Image Segmentation using Text and Image prompts

HARDIK(22886), NIDHI(22947)

INTRODUCTION

- Introduces segmentation via arbitrary text or image prompts, avoiding re-training for new queries.
- Utilizes CLIP model with a transformer-based decoder for versatile segmentation tasks.
- Trained on the enhanced PhraseCut dataset for precise binary segmentation maps.
- Features dynamic adaptation through a novel hybrid input method for varied tasks.
- Increases system flexibility and application potential across diverse scenarios.
- Paves the way for advanced image analysis and interaction capabilities.

SEGMENTATIONS

Zero-shot segmentation involves segmenting objects or scenes that the model has not explicitly been trained to recognize, using only textual descriptions without prior visual examples.

ZERO-SHOT

One-shot segmentation refers to the ability of a model to segment specific objects or scenes from an image based on learning from just one example or image of that object or scene.

ONE-SHOT

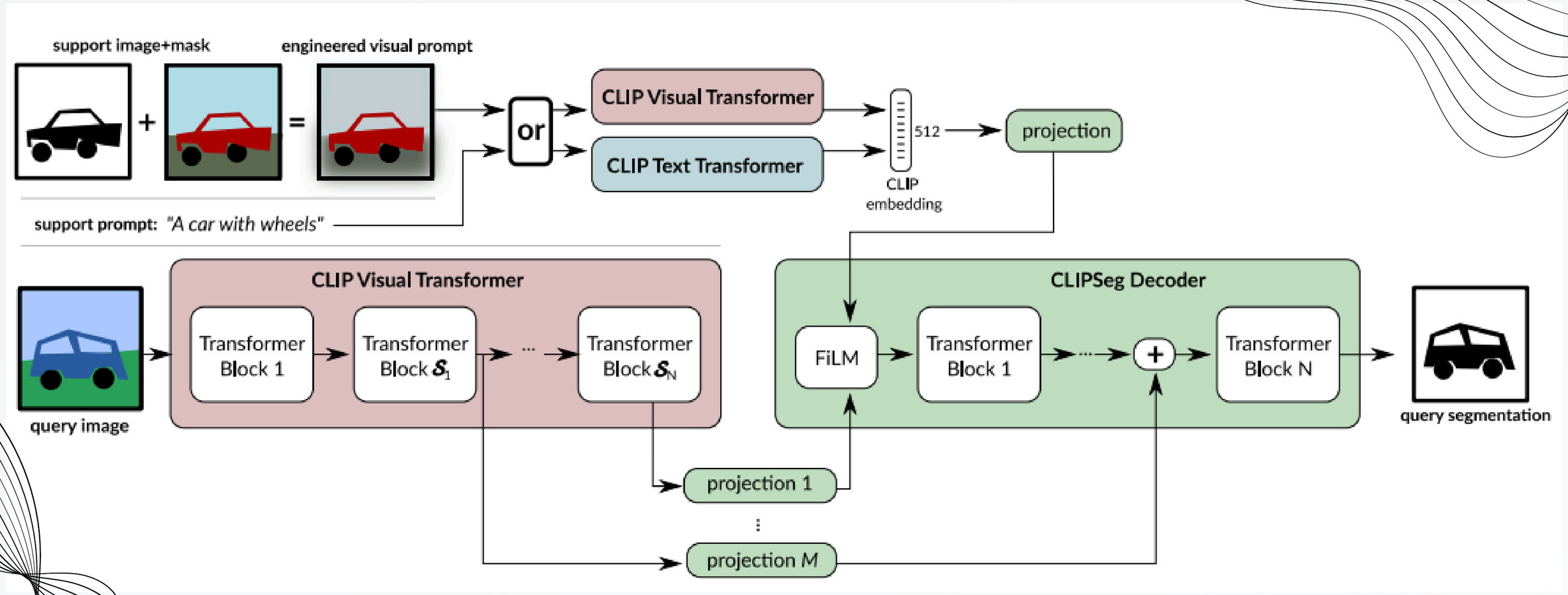
A referring expression segmentation task involves segmenting specific parts of an image that are uniquely described by a textual phrase, pinpointing objects based on descriptive language.

