DeCLIP: Decoupled Learning for Open-Vocabulary Dense Perception

Junjie Wang[1], Bin Chen[2,3,*], Yulin Li[1], Bin Kang[3] ,
Yichi Chen[3] , Zhuotao Tian[1*]

[1] HIT, Shenzhen
[2] University of Chinese Academy of Sciences

Why do we propose DeCLIP?

Training Set

Closed-Set Dense Prediction

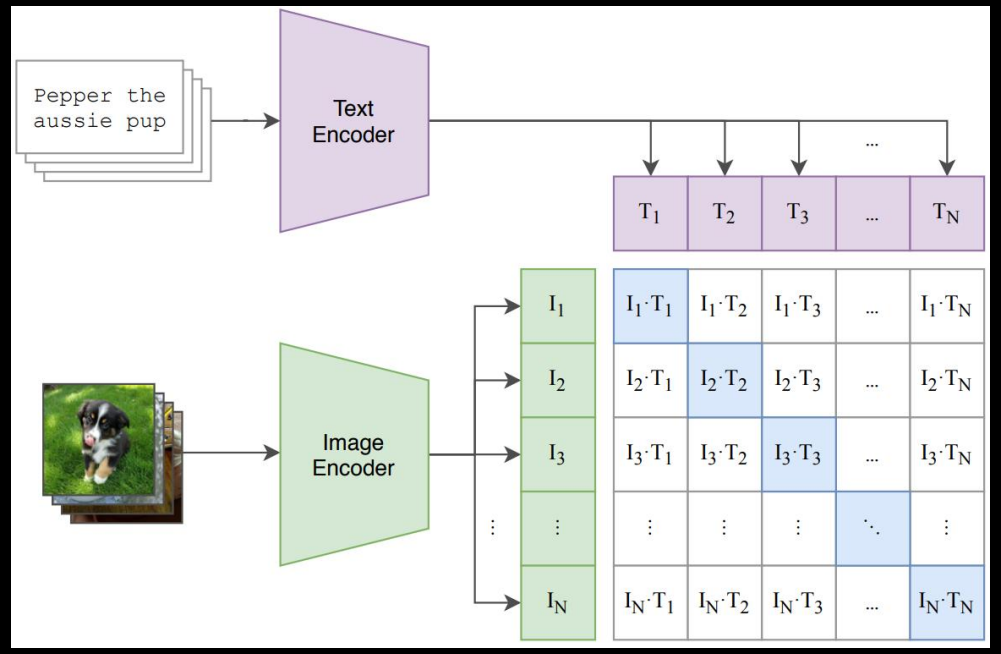
Test Set (Predefined Categories)

