MaskInversion: Localized Embeddings via Optimization of Explainability Maps

Walid Bousselham[1,2], Sofian Chaybouti[1,2], Christian Rupprecht[3], Vittorio Ferrari[4], Hilde Kuehne[1,2,5]

University of Tuebingen[1], Tubingen AI Center[2], University of Oxford[3], Meta[4], MIT-IBM Watson AI Lab[5]

ICLR

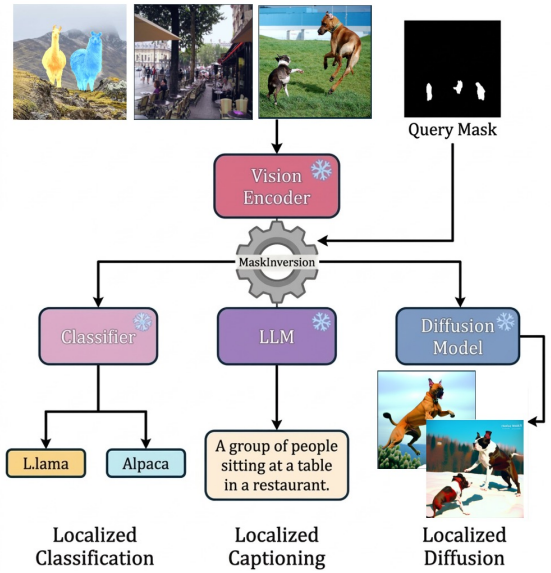
Motivation

The Problem:

- ✗ VLMs struggle with region-level understanding.
- ✗ Naive cropping loses context
- ✗ masking destroys image distribution.

Previous Methods

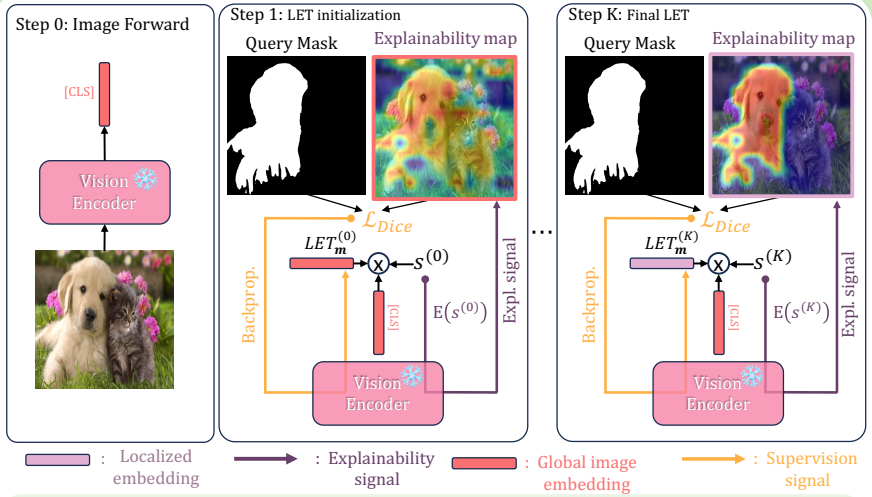
- ✗ require expensive fine-tuning (e.g., AlphaCLIP)
- ✗ modifying input pixels causing domain gaps.



Our Solution: Extract region-specific embeddings from frozen foundation models **at test time**.

How: Iteratively **optimize** an embedding token so its explainability map matches a user-provided query mask.

Method



MaskInversion consist of 3 steps:

1. **Initialize**: Start with the VLM's global [CLS] token
2. **Explain**: Generate a spatial explainability map
3. **Optimize**: Update only the token using a Soft Dice Loss to align the explainability map with the query mask.

Key benefits of MaskInversion:

- ✅ Frozen Backbone
- ✅ Context-Aware
- ✅ Drop-in replacement
- ✅ Multi-mask friendly

Gradient Decomposition

Computing gradients iteratively is **slow**.

⟹ Gradient Decomposition to reduce the explainability map generation to a simple dot product, drastically speeding up multi-mask processing.

$$\nabla \mathbf{A} = \frac{\partial s}{\partial \mathbf{A}} = \frac{\partial \bar{\mathbf{z}} \cdot \left(LET_{\mathbf{m}}^{(k)}\right)^T}{\partial \mathbf{A}} = \frac{\partial \bar{\mathbf{z}}}{\partial \mathbf{A}} \cdot \left(LET_{\mathbf{m}}^{(k)}\right)^T \in \mathbb{R}^{h \times n \times n}$$

Results

Method	Backbone	PhraseCut (*Acc@1*)	RefCOCO (*mIoU*)	PascalVOC (*Acc@1*)	MSCOCO (*Acc@1*)
CLIP	ViT-B/16	14.4	18.9	40.1	25.0
Crop	ViT-B/16	15.1	18.5	27.9	23.9
Masked Crop	ViT-B/16	48.3	52.9	75.0	38.2
RedCircle	ViT-B/16	21.5	43.2	47.5	28.8
FGVP	ViT-B/16	35.9	43.2	71.8	35.9
AlphaCLIP	ViT-B/16	34.0	44.0	52.6	30.9
MaskInversion (Ours)	**ViT-B/16**	**57.2**	**56.8**	**85.4**	**44.7**
MaskInversion (Ours)	**ViT-L/14**	**60.2**	**56.7**	**91.0**	**56.0**
MaskInversion (Ours)	**ViT-H/14**	**64.0**	**61.8**	**93.5**	**63.7**

Localized Diffusion

Localized Caption