REFERRING EXPRESSION

CLIPSeg

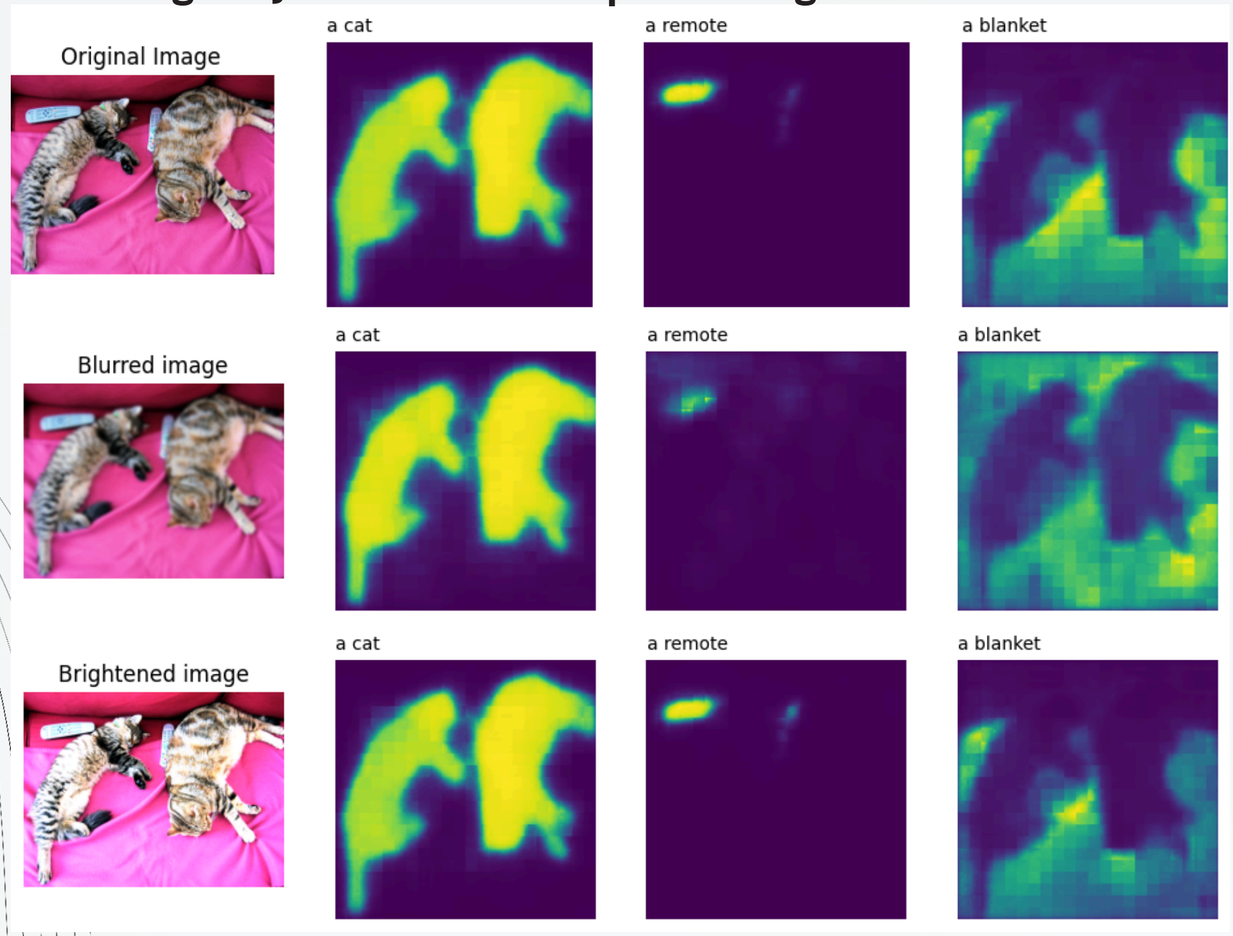
- **Backbone:** CLIPSeg extends (ViT-B/16)CLIP with a compact decoder architecture.
- **Decoder:** Features U-Net-inspired skip connections for efficient feature utilization.
- **Incorporation**: Utilizes specific CLIP encoder activations for segmentation.
- **Conditional Input:** Receives segmentation target info from text queries or images.
- **Efficiency:** Maintains a low parameter count with 1,122,305 trainable parameters.
- **Flexibility**: Supports variable image sizes via positional embedding interpolation.
- **Interpolation**: Employs image-text interpolation for data augmentation during training.

