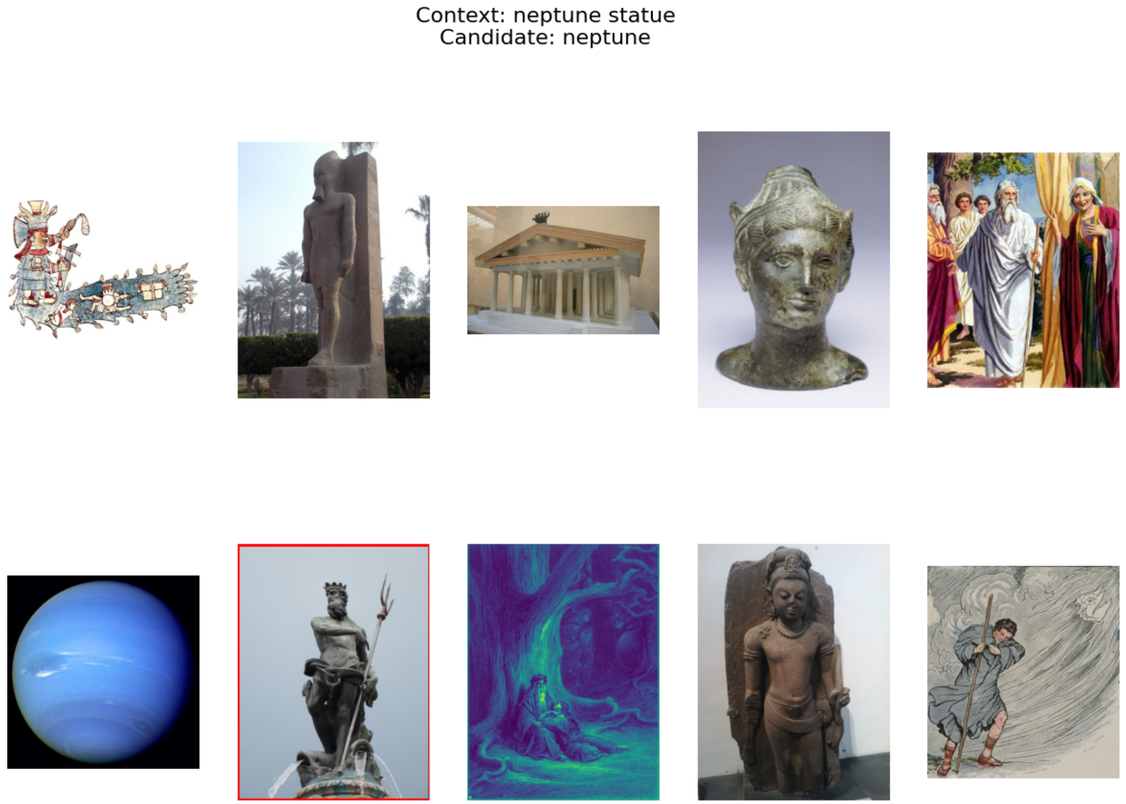
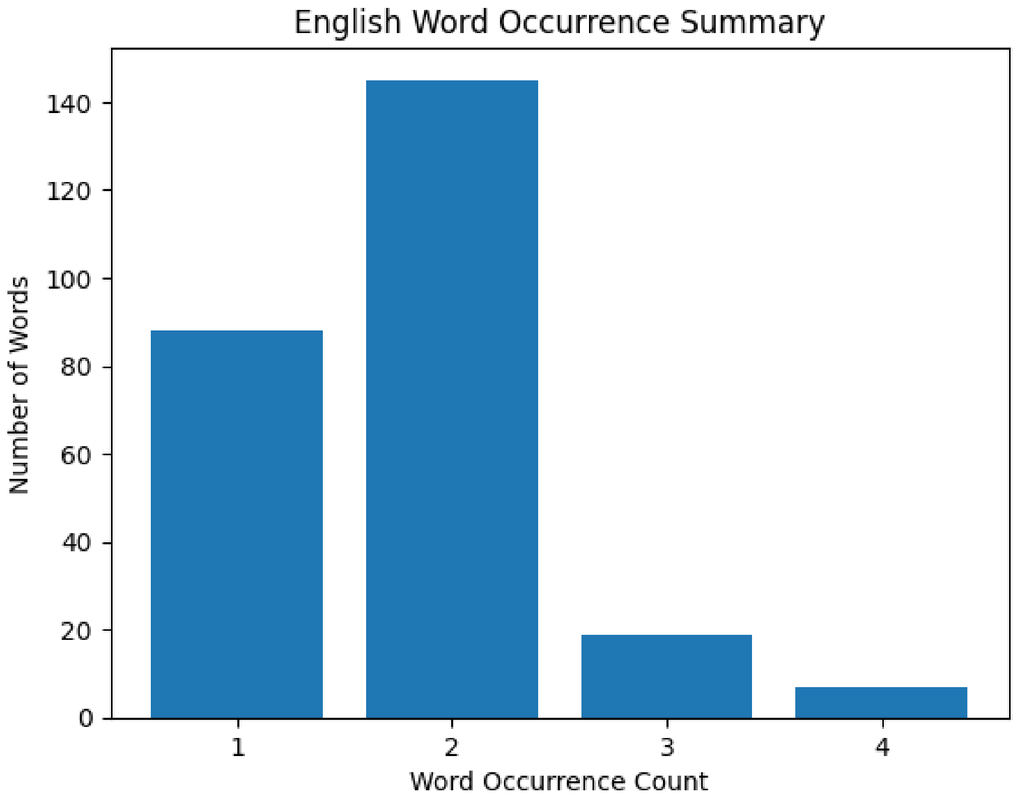
Homework 3: Visual WSD