Training Set

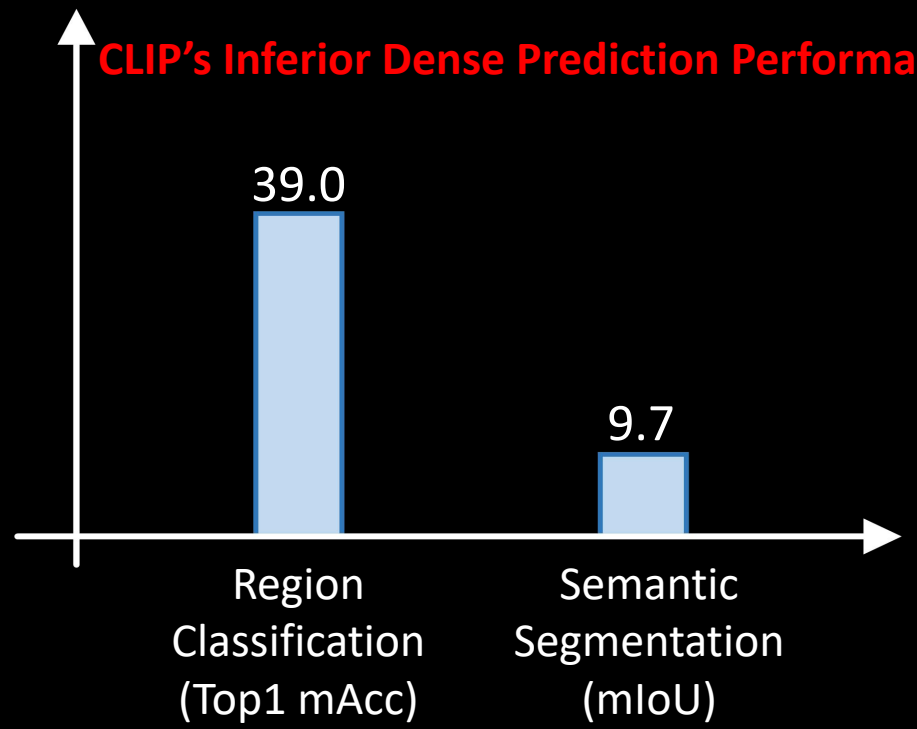
Open-Vocabulary Dense Prediction

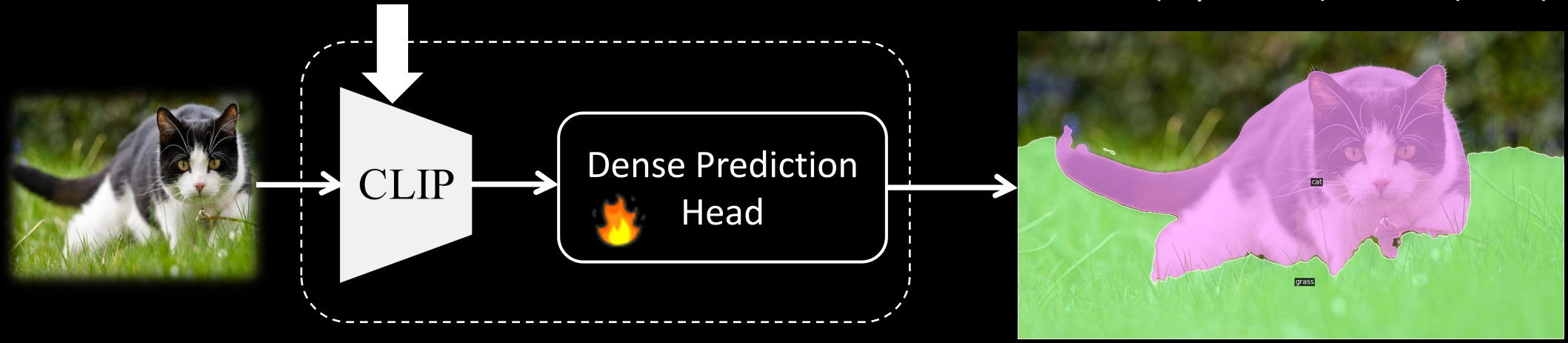
Test Set (Any Category)

Why do we propose DeCLIP?