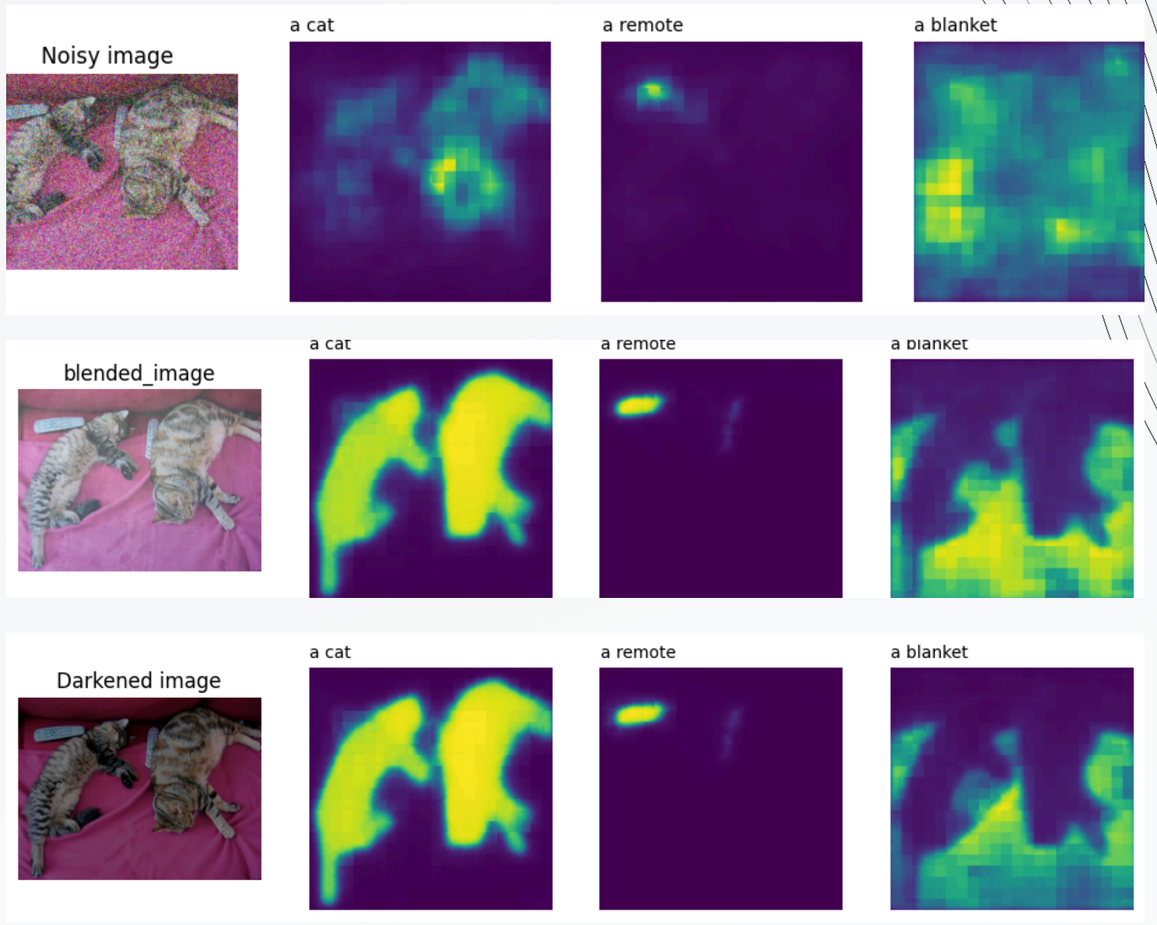
ARCHITECTURE



EXPERIMENT

Slightly distorted input image

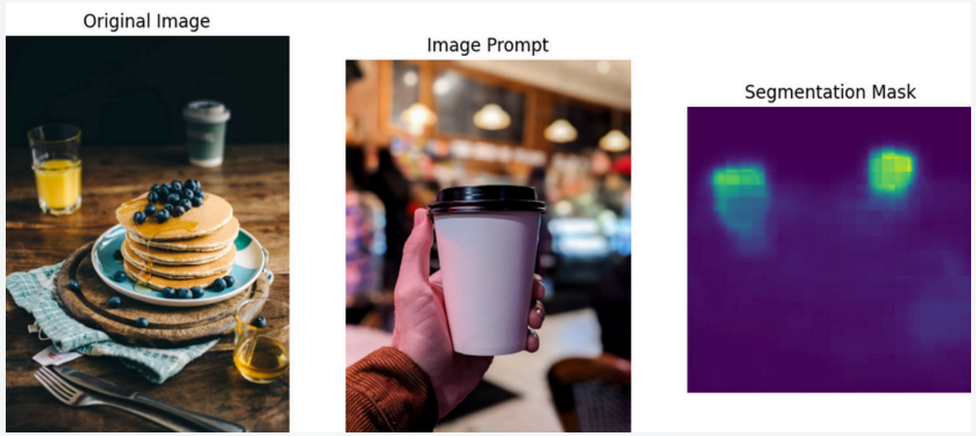


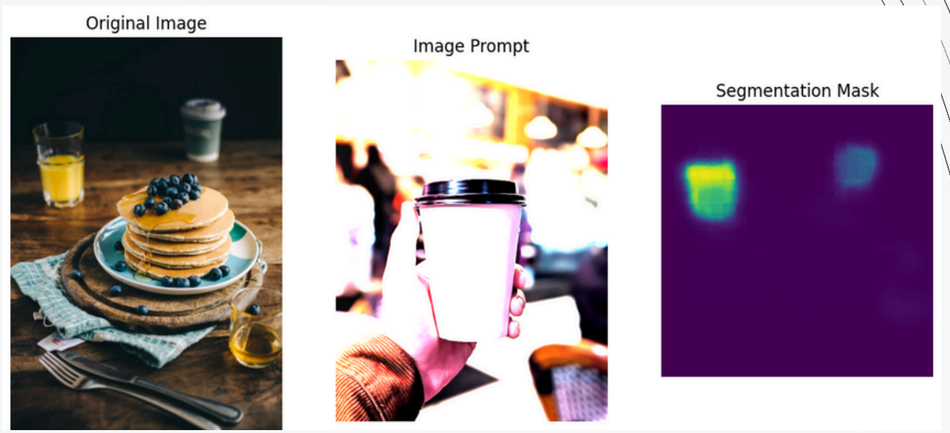