Task challenges

- Handling cross-modality.
- Out Of Vocabulary words.
- Multilingual data.

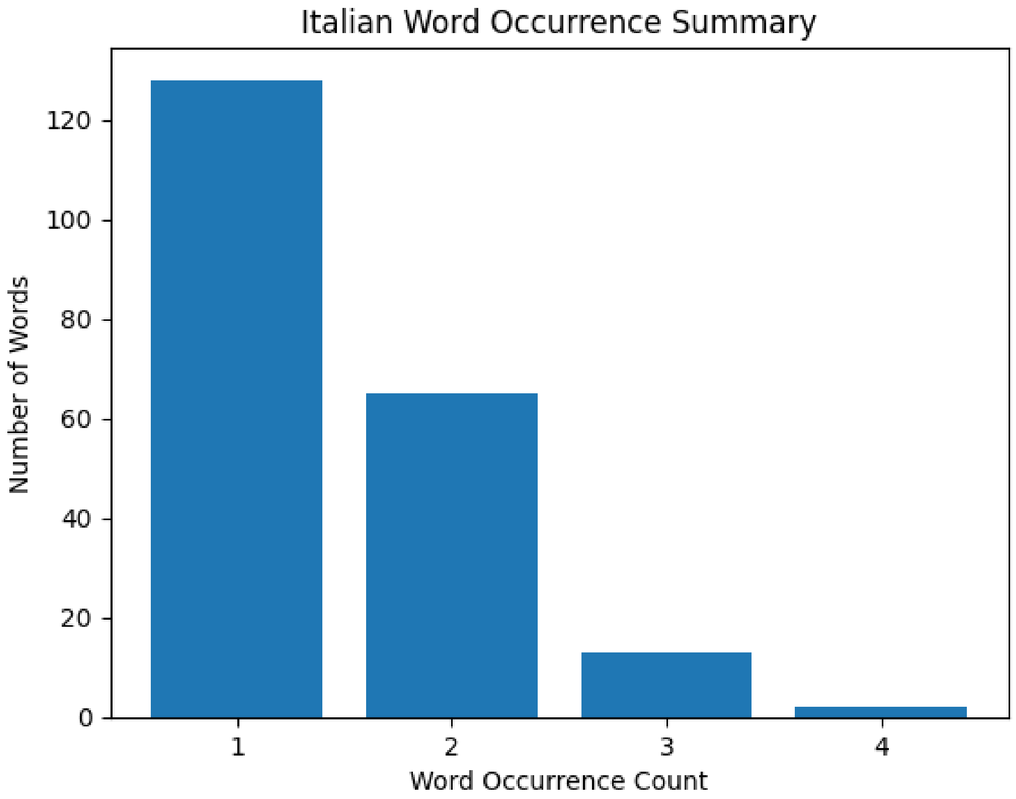


Test Dataset - English



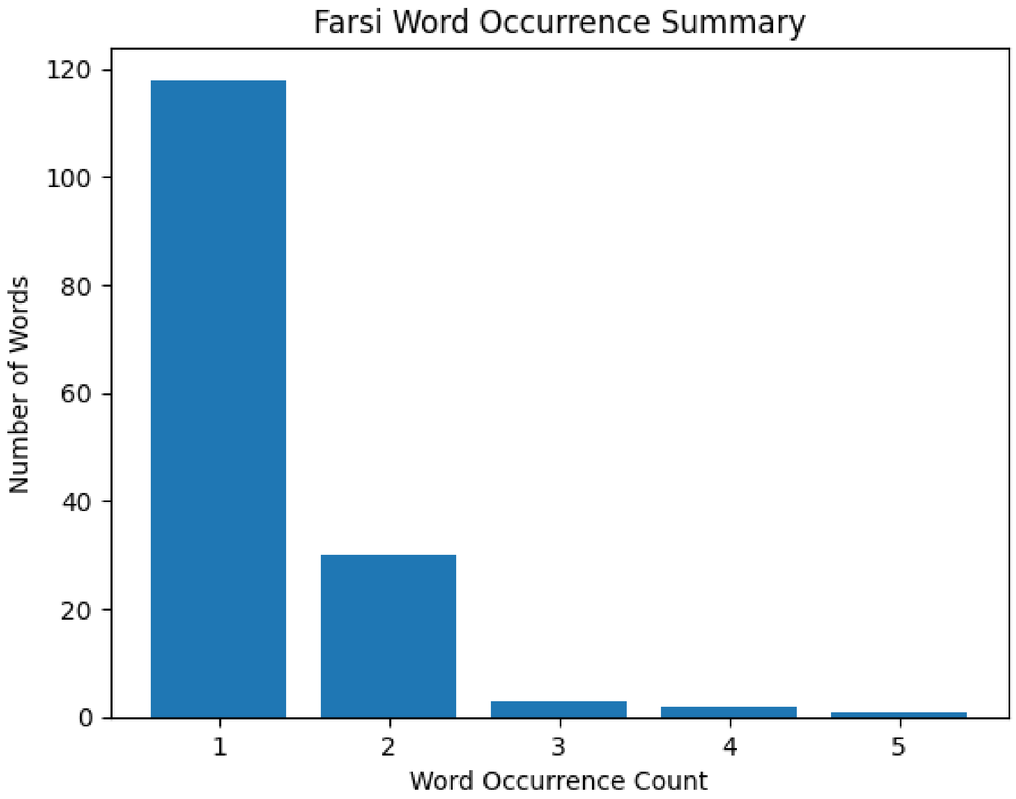
463 Samples

