PROJECT PAGE

NERVE: Neighbourhood & Entropy-guided Random Walk for training free open-Vocabulary sEgmentation

Kunal Mahatha José Dolz Christian Desrosiers

LIVIA, International Laboratory on Learning Systems (**ILLS**), École de technologie supérieure (**ÉTS**) Montréal

Contents

01 Motivation

Problem:

- Semantic segmentation assumes fixed classes
- The real world has unseen categories

Open-Vocabulary Segmentation helps — but:

❌ Weak spatial coherence
❌ Expensive refinement
❌ Uniform attention fusion

Goal. How can we perform open-vocabulary semantic segmentation without any training data, while preserving spatial coherence and computational efficiency?

02 Key Idea

Treat segmentation as a stochastic random walk on a graph of image tokens

Components:

- **CLIP** → semantic probabilities
- **Diffusion attention** → structural relationships
- **Random walk** → propagation

03 Problem Formulation

Given:

- Token $X = \{x_i\}_{i=1}^{N}$
- Text prompts $\{y_k\}_{k=1}^{K}$

Compute — $P(i,j) = P(token(i) \rightarrow class(k))$

❌ without training

❌ without adaptation

04 Methodology

1 Random Walk Computation

2 Node-to-Node Transition Matrix

3 Uncertainity-weighted Affinities

4 Transition Probability Matrix

4.1 Random Walk Computation

We model the segmentation as a stochastic process on a graph:

- Nodes = Spatial Tokens
- Edges = Affinity between tokens
- Label propagate through transitions

At each step:

- Move with probability α
- Generate label with $1 - \alpha$

Expected label probabilities:

$$P_L = \frac{1-\alpha}{1-\alpha^{L+1}} \left(\sum_{t=0}^{L} \alpha^t S^t \right) G$$

4.2 Node-to-Node Transition Matrix

Global Affinity Matrix

$$[A_{global}]_{i,j} := \frac{\langle Q_i, K_j \rangle}{\|Q_i\| \cdot \|K_j\|}$$

Local Affinity Matrix

$$[A_{local}]_{i,j} := \begin{cases} \epsilon_{self}, & i = j, \\ \frac{\langle Q_i, K_j \rangle}{\|Q_i\| \cdot \|K_j\|}, & j \in N(i), \\ 0, & else \end{cases}$$