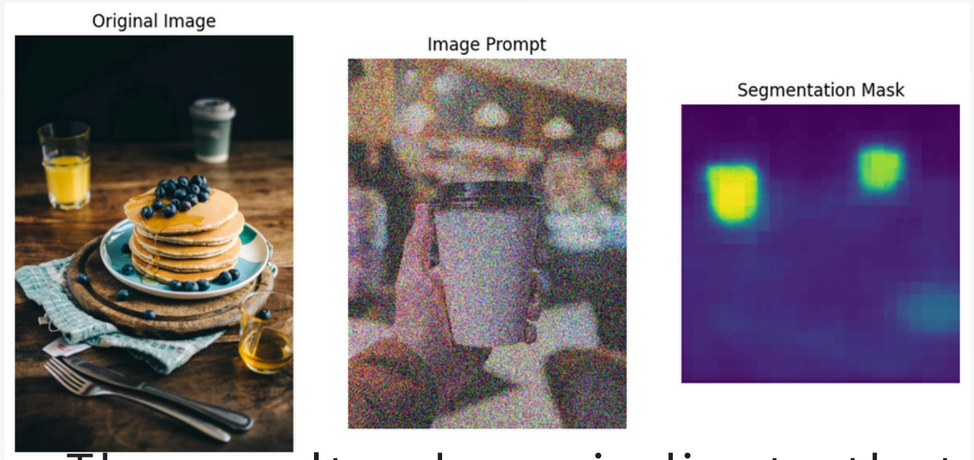
The results above indicate that CLIPSeg performs well even when presented with slightly distorted input images.

