NeurIPS 2026
Journal Track

ILLINOIS

TextRegion

Text-Aligned Region Tokens from Frozen Image-Text Models

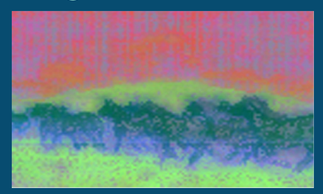
Yao Xiao, Qiqian Fu, Heyi Tao, Yuqun Wu, Zhen Zhu, Derek Hoiem
University of Illinois Urbana-Champaign

How can we extend image-text models to understand local details without training?

Frozen SAM2 masks



Image–text features



Region matching “white horse”



No additional training

Only the final attention layer changes to pool features within SAM2 masks.

Works across model families

CLIP, SigLIP2, and Perception Encoder all produce text-aligned region tokens.

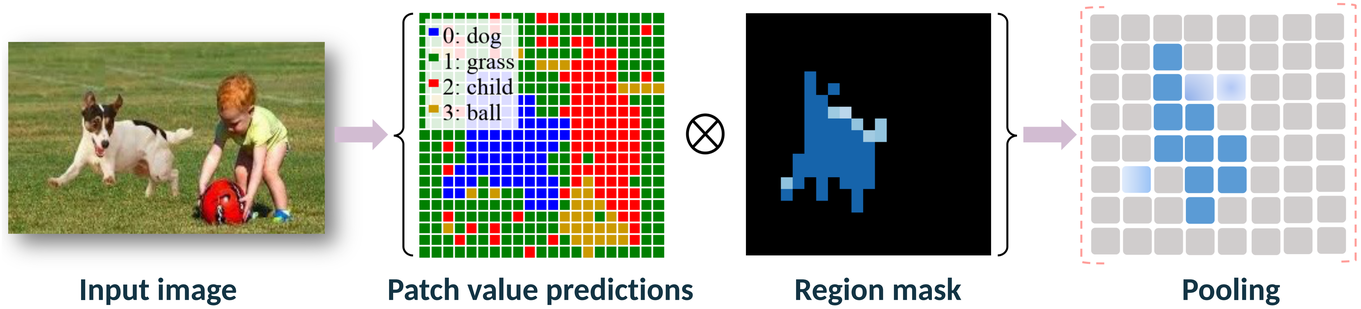
SOTA results on three tasks

Open-world segmentation; Referring expressions; Multiple Object Grounding.

Paper and code

arXiv 2505.23769 github.com/avaxiao/TextRegion

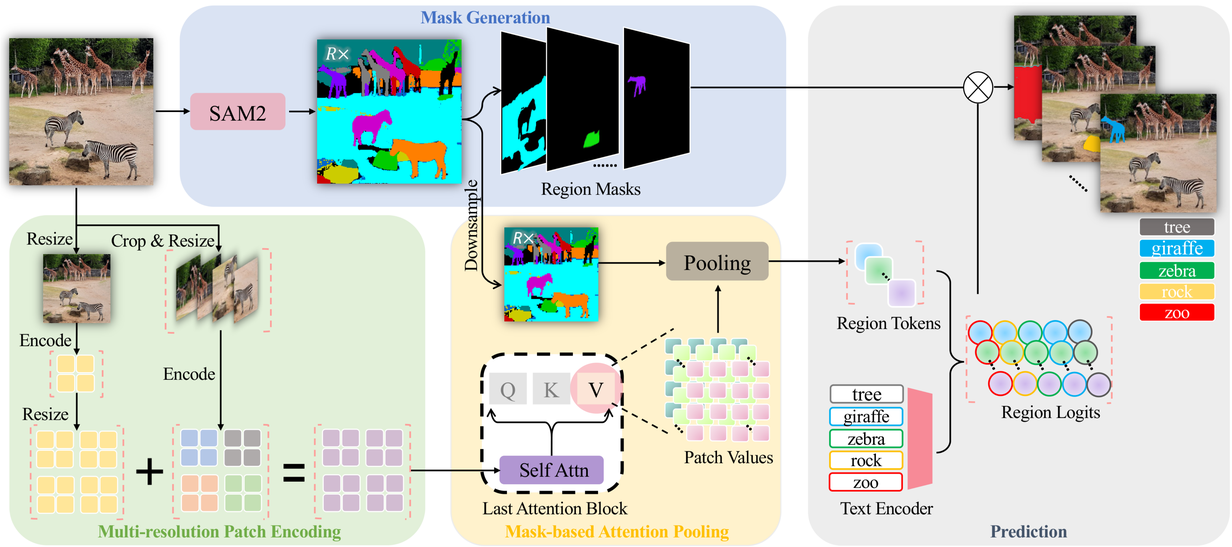
How does a region get a text label?



Input image · Patch value predictions · Region mask · Pooling

Pool patch values inside the SAM2 mask, then match the region token to a text label such as “dog.”

The TextRegion framework



Open-world segmentation

CLIP ViT-B/16 · mIoU (%)

Method	Training-free?	VOC21	COCO-Stuff	ADE20K
SAM-CLIP	✗	60.6	**31.5**	17.1
MaskCLIP	✓	43.4	16.7	11.9
TextRegion	✓	**70.9**	28.7	**22.8**

Referring expressions

“left guy”

Ground truth **TextRegion**



RefCOCO validation · zero-shot accuracy (%)

CLIP baselines		TextRegion		
38.0	46.1	**48.7**	**47.3**	**45.0**
Red Circle	FGVP	CLIP	PE-Core	SigLIP2