Test Dataset - Italian



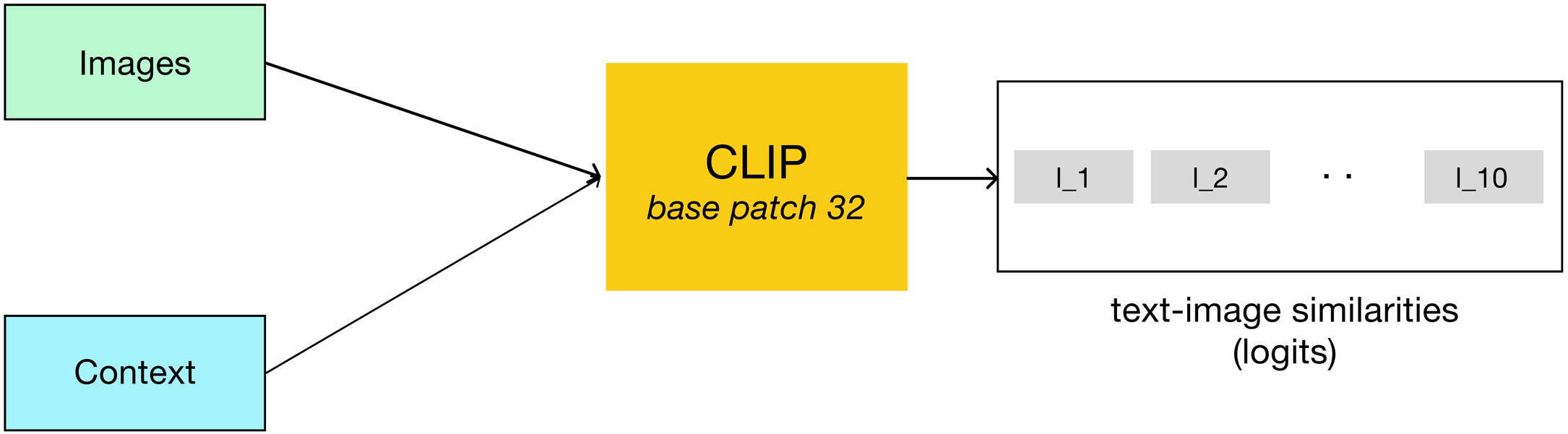
305 Samples

Test Dataset - Farsi



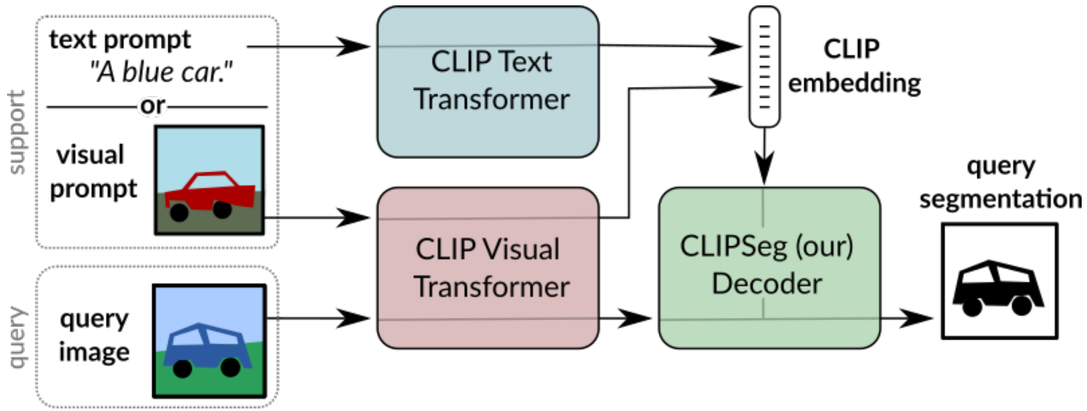
200 Samples

Baseline approach - CLIP



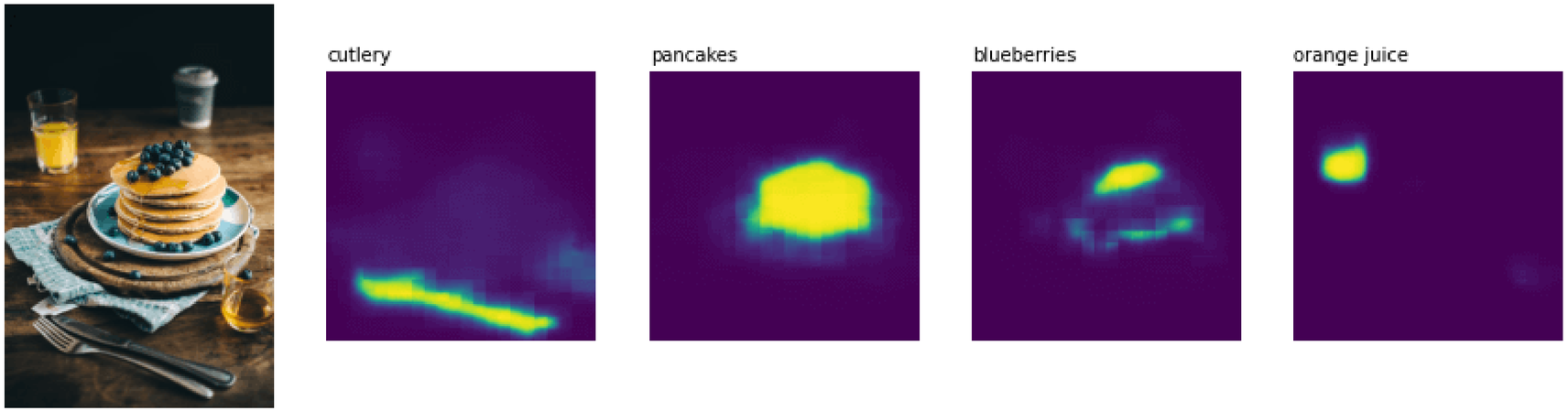
CLIPSeg

- CLIP-based image segmentation model (ViT-B/16).
- Uses a transformer based decoder that combines image and prompt embeddings to output a binary segmentation mask.
- Proved good performances in zero-shot and one-shot environents.

