Emergent Open-Vocabulary Semantic Segmentation from Off-the-shelf Vision-Language Models

Jiayun Luo, Siddhesh Khandelwal, Leonid Sigal, and Boyang Li

{luoj0028, boyang.li}@ntu.edu.sg, {skhandel, lsigal}@cs.ubc.ca

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE · UBC

CVPR JUNE 17-21, 2024 SEATTLE, WA

Paper · Repo

Motivation

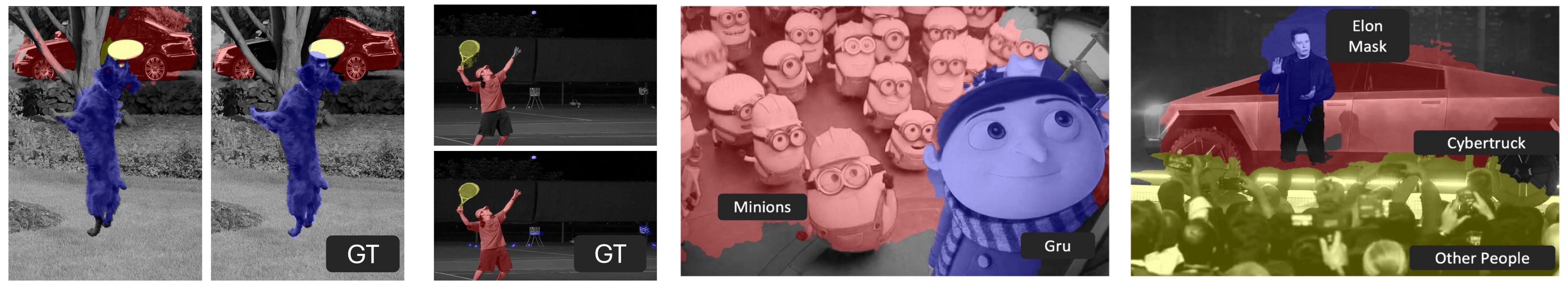
Q: What is the white object on the grey table?

BLIP: Paper.

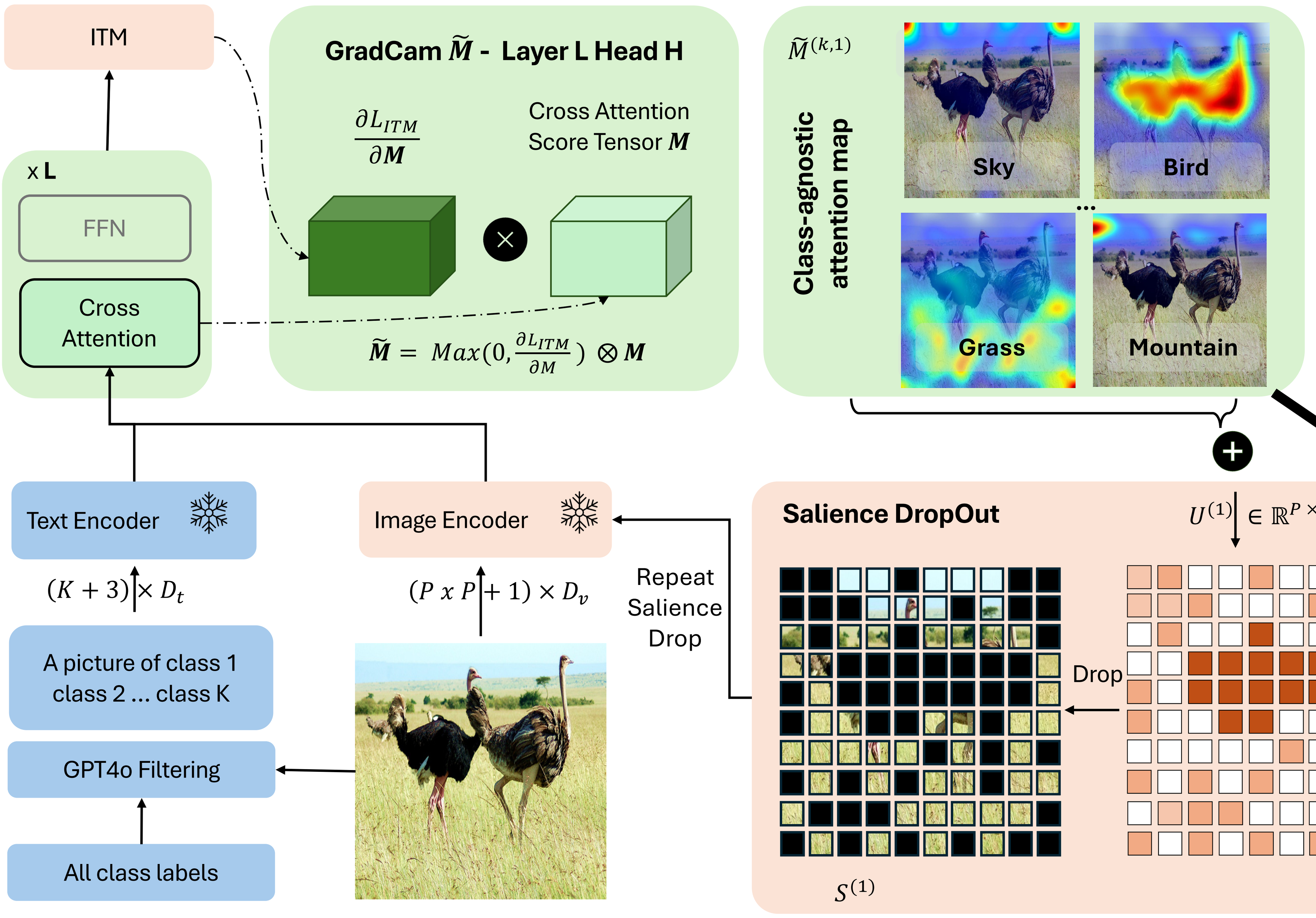
Vision Language Models (VLMs) can identify objects of interest, but isolating this localization capability to create usable segmentation masks without continuously posing questions is unexplored.

