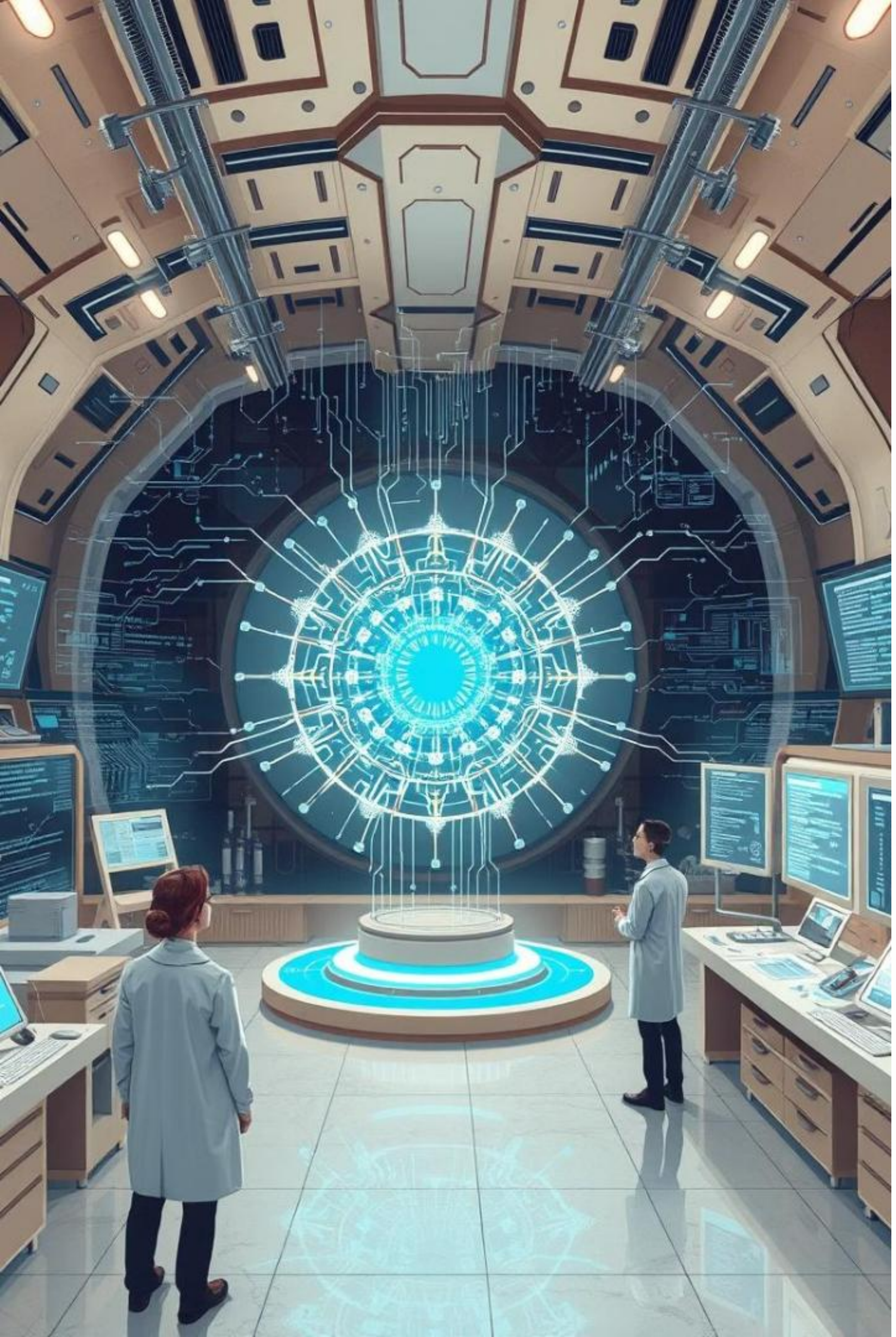


# CLIPSeg: Flexible Prompt-Based Image Segmentation System

A presentation on CLIPSeg - a model that segments images based on any text or image prompt.

**Presented by: Imene Bouaziz - Mohamed Amine Charfi**

# Plan Overview

1. Problem Statement
2. Proposed Solution
3. Solution Architecture
4. Datasets Used
5. Results
6. Conclusion & Future Outlook