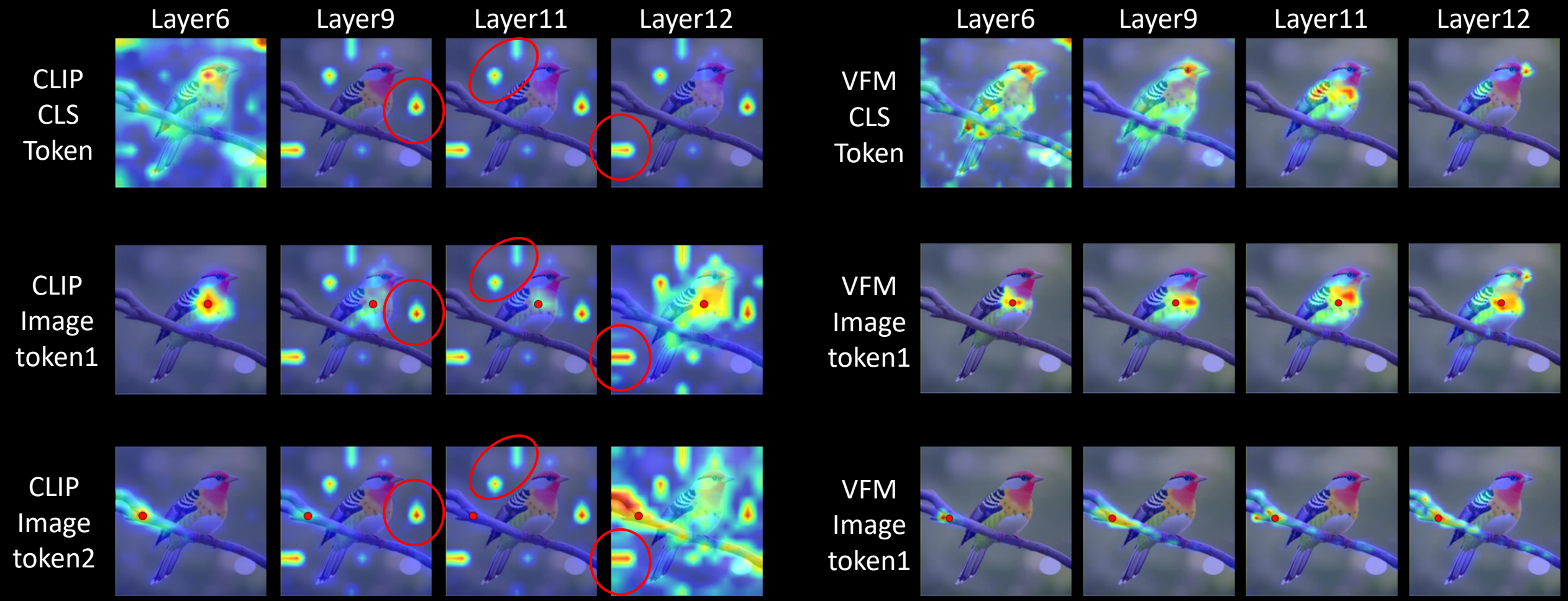


Contrastive Language-Image Pre-training (CLIP)

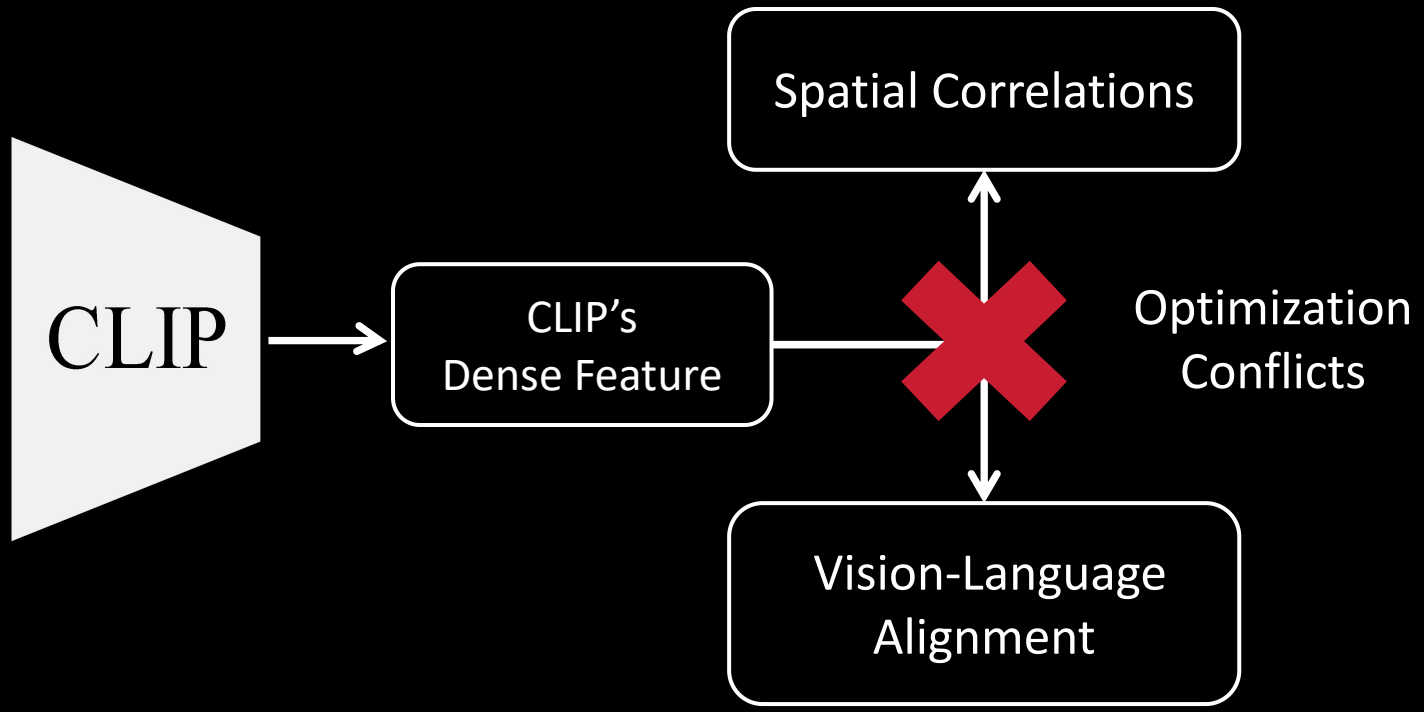




Why do we propose DeCLIP?