Main Contributions

- We propose to combine **text-to-image attention, GradCAM, and Salience DropOut** to iteratively acquire accurate segmentation of arbitrary classes from **pretrained VLM.**
- We replace the densely annotated validation set for hyperparameter tuning, which is needed by most existing methods, with a **contrastive reward function based on CLIP.** This reward function, coupled with random search, finds a good set of hyperparameters for OVSS.
- The proposed method, PnP-OVSS, is simple to use, **requires no extra finetuning**, and delivers high performance. Its success hints at a new direction for open-vocabulary segmentation tasks leveraging large VLMs.



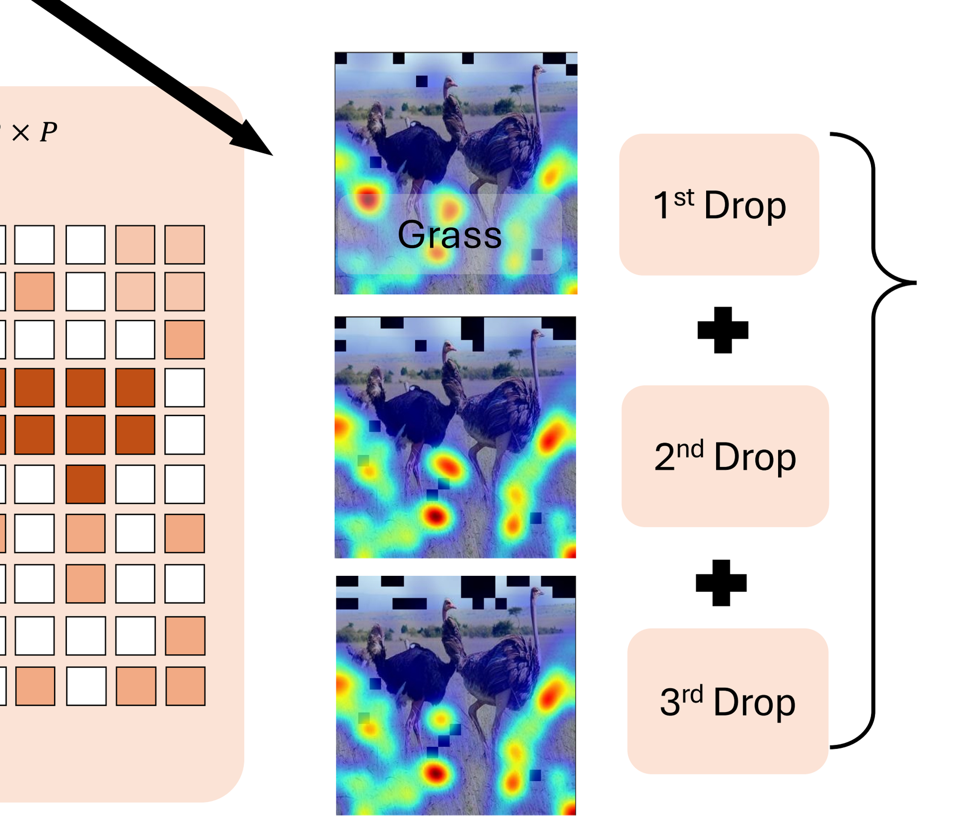
Method

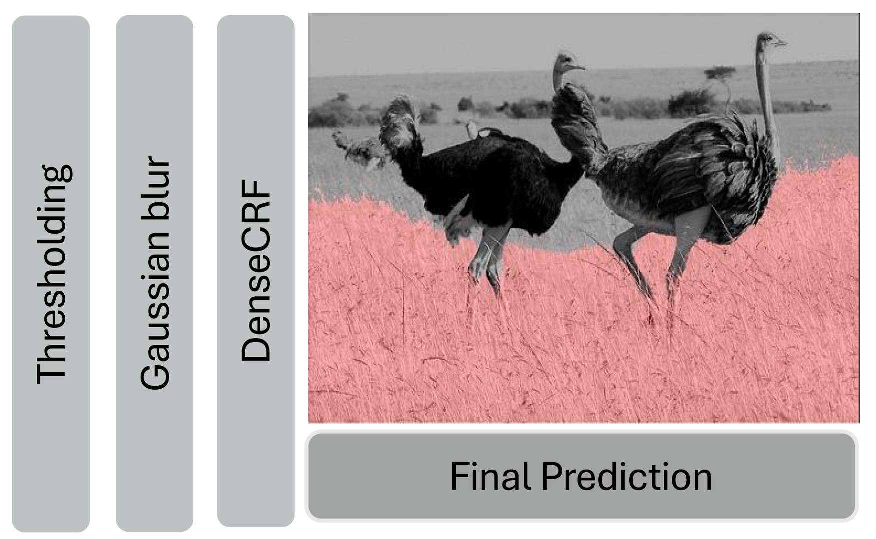


PnP-OVSS has five major steps.

1. We prompt GPT4o to output the categories that appear in each image.
2. We extract a cross-attention salience map for each predicted category from a VLM.
3. We sharpen the salience map by weighing it with the image-text matching gradient in the style of GradCAM. This highlights the salient region of the objects but results in incomplete masks.
4. We apply Salience DropOut which iteratively drops salient image patches to move the attention to previously unattended object areas. For each drop iteration, we black out N remaining image patches with the highest aggregated attention value, run step 2 and 3 again, and sum up the attention maps from all iterations. Salience DropOut helps complete the salience maps.
5. We apply Thresholding, Gaussian Blur, and Dense CRF for fine-grained adjustment.

* Layer L Head H and Threshold T are automatically tuned with our weakly supervised reward function based on CLIP.





$$S^{(t)} = S^{(t-1)} \setminus \{(i,j) \mid U_{ij}^{(t)} \in A\} \quad (2)$$

$$A = TOP10\left(\left\{U_{ij}^{(t)} \middle| (i,j) \in S^{(t-1)}\right\}\right) \quad (3)$$