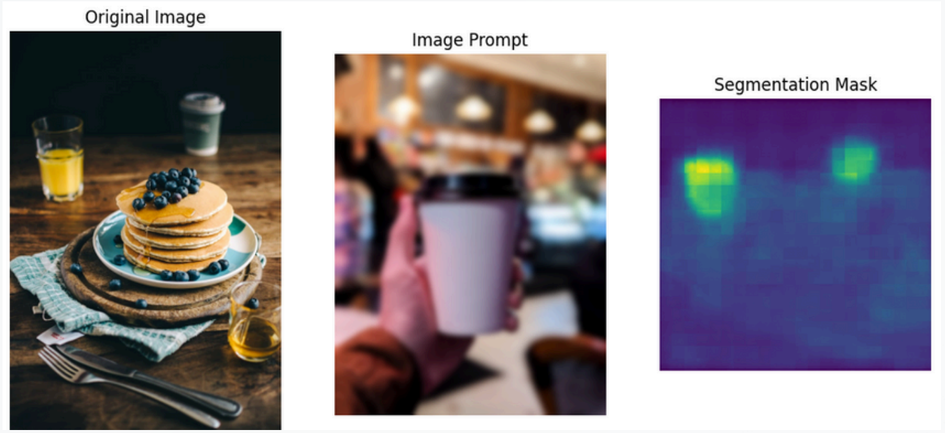
EXPERIMENT

Slightly distorted input image prompt





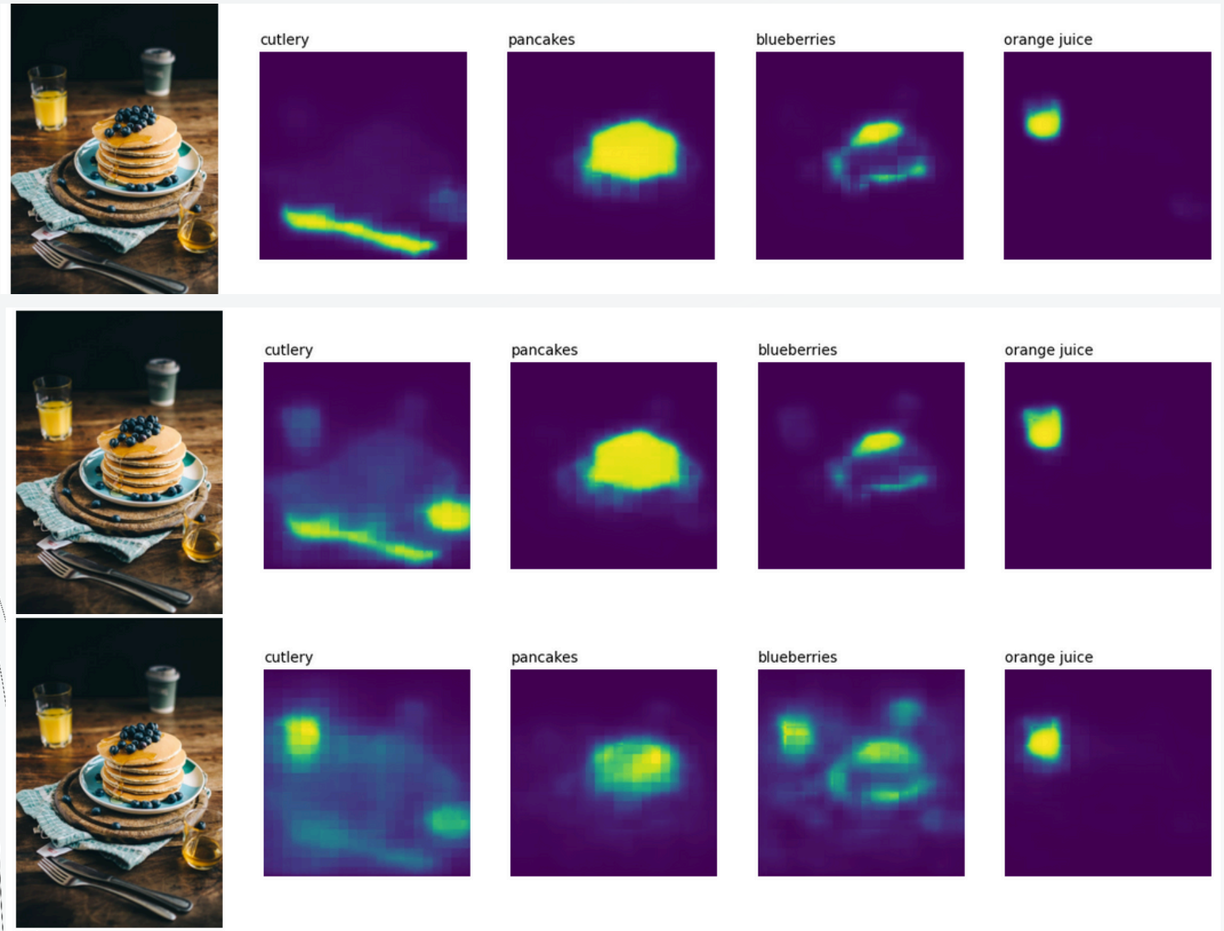


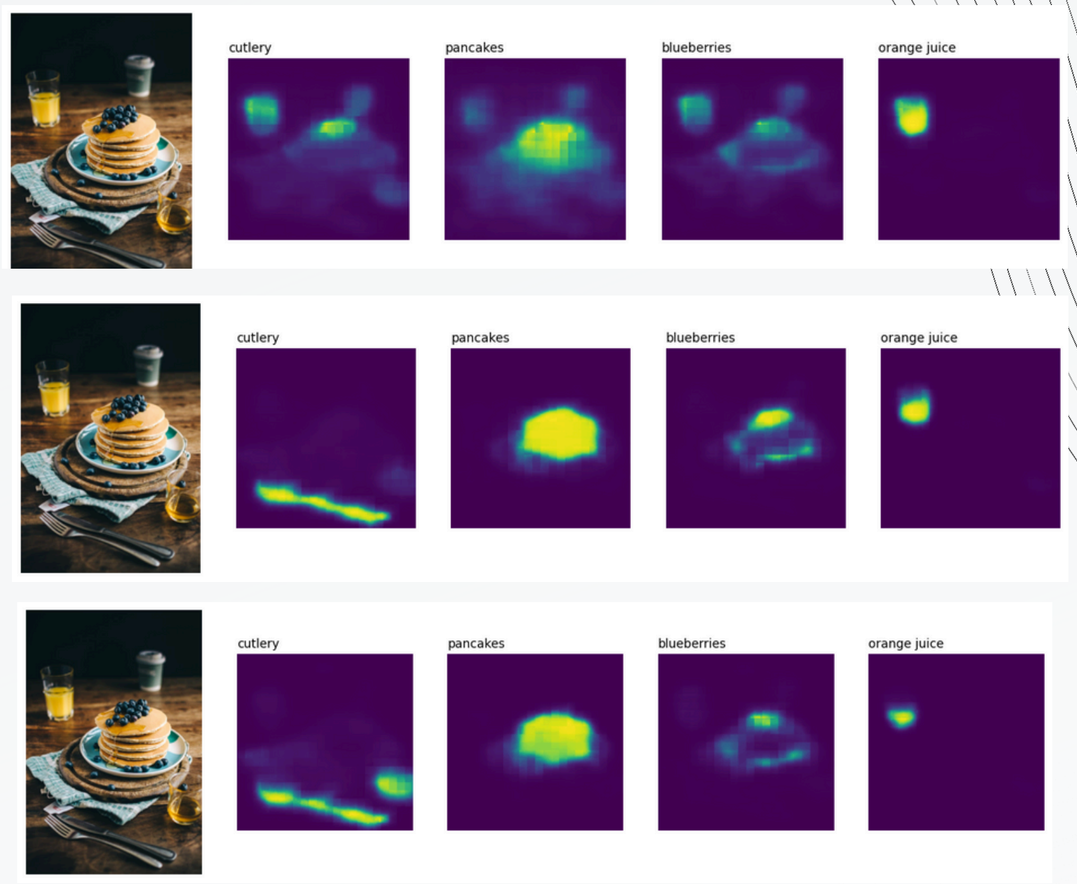


- The results above indicate that CLIPSeg performs well even when presented with slightly distorted input image prompts.

EXPERIMENT

Segmentation Outputs for different extract layers





- The above results show the segmentation outputs derived from different extract layers.