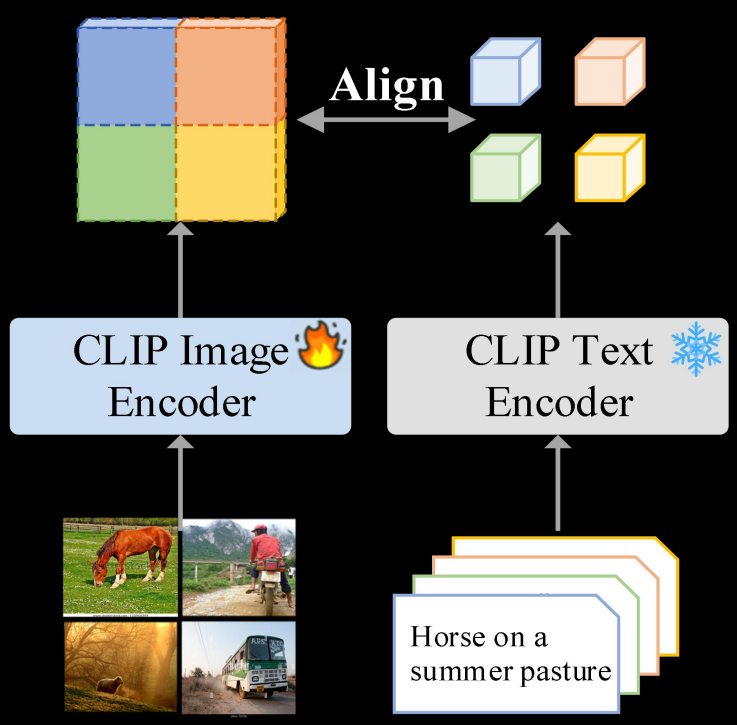


Why do we propose DeCLIP?

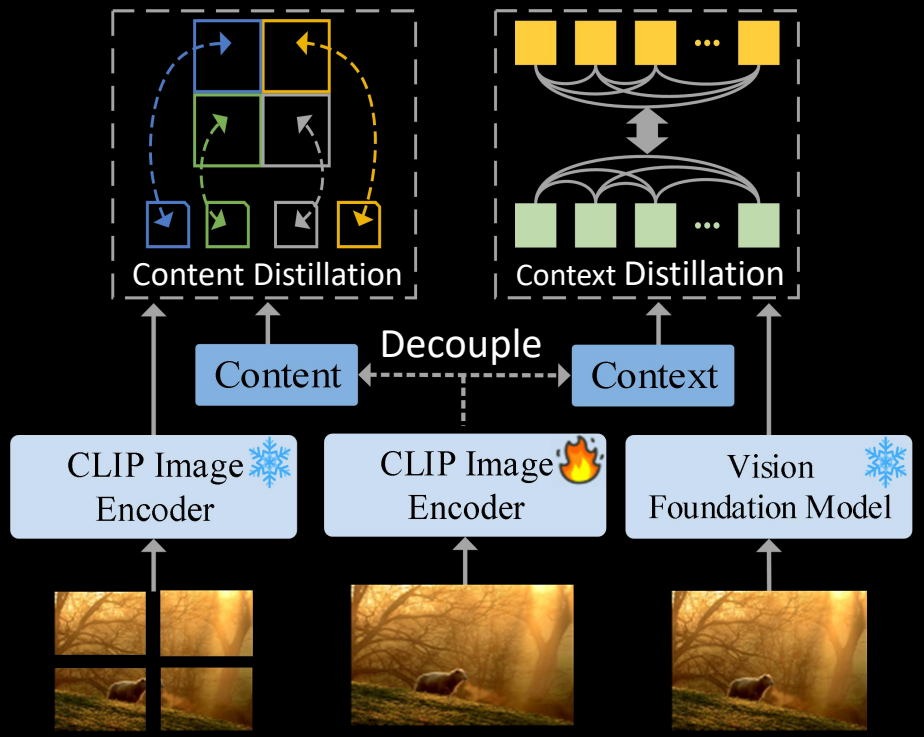


Distillation Type	Region Classification (mAcc)		Semantic Segmentation (mIoU)	
	COCO (Thing)	COCO (Stuff)	Context59	CityScape
Self Distillation [48]	69.5	44.6	29.4	25.6
Self+VFM Distillation [22]	65.6 (-3.9)	41.3 (-3.3)	32.4 (+3.0)	28.7 (+3.1)
Self+VFM+Decouple	75.0 (+5.5)	51.8 (+7.2)	35.3 (+5.9)	32.3 (+6.7)