Experiments

Method	Finetuning VLMs	HT on Dense Labels	Short-side Resolution	Pascal VOC-20	Pascal Context-59	COCO Object-80	COCO Stuff-171	ADE 20K-150
Group 1: Methods that require weakly supervised finetuning on image-text data								
ViL-Seg† [43]	✓	✓	-	37.3	18.9	-	18.0	-
CLIPpy [51]	✓	✓	224	52.2	-	**32.0**	25.5★	13.5
SegClip [44]	✓	✓	224	52.6	24.7	26.5	-	-
GroupVit (by [51])	✓	✓	224	28.1	14.8	12.9	-	6.2
GroupVit (by [6])	✓	✓	448	50.4	18.7	27.5	15.3	9.2
GroupVit [74]	✓	✓	448	52.3	22.4	24.3	-	-
ViewCo [52]	✓	✓	448	52.4	23.0	23.5	-	-
OVSegmentor [75]	✓	✓	448	53.8	20.4	25.1	-	5.6
TCL [6] +PAMR [2]	✓	✓	448	55.0	30.4	31.6	22.4	17.1
PACL [46]	✓	✓	224×4	**72.3**	**50.1**	-	**38.8**	**31.4**
Group 2: Methods that require finetuning but not real image-text data								
MaskClip w/ ST [90]	✓	✓	336	-	**31.1**	-	**18.0**	-
MaskClip w/ ST (by [6])	✓	✓	448	38.8	23.6	20.6	16.4	9.8
ZeroSeg* [8]	✓	✓	448	37.3	19.7	17.8	-	
ZeroSeg [8]	✓	✓	448	**40.8**	20.4	**20.2**	-	-
Group 3: Methods that require no finetuning								
MaskClip (by [51])	×	×	224	22.1	-	13.8	8.1	6.8
MaskClip [90]	×	×	336	-	25.5	-	14.6	-
Reco [58]	×	×	320	-	**27.2**	-	**27.2★**	-
Reco (by [6])	×	×	448	**25.1**	19.9	**15.7**	14.8	11.2
PnP-OVSS with different VLMs								
$\text{BLIP}_{\text{Flickr}}$	×	×	224	46.3	27.5	32.3	17.3	13.6
$\text{BLIP}_{\text{Flickr}}$	×	×	336	**51.3**	**28.0**	**36.2**	**17.9**	**14.2**
BridgeTower	×	×	336	42.4	25.3	30.4	15.7	14.8

Table 3. Zero-shot semantic segmentation performance in mIoU. Group 3 contains the most similar baselines that serve as fair comparisons to PnP-OVSS. Groups 1 and 2 benefit from additional training, extra image-text data, and hyperparameter tuning on dense labels. We use the word "by" followed by a paper citation to indicate results of the same technique reported by different papers. ★ CLIPpy tests on 133 categories of COCO Stuff and Reco tests on 27 super categories while we test all 171 classes of COCO Stuff. ViL-Seg† is tested on subset of classes on the three datasets, as detailed in the supplementary.

Takeaway: PnP-OVSS outperform comparable training-free methods in Group 3 with +26.2% mIoU on Pascal VOC, +20.5% mIoU on MS COCO, +3.1% mIoU on COCO Stuff and +3.0% mIoU on ADE20K

Automatic Hyperparameter Tuning

Hyperparameters	Start	End	Step	Solution
BLIP				
Layer	1	12	1	8
Head	1	12	1	10
Attention Threshold	0.05	0.5	0.1	0.15
BridgeTower				
Layer	1	6	1	2
Head	1	16	1	8
Attention Threshold	0.05	0.5	0.1	0.15
Gaussian Blur				
Standard Deviation	0.01	0.11	0.02	0.05

Table 2. Search space for hyperparameters

$$Reward = \sum_{k \in K(I)} \mathbb{1}\left[\Pr\left(M^{(k)} \otimes I, k\right) > \Pr(\mathbf{0}, k)\right] \quad (4)$$

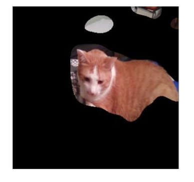
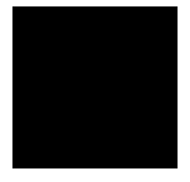
$$\Pr(I,k) = \frac{\exp(f(I,k))}{\sum_{k' \in K(I)} \exp(f(I,k'))} \quad (5)$$

I: *Image* k: *Class of interest* $K(I)$: *All classes in the dataset*

$\mathbf{0}$: *Black Image*

f: *Image and label similarity calculated with CLIP*

Given GT class labels of the image, Reward +1 when

Pr( , cat) > Pr( , cat)

Takeaway: Dense-label-free hyperparameter tuning for segmentation task is possible with our contrastive reward function based on CLIP