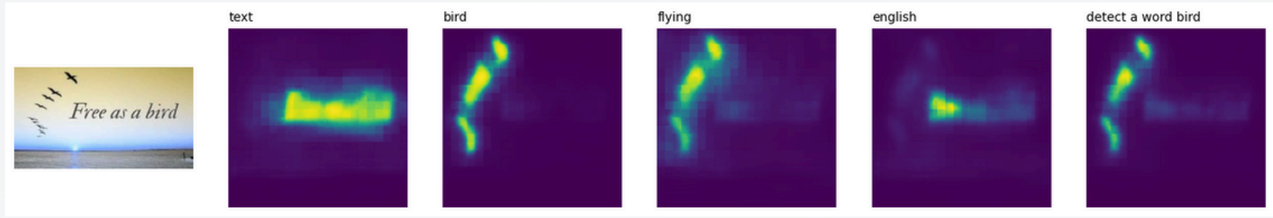
EXPERIMENT

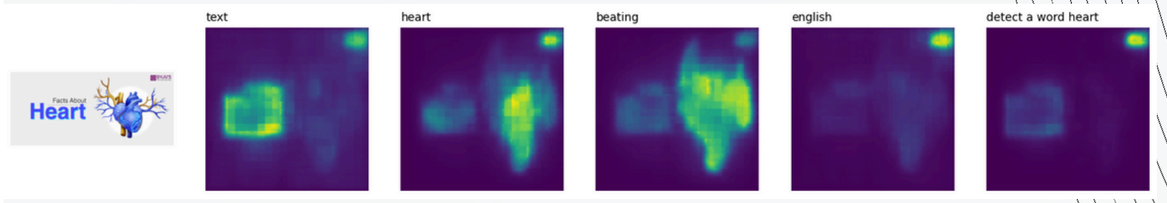
Extract_layers	Prompt 1	Prompt 2	Prompt 3	Prompt 4
(3,5,8)	0.51	0.97	0.719	0.676
(2,4,7)	0.0	0.0	0.0	0.767
(4,5,6)	0.0	0.0	0.0	0.671
(3,6,9)	0.88	0.94	0.797	0.967
(5,7,8)	0.56	0.87	0.00031	0.437

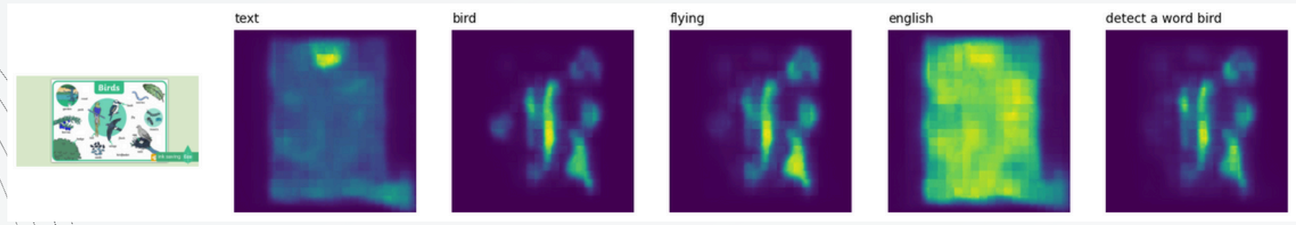
- IoU computed between the segmentation outcomes derived from layers (3,7,9) and the above mentioned extract layers.

