1], Yuanhong Zhang[1], Weizhan Zhang[†1], Lan Ma[2], Yuan Gao[2], Jiangyong Ying[2], Yudeng Xin[3]

[1] Xi'an Jiaotong University [2] China Telecom [3] University of Melbourne

Introduction

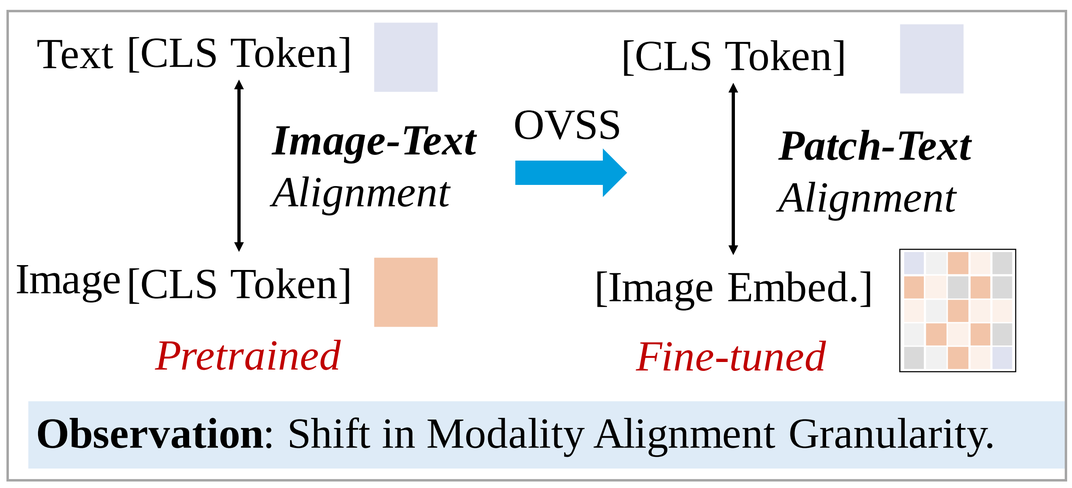
Fine-tuning for open-vocabulary segmentation on a limited set of categories can hurt generalization by constraining the modality alignment space.



Observation: Fine-tuning Limits Cross-modal Generality.

? ***Can we extract the modality alignment information preserved in the pre-trained CLIP to enhance the PEFT process?***

Pretrained CLIP captures global image-text alignment, but segmentation requires precise pixel-level alignment.



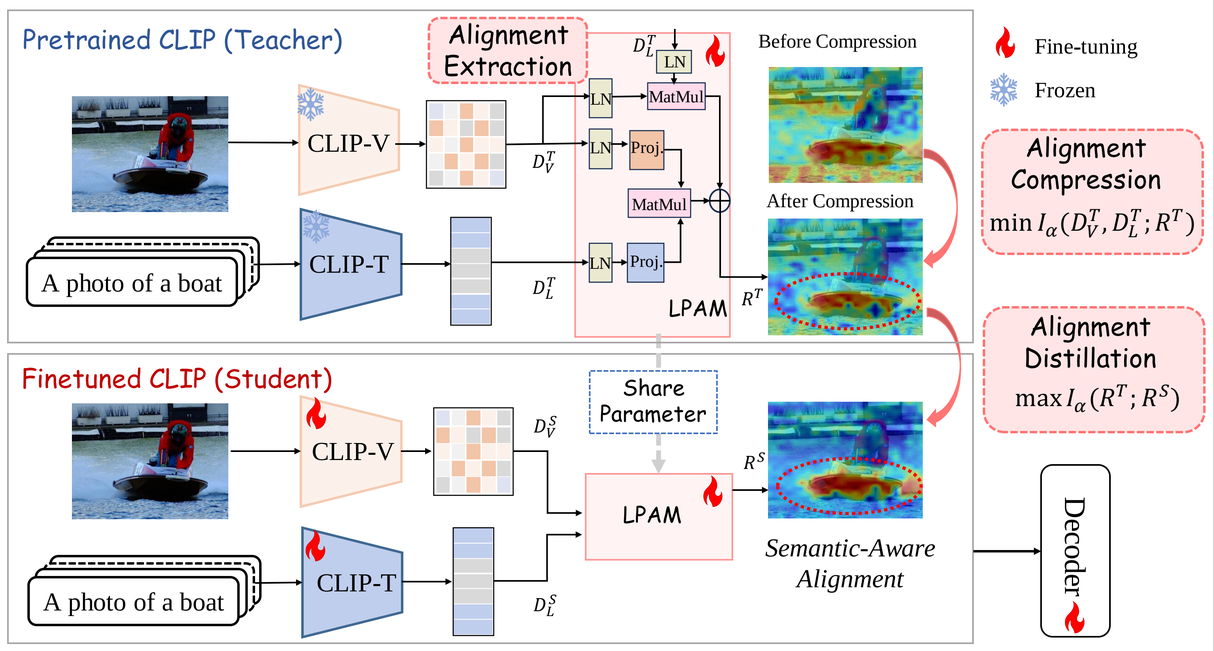
Observation: Shift in Modality Alignment Granularity.