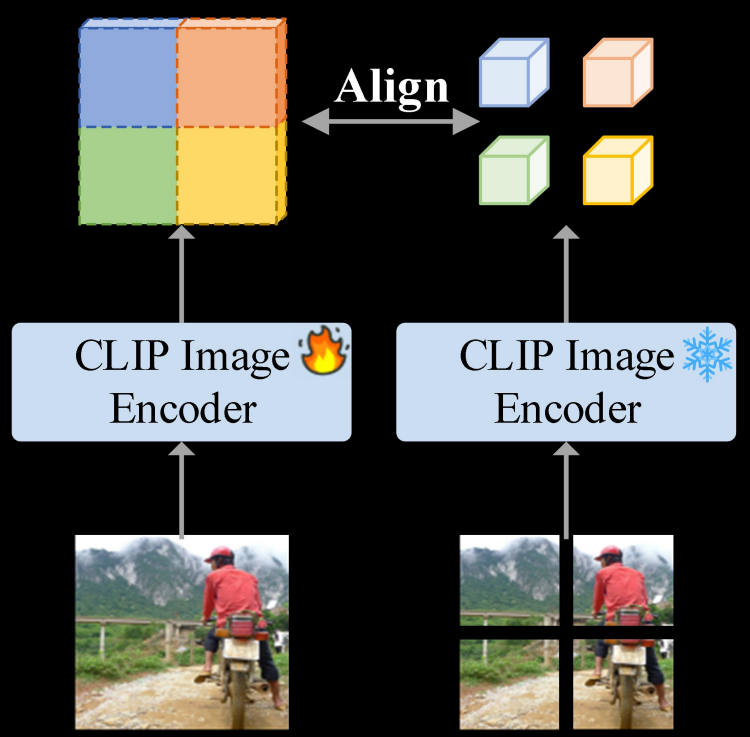
Why do we propose DeCLIP?