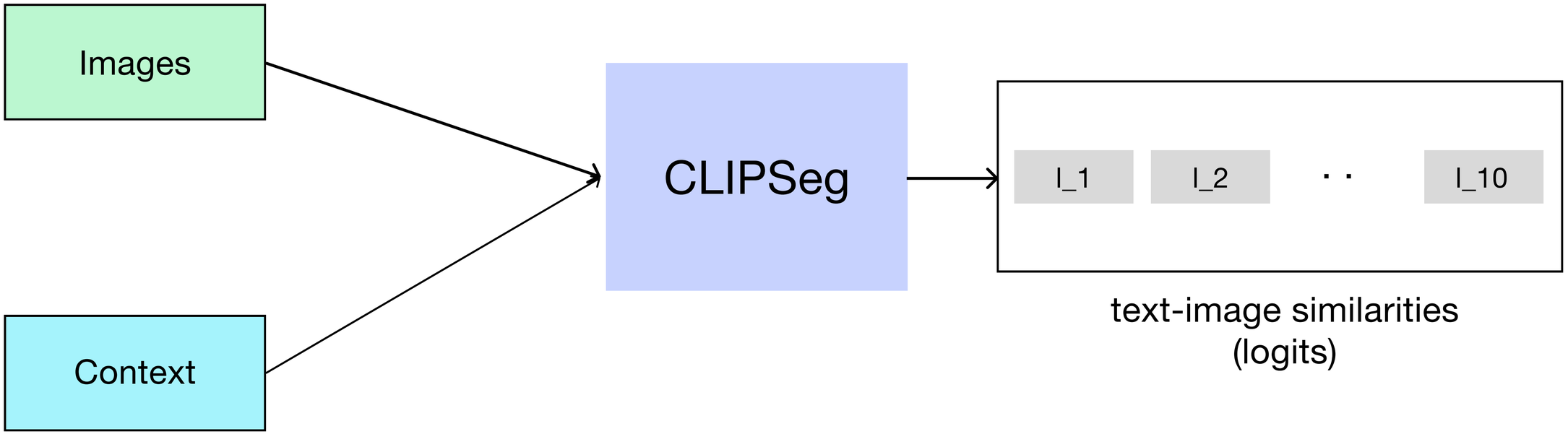


CLIPSeg - Why?

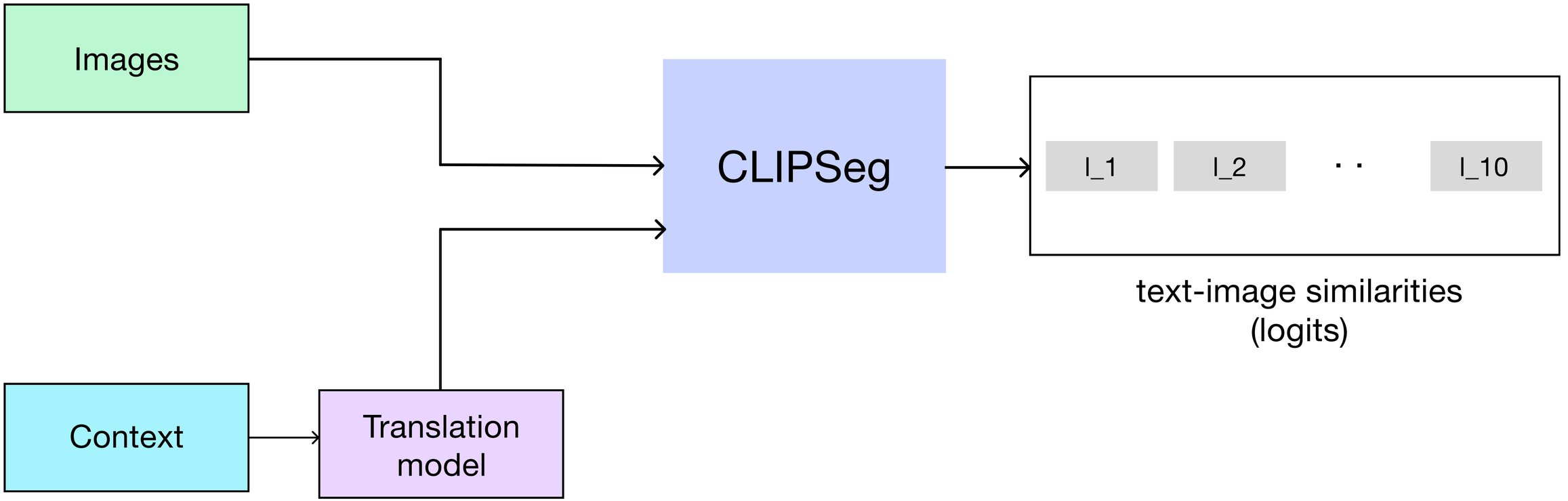
Segmentation models might have a superior ability at understanding a scene and the various elements that compose it.