Challenges:

- How to extract refined pixel-level alignment relations from noisy pretrained representations?
- How to effectively transfer them to guide fine-tuning while preserving modality alignment?

Methodology

■ Overview of InfoCLIP



InfoCLIP introduces an asymmetric adaptation framework: (1) LPAM for fine-grained patch-text alignment, (2) an information bottleneck to reduce noise, and (3) mutual information transfer to preserve modality alignment.

■ Formulation

- Let D_V^T / D_V^S and D_L^T / D_L^S denote the image embeddings and text embeddings from the teacher (pre-trained CLIP) and student (fine-tuned CLIP), respectively.
- To suppress noise and retain semantic-aware alignmen R^T:

$$\min I_\alpha(D_V^T, D_L^T; R^T) \Rightarrow \mathcal{L}_c = -log_2\|G_R^T\|_F^2 + log_2\|G_{VLR}^T\|_F^2$$

- To maximize the extracted information:

$$\max I_\alpha(R^T; R^S) \Rightarrow \mathcal{L}_d = log_2\|G_R^T\|_F^2 + log_2\|G_R^S\|_F^2 - log_2\|G_R^{TS}\|_F^2$$

- **The overall objectives:**

$$\mathcal{L}_{overall} = \mathcal{L}_{task} + \lambda_1\mathcal{L}_c + \lambda_2\mathcal{L}_d$$

Experiments

Model	VLM	Add. Backbone	Training Dataset	A-847	PC-459	A-150	PC-59	PAS-20	PAS-20^b
Without Distillation									
LSeg (2022a)	CLIP ViT-B/32	ViT-L/16	PASCAL VOC-15	-	-	-	-	52.3	-
OpenSeg (2022)	ALIGN	Eff-B7	COCO Panoptic	8.1	11.5	26.4	44.8	-	70.2
OVSeg (2023)	CLIP ViT-L/14	Swin-B	COCO-Stuff	9.0	12.4	29.6	55.7	94.5	-
SAN (2023b)	CLIP ViT-L/14	-	COCO-Stuff	12.4	15.7	32.1	57.7	94.6	-
ODISE (2023a)	CLIP ViT-L/14	Stable Diffusion	COCO-Stuff	11.1	14.5	29.9	57.3	-	-
SED (2024)	CLIP ConvNeXt-L	-	COCO-Stuff	13.9	22.6	35.2	60.6	96.1	-
FC-CLIP (2023)	CLIP ConvNeXt-L	-	COCO Panoptic	14.8	18.2	34.1	58.4	95.4	-
CAT-Seg (2024)	CLIP ViT-L/14	-	COCO-Stuff	16.0	23.8	37.9	63.3	97.0	82.5
With Distillation									
MAFT (2024)	CLIP ViT-L/14	Mask2Former	COCO-Stuff	12.7	16.2	33.0	59.0	92.1	-
MAFT (2024)	CLIP ConvNeXt-L	Mask2Former	COCO-Stuff	13.1	17.0	34.4	57.5	93.0	-
MAFT+ (2024)	CLIP ConvNeXt-L	Mask2Former	COCO-Stuff	15.1	21.6	36.1	59.4	96.5	-
InfoCLIP (ours)	CLIP ViT-L/14	-	COCO-Stuff	**16.6**	**24.6**	**38.5**	**63.5**	**97.5**	**83.1**

- **Alignment Compression**