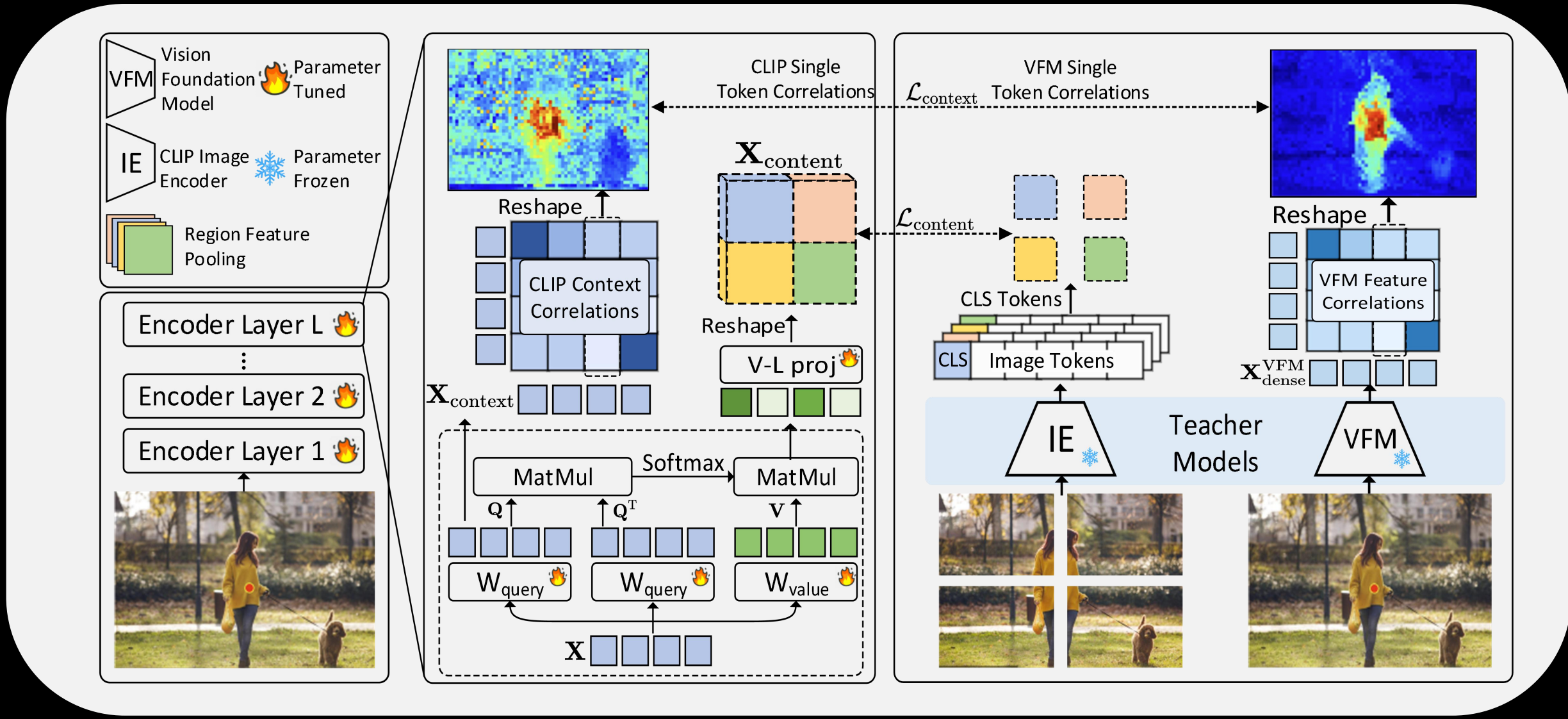
Using Pseudo Regions



Decouple Distillation (ours)



Using Self-Distillation


VFM
Vision Foundation Model
Parameter Tuned
IE
CLIP Image Encoder
Parameter Frozen
Region Feature Pooling
Encoder Layer L
Encoder Layer 2
Encoder Layer 1
CLIP Single Token Correlations
$\mathcal{L}_{\text{context}}$
VFM Single Token Correlations
$\mathbf{X}_{\text{content}}$
$\mathcal{L}_{\text{content}}$
Reshape
CLIP Context Correlations
Reshape
V-L proj
$\mathbf{X}_{\text{context}}$
CLS Tokens
CLS
Image Tokens
Reshape
VFM Feature Correlations
$\mathbf{X}_{\text{dense}}^{\text{VFM}}$
MatMul
Softmax
MatMul
Q
$\mathbf{Q}^{\text{T}}$
V
W_{query}
W_{query}
W_{value}
X
IE
Teacher Models
VFM

Experiment Results

Open-Vocabulary Object Detection

(a) OV-COCO benchmark

Method	Supervision	Backbone	AP_{50}^{Novel}
ViLD [14]	CLIP	RN50	27.6
Detic [64]	Caption	RN50	27.8
OV-DETR† [59]	CLIP	RN50	29.4
BARON-KD [47]	CLIP	RN50	34.0
SAS-Det [61]	CLIP	RN50	37.4
OV-DQUO† [46]	CLIP	RN50	39.2
RegionCLIP [62]	Captions	RN50x4	39.3
CORA† [50]	CLIP	RN50x4	41.7
OV-DQUO† [46]	CLIP	RN50x4	45.6
RO-ViT [20]	CLIP	ViT-L/16	33.0
CFM-ViT [19]	CLIP	ViT-L/16	34.1
F-ViT [48]	CLIP	ViT-B/16	37.6
F-ViT [48]	CLIP	ViT-L/14	44.3
F-ViT+DeCLIP	CLIP	ViT-B/16	41.1 (+3.5)
F-ViT+DeCLIP	CLIP	ViT-L/14	46.2 (+1.9)
OV-DQUO+DeCLIP†	CLIP	ViT-B/16	**46.1** (+6.9)
OV-DQUO+DeCLIP†	CLIP	ViT-L/14	**48.3** (+2.7)

(b) OV-LVIS benchmark

Method	Supervision	Backbone	mAP_r
ViLD [14]	CLIP	RN50	16.3
OV-DETR† [59]	CLIP	RN50	17.4
BARON-KD [47]	CLIP	RN50	22.6
RegionCLIP [62]	Caption	RN50x4	22.0
CORA+† [50]	Caption	RN50x4	28.1
F-VLM [23]	CLIP	RN50x64	32.8
F-ViT [48]	CLIP	ViT-B/16	25.3
OV-DQUO† [46]	CLIP	ViT-B/16	29.7
Detic [64]	Caption	Swin-B	33.8
RO-ViT [20]	CLIP	ViT-H/16	34.1
F-ViT [48]	CLIP	ViT-L/14	34.9
CoDet [32]	Caption	ViT-L/14	37.0
OV-DQUO† [46]	CLIP	ViT-L/14	39.3
F-ViT+DeCLIP	CLIP	ViT-B/16	26.8 (+1.5)
F-ViT+DeCLIP	CLIP	ViT-L/14	37.2 (+2.3)
OV-DQUO+DeCLIP†	CLIP	ViT-B/16	**31.0** (+1.3)
OV-DQUO+DeCLIP†	CLIP	ViT-L/14	**41.5** (+2.2)