4.3 Uncertainity-weighted Affinities

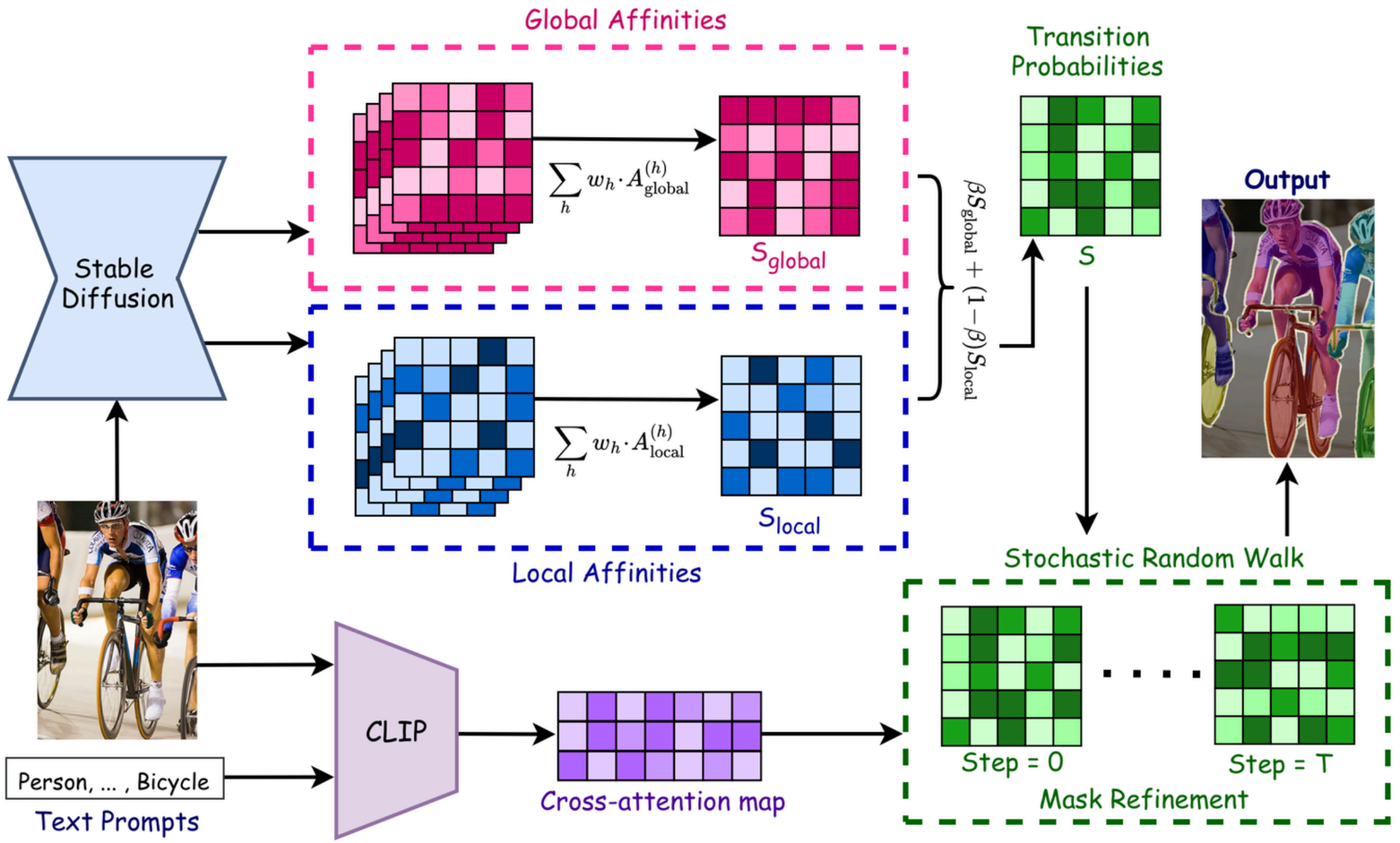
$$w_h := \frac{\exp(-c \cdot H^{(h)})}{\sum_{h'} \exp(-c \cdot H^{(h')})}$$

$$A_{weighted} := \sum_{h=1}^{H} w_h \cdot A^{(h)}$$

4.4 Transition Probability Matrix

$$S := \beta S_{global} + (1 - \beta) S_{local}$$

4.5 Proposed Pipeline



05 Experiments

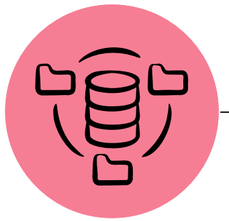
1 Setup

2 Main Results

3 Ablations

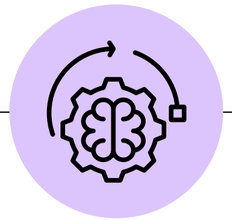
4 Qualitative Results

5.1 Setup



Dataset

- VOC
- ADE20K
- Context
- COCO-Stuff
- COCO-Object



Backbones

- CLIP - ViT-L/14
- Stable Diffusion 2.1

Metric

- mIoU

5.2 Main Results

Table 1. Quantitative evaluation across five datasets and two of their variants.