# The Problem with Classical Segmentation

### Fixed Classes

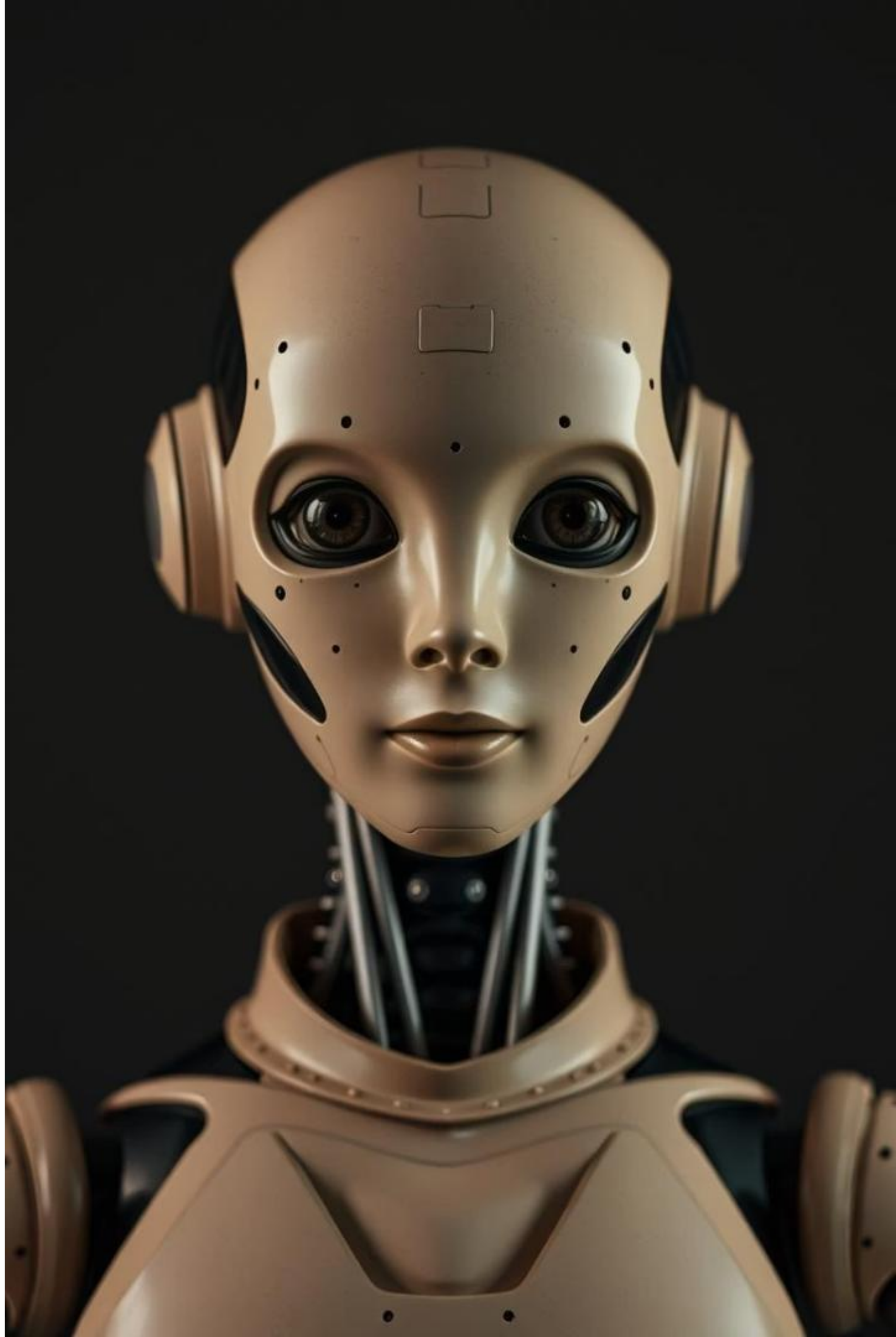
Classical models trained on fixed object categories only

### Limited Adaptability

Cannot handle new objects or contexts without retraining

### Core Challenges

- Zero-shot segmentation
- One-shot segmentation
- Referring expression segmentation



# CLIPSeg: The Proposed Solution

## Prompt-Based

Segments images from any text or image prompt

## Built on CLIP (Contrastive Language-Image Pretraining)

Uses shared embedding space for images and text

## Binary Segmentation

Foreground vs background output

## Multi-Task

Handles multiple segmentation tasks in one model

# CLIPSeg Architecture Overview

## Backbone

CLIP ViT-B/16:  Pretrained transformer encodes both images and text into a **joint semantic space**

## Lightweight Transformer Decoder

- 3 transformer blocks, with **U-Net-style skip connections**
- FiLM conditioning with prompts
- Only ~1.1M trainable parameters

## Prompt Types

Text prompt : encoded with CLIP's text transformer

Image prompt: processed using engineered visual cues

# Visual Prompt Engineering

### Support Images

Highlight target object for better segmentation

### Techniques Tested

- Cropping object
- Blurring background
- Darkening background

### Best Results

Combining all three techniques

# Datasets Used for Training & Evaluation

## Main Dataset

**PhraseCut:** 340K phrases with segmentation masks

**PhraseCut+ (PC+):** Enhanced with:

- Visual prompts
- Negatives samples
- Text-image interpolation during training

## Additional Benchmarks

- **Pascal-VOC :** for zero-shot segmentation
- **Pascal-5i & COCO-20i :** for one-shot segmentation
- **LVIS :** for generalization to abstract prompts

# Results Summary Across Tasks

| Task | Dataset | Performance |
|---|---|---|
| Referring Expression | PhraseCut | Outperforms classical methods |
| Zero-Shot | Pascal-VOC | Good on unseen classes |
| One-Shot | Pascal-5i & COCO-20i | Competitive with SOTA models |
| Abstract Prompts | LVIS | Handles conceptual queries well |

# Analysis of Model Components

## Key Findings

- No CLIP pretraining → huge performance drop
- Poor visual prompts → weaker results
- Smaller decoder / fewer layers → worse accuracy
- Visual & text prompts complement each other

## Limitations

- Only image data, no video
- Depends on undisclosed CLIP training data
- May inherit dataset biases

# Conclusion & Future Outlook

### CLIPSeg Strengths

Universal prompt-guided segmentation model

### Capabilities

- Referring expression segmentation
- One-shot segmentation
- Zero-shot segmentation
- Free-form conceptual prompts

### Potential Applications

Robotics, Human-computer interaction, no-code vision tools

### Future Work

- Extend to video
- Multimodal prompts (audio)
- Reduce training data dependence

# Demo

# Thank you for your attention

**Colab:** https://colab.research.google.com/drive/1yOOWX48ZOikr4SbxH_6kpyNQU9q1XWvy?usp=sharing

**Github:** "Image Segmentation Using Text and Image Prompts".

**Hugging Face:** CIDAS/clipseg-rd64-refined · Hugging Face