Method	COCO			Objects365 [40]		
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
Supervised Baseline [14]	46.5	67.6	50.9	25.6	38.6	28.0
ViLD [14]	36.6	55.6	39.6	11.8	18.0	12.6
DetPro [13]	34.9	53.8	37.4	12.1	18.8	12.9
BARON [47]	36.2	55.7	39.1	13.6	21.0	14.5
F-VLM [23]	37.9	61.6	41.2	16.2	27.4	17.5
CoDet [32]	39.1	57.0	42.3	14.2	20.5	15.3
RO-ViT [21]	-	-	-	17.7	27.4	19.1
CLIPSelf [48]	40.5	63.8	44.3	19.5	31.3	20.7
DeCLIP	**41.0**	**64.6**	**44.8**	**20.0**	**32.2**	**21.2**

Experiment Results

Open-Vocabulary Semantic Segmentation

Method	Backbone	Training Set	ADE847	Context459	ADE150	Context59	VOC20	VOC21
ZegFormer† [11]	ViT-B/16	COCO-Stuff	5.6	10.4	18.0	45.5	89.5	65.5
ZSseg [55]	ViT-B/16	COCO-Stuff	7.0	-	20.5	47.7	88.4	-
OVSeg [28]	ViT-L/14	COCO-Stuff	9.0	12.4	29.6	55.7	94.5	-
SAN [56]	ViT-L/14	COCO-Stuff	13.7	17.1	33.3	60.2	95.5	-
ODISE [54]	ViT-L/14	COCO-Panoptic	11.1	14.5	29.9	57.3	-	84.6
MAFT [17]	ConvNeXt-L	COCO-Stuff	13.1	17.0	34.4	57.5	93.0	-
FC-CLIP [58]	ConvNeXt-L	COCO-Panoptic	14.8	18.2	34.1	58.4	95.4	81.8
FrozenSeg [7]	ConvNeXt-L	COCO-Panoptic	14.8	19.7	34.4	-	-	82.5
CAT-Seg [10]	ViT-B/16	COCO-Stuff	12.0	19.0	31.8	57.5	94.6	77.3
CAT-Seg [10]	ViT-L/14	COCO-Stuff	16.0	23.8	37.9	63.3	97.0	82.5
CAT-Seg+DeCLIP	ViT-B/16	COCO-Stuff	15.3 (+3.3)	21.4 (+2.4)	36.3 (+4.5)	60.6 (+3.1)	96.6 (+2.0)	81.3 (+4.0)
CAT-Seg+DeCLIP	ViT-L/14	COCO-Stuff	**17.6** (+1.6)	**25.9** (+2.1)	**40.7** (+2.8)	**63.9** (+0.6)	**97.7** (+0.7)	**83.9** (+1.4)