CLIPSeg approach

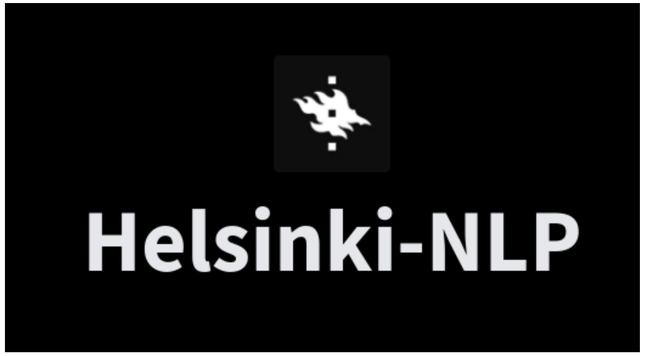


CLIPSeg + Translation approach



Translation models

- Deal with multilingual data.
- *opus-mt-it-en* for Italian test set.
- *mt5-base-parsinlu* for persian test set.



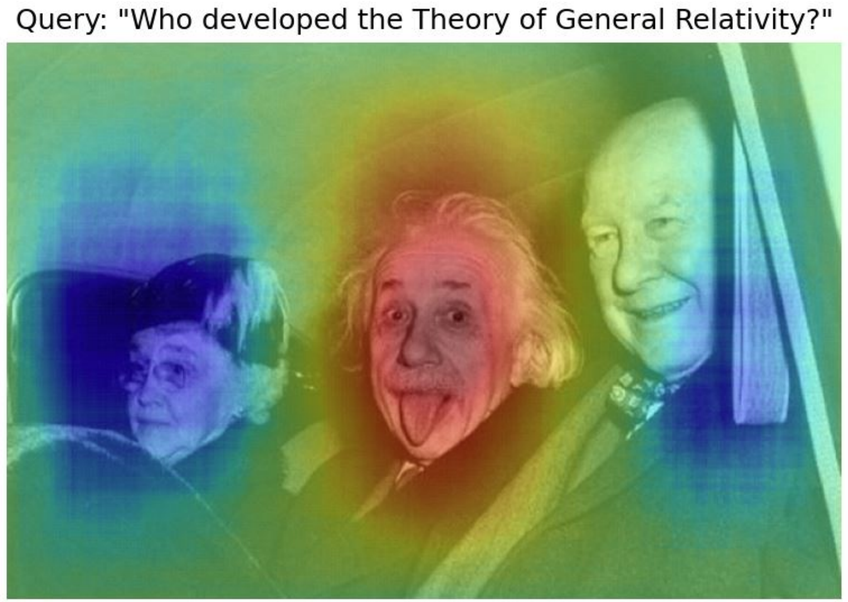
Results (accuracy)

	English	Italian	Farsi
CLIP	58.31%	-	-
CLIPSeg	**63.28%**	18.03%	9.5%
CLIPSeg + Translation	-	**50.49%**	**32%**