Method		Post.	V21	PC60	C-Obj	V20	ADE	PC59	C-Stf	Avg
CLIP	ICML'21	No	18.6	7.8	6.5	49.1	3.2	11.2	5.7	13.6
MaskCLIP	ECCV'22	No	43.4	23.2	20.6	74.9	16.7	26.4	16.7	30.3
GroupViT	CVPR'22	No	52.3	18.7	27.5	79.7	15.3	23.4	15.3	30.7
CLIP-DIY	WACV'24	No	59.0	--	30.4	--	--	--	--	--
GEM	CVPR'24	No	46.2	--	--	--	15.7	32.6	--	--
SCLIP	ECCV'24	No	59.1	30.4	30.5	80.4	16.1	34.2	22.4	38.2
PixelCLIP	NeurIPS'24	No	--	-	--	85.9	20.3	37.9	23.6	--
NACLIP	WACV'25	No	58.9	32.2	33.2	79.9	17.4	35.2	23.3	39.4
iSeg	Arxiv'24	No	68.2	30.9	38.4	--	--	--	--	--
NERVE	**Ours**	**No**	**69.7**(+1.5)	**37.7**(+5.5)	**43.3**(+4.9)	**90.1**(+4.2)	**24.0**(+3.7)	**43.4**(+5.5)	**28.8**(+5.2)	**48.1**(+4.3)
SCLIP	ECCV'24	Yes	61.7	31.5	32.1	83.5	17.8	36.1	23.9	40.1
ClearCLIP	ECVA'24	Yes	46.1	26.7	30.1	80.0	15.0	29.6	19.9	35.3
ProxyCLIP	ECCV'24	Yes	60.6	34.5	39.2	83.2	22.6	37.7	25.6	43.3
OVDiff	CVPR'24	Yes	--	--	--	80.9	14.1	32.2	20.3	--
CaR	CVPR'24	Yes	67.6	30.5	36.6	91.4	17.7	39.5	--	--
NACLIP	WACV'25	Yes	64.1	35.0	36.2	83.0	19.1	38.4	25.7	42.5
NERVE	**Ours**	**No**	**69.7**(+2.1)	**37.7**(+2.7)	**43.3**(+4.1)	90.1(-1.3)	**24.0**(+1.4)	**43.4**(+3.9)	**28.8**(+3.1)	**48.1**(+2.3)

5.3 Ablation

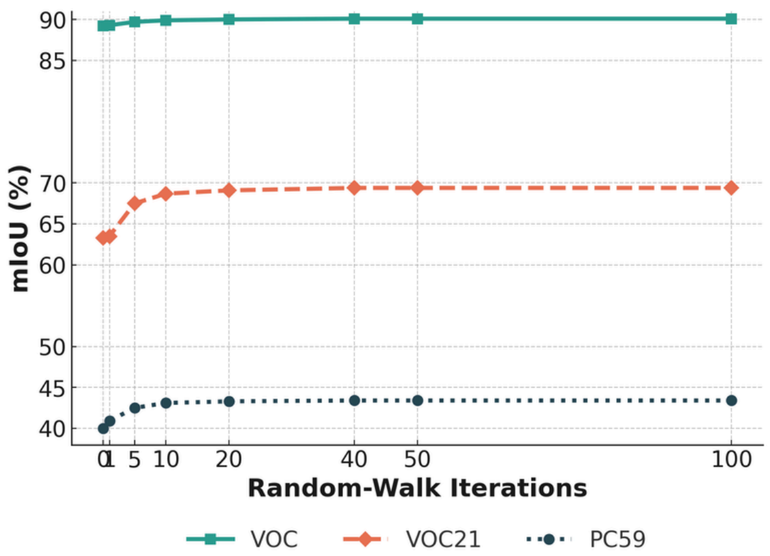


Figure 1. Impact of random-walk iteration count on segmentation mIoU

Type	VOC	Context	Object
Global	63.3	35.0	39.8
Local	58.7	34.5	37.6
Global + Local + RW	**69.7**	**37.6**	**43.4**

Table 2. Ablation on different affinity maps

Type	VOC	Context	Object
Single	67.2	36.6	42.0
Mean	66.9	36.5	41.7
Weighted Mean	**69.7**	**37.6**	**43.3**

Table 3. Comparison of multi-head attention fusion strategies in Stable Diffusion

5.4 Qualitative Result

Figure 2. Progressive segmentation refinement via stochastic random walk

06 Conclusion

We proposed NERVE:

- Training-free
- Spatially coherent
- Efficient
- State-of-the-art

Key insight. Stochastic propagation + entropy weighting improves segmentation without supervision.

Thankyou for your **attention!**