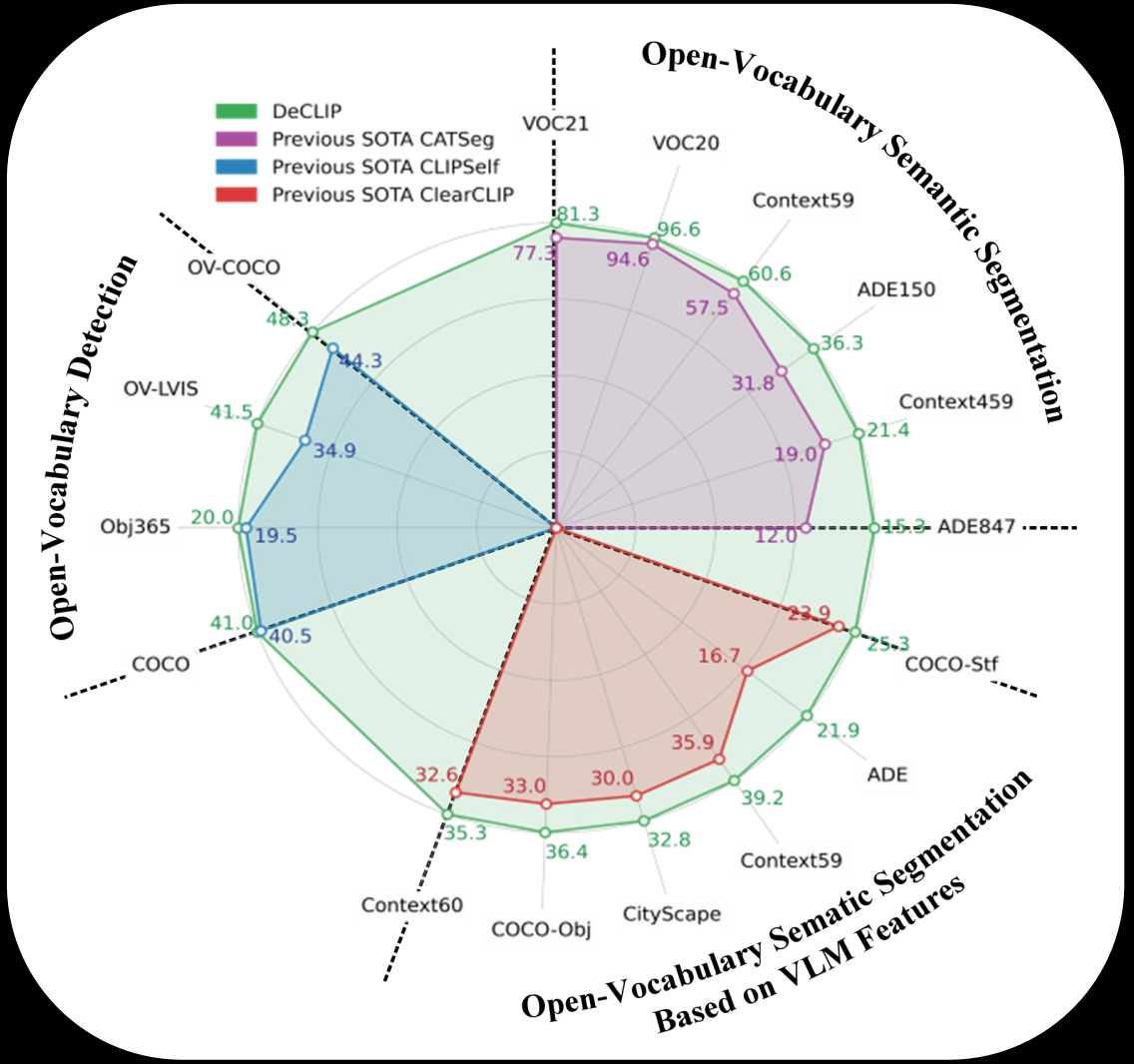
Experiment Results

Open-vocabulary Semantic Segmentation Base on VLM features

Method	With a background category			Without background category					Avg
	VOC21	Context60	COCO-Obj	VOC20	CityScape	Context59	ADE	COCO-Stf	
CLIP [36]	18.8	9.9	8.1	49.4	6.5	11.1	3.1	5.7	14.1
MaskCLIP [63]	43.4	23.2	20.6	74.9	24.9	26.4	11.9	16.7	30.3
GroupViT [53]	52.3	18.7	27.5	79.7	18.5	23.4	10.4	15.3	30.7
ReCo [42]	25.1	19.9	15.7	57.7	21.6	22.3	11.2	14.8	23.5
TCL [3]	51.2	24.3	30.4	77.5	23.5	30.3	14.9	19.6	33.9
OVSeg [28]	53.8	20.4	25.1	-	-	-	5.6	-	-
SCLIP [45]	59.1	30.4	30.5	80.4	32.2	34.2	16.1	22.4	38.2
ClearCLIP [24]	51.8	32.6	33.0	80.9	30.0	35.9	16.7	23.9	38.1
CLIP-DINOiser [51]	**62.1**	32.4	34.8	80.9	31.7	35.9	20.0	24.6	40.3
DeCLIP (ours)	59.7	**35.3**	**36.4**	**85.0**	**32.8**	**39.2**	**21.9**	**25.3**	**41.9**

Experiment Results

Summary of Our Proposed DeCLIP in Dense Prediction



Thanks for your listening

All code and checkpoints will be released after the review process