Qualitative analysis

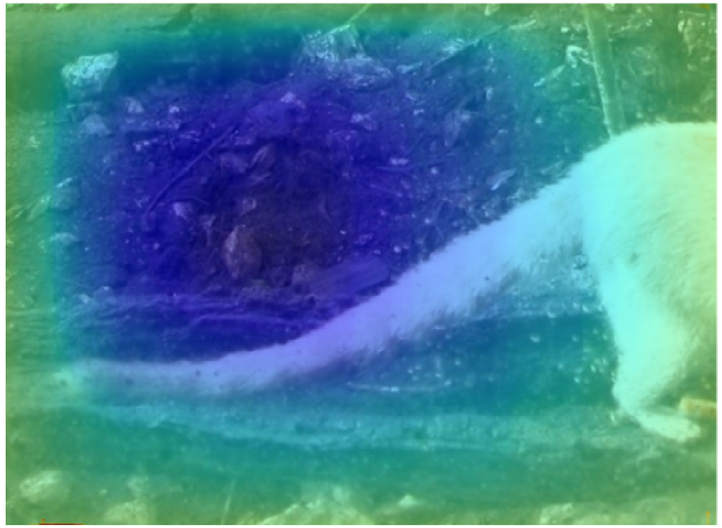
Saliency maps

- Which regions of the image are closer to the text query?
- Compute a query-image similarity for the whole image, and compare it to various crops of the same image.
- *Reason:* compare CLIP vs CLIPSeg capability at recognizing relevant parts of the image.

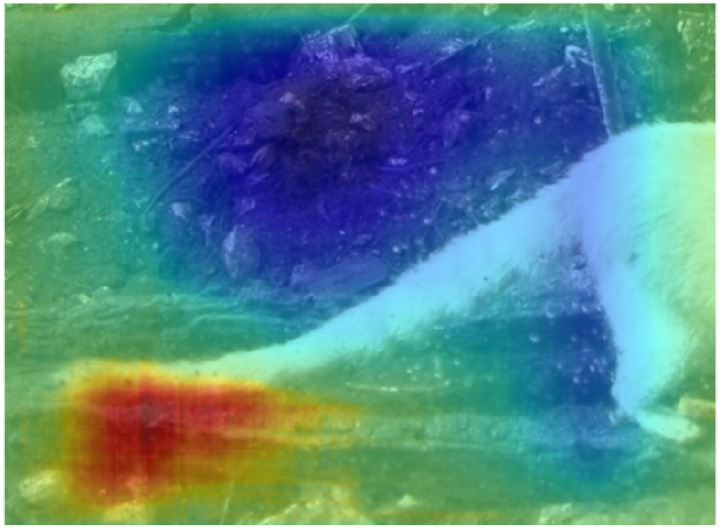


Saliency maps - comparison

CLIP



CLIPSeg



Query: “animal tail”