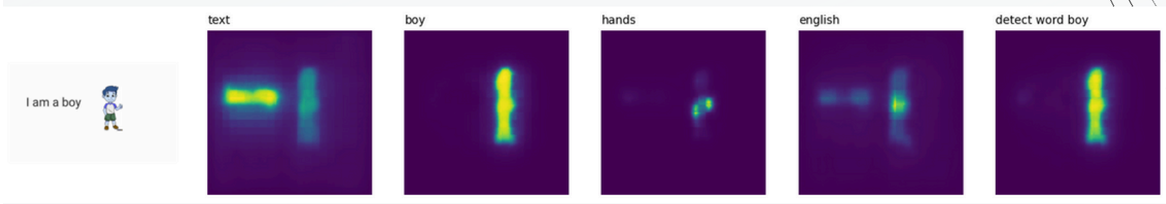
EXPERIMENT

Segmentation for a word





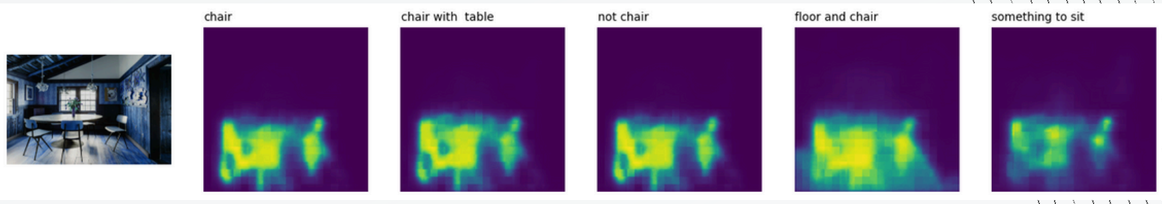


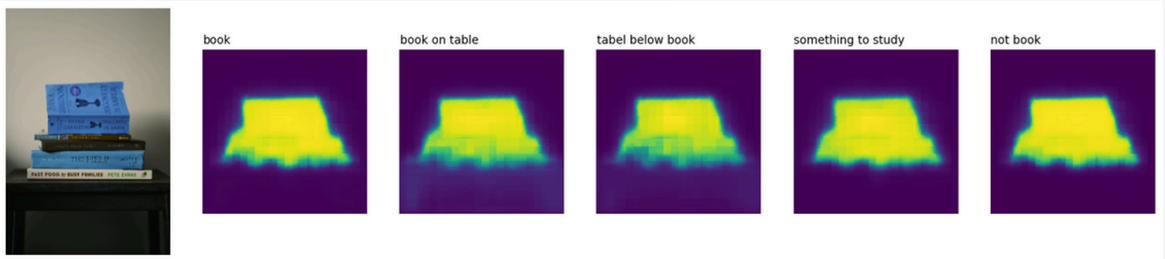


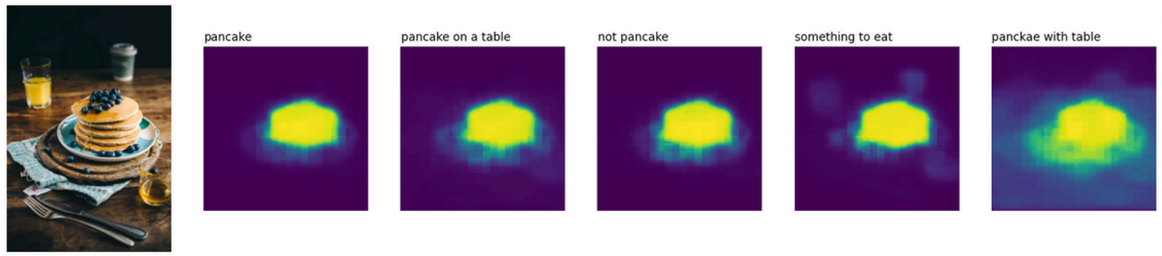
- The results above demonstrate that the CLIPSeg model struggles to segment images based on specific textual prompts accurately. Specifically,if the text prompt specifies a particular word present in the text of the input image.

EXPERIMENT

Different prompt for same segmentation







The results above demonstrate that the CLIPSeg model struggles to segment images accurately when a negative textual prompts, such as those containing negations of words (e.g., "not"), are provided.

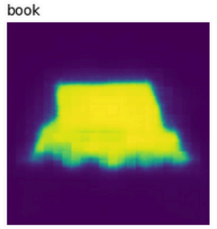
EXPERIMENT

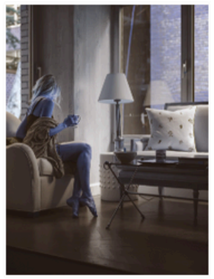
Facial Expression

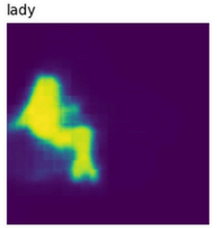
The results above demonstrate that the CLIPSeg model struggles to segment images based on specific facial expression accurately.

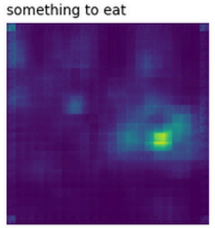
EXPERIMENT

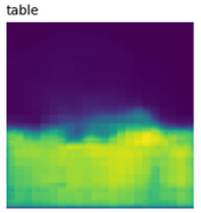
When object is not present











The experiments above demonstrate that the CLIPSeg model struggles to segment images accurately when the specified prompt is not present in the input image.

CONCLUSION

Although CLIPSeg is recognized as a state-of-the-art model, it faces difficulties in certain edge cases. The model excels in processing distorted input images and prompts, demonstrating robustness in these scenarios. However, while it is capable of detecting text in general, it struggles to identify specific text. Furthermore, its performance declines when the targeted object is absent from the input image. CLIPSeg also fails to differentiate between positive and negative prompts and cannot detect facial expressions, highlighting areas for potential improvement.

REFERENCES

1.Lüddecke T, Ecker A. Image segmentation using text and image prompts. InProceedings of the IEEE/CVF conference on computer vision and pattern recognition 2022 (pp. 7086-7096).
2.https://github.com/timojl/clipseg (The github link for the code we used for our project)