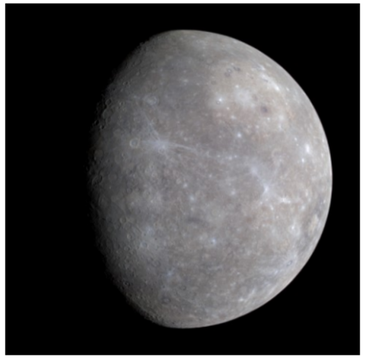
Saliency maps - comparison

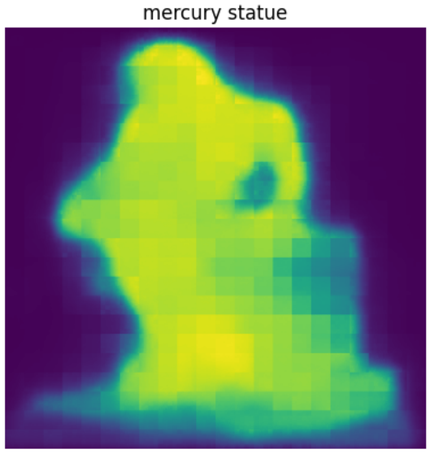
CLIP

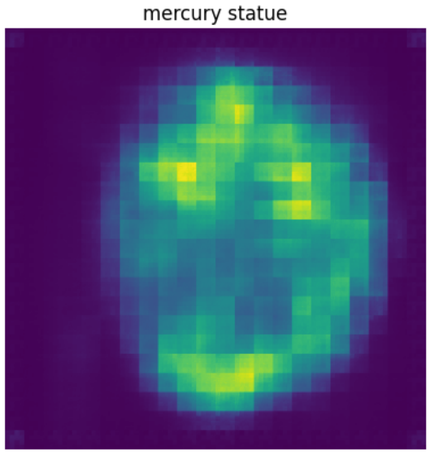
CLIPSeg

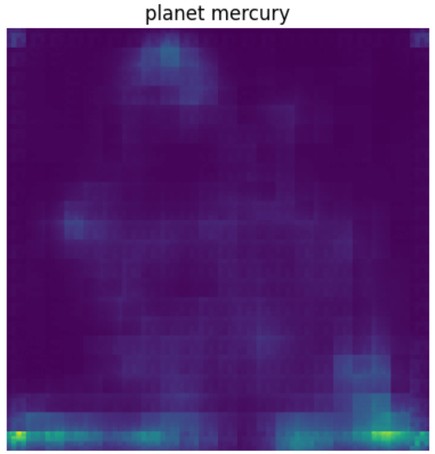
Query: “glutton hungry”

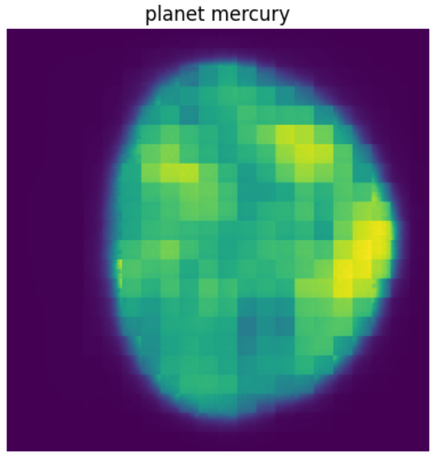
Segmentation



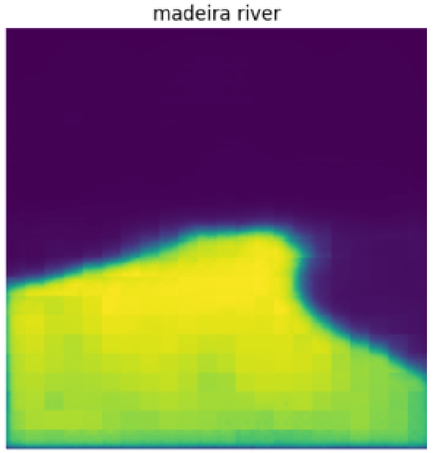


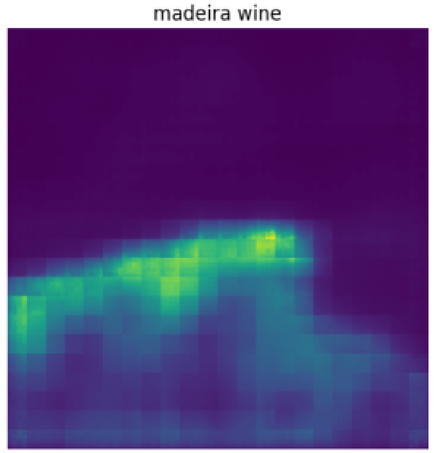


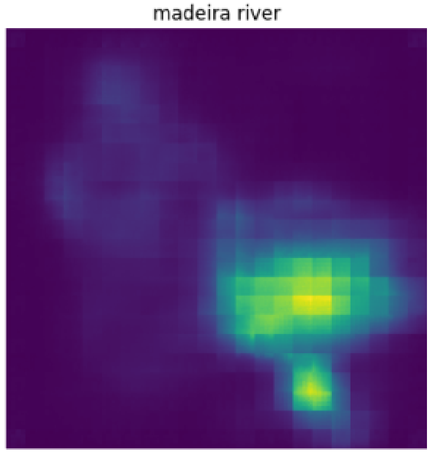


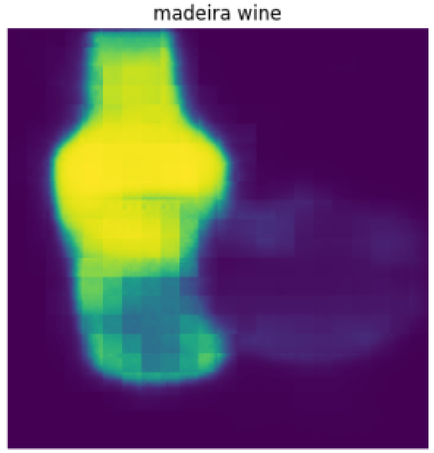


Segmentation









END