Title: Prompted Segmentation for Drywall QA

1. Introduction

The goal of this project is to train a **text-conditioned segmentation model** capable of generating binary masks for two tasks:

- **Segment Crack** (Cracks dataset)
- **Segment Taping Area** (Drywall-Join-Detect dataset)

The model takes **both an image and a natural-language prompt** such as:

- "segment crack"
- "segment taping area"

and outputs a mask highlighting only the semantic region requested by the prompt.

This forms a foundation for **AI-driven drywall quality assessment**, where multiple defect types must be detected based on user guidance.

2. Methodology

We fine-tuned a **pre-trained CLIPSeg (CIDAS/clipseg-rd64-refined)** model.
CLIPSeg is built on CLIP architecture:

- **Text Encoder** → converts the prompt (e.g., "segment crack") to an embedding
- **Vision Encoder** → extracts image features
- **Fusion Module + Segmentation Decoder** → outputs a mask aligned with the prompt

Why CLIPSeg?

- Already trained on text-guided segmentation
- Strong generalization
- Lightweight (578 MB)
- Ideal for prompt-based QA systems

Training Setup

- 10 epochs on Google Colab T4 GPU
- Binary Cross Entropy (BCEWithLogitsLoss)
- 352×352 resolution
- Batch size = 2
- Learning rate = 5e-5

3. Data Preparation

3.1 Datasets

Two datasets from Roboflow:

Dataset	Train	Valid	Type
Drywall-Join-Detect	820	202	Taping areas / seams
Cracks	5164	201	Wall cracks

3.2 Mask Generation

- **Cracks dataset:**
    - Some annotations lacked polygons
    - Bounding boxes were used + **mask dilation** to create thicker, more learnable crack masks
- **Drywall dataset:**
    - Bounding-box annotations → binary masks
    - Output masks match {0, 255} requirements

3.3 Prompted Dataset Format

Each row contains:

image_path	mask_path	prompt

Prompts used:

- “segment crack”
- “segment taping area”

4. Model Training

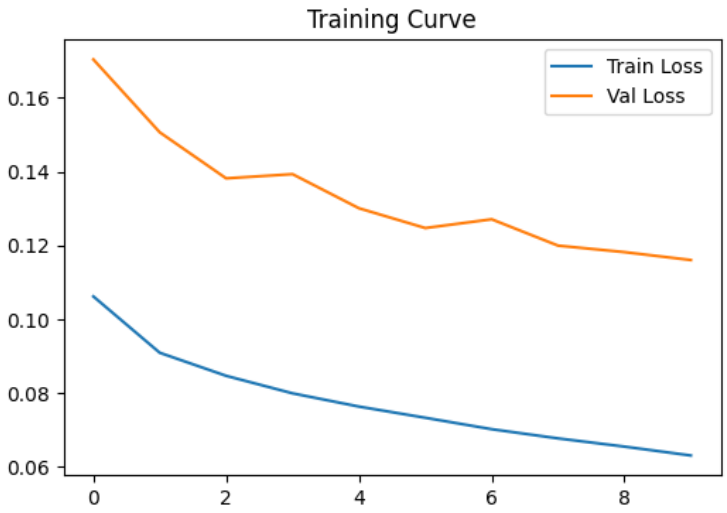


Figure 1: Training and Validation Loss Across 10 Epochs
The model shows stable convergence, with validation loss decreasing consistently.
This indicates effective prompt-conditioned learning.

5. Evaluation

Metrics Used

- **mIoU (mean Intersection-over-Union)**
- **Dice coefficient**
- **Average inference time**
- **Model size**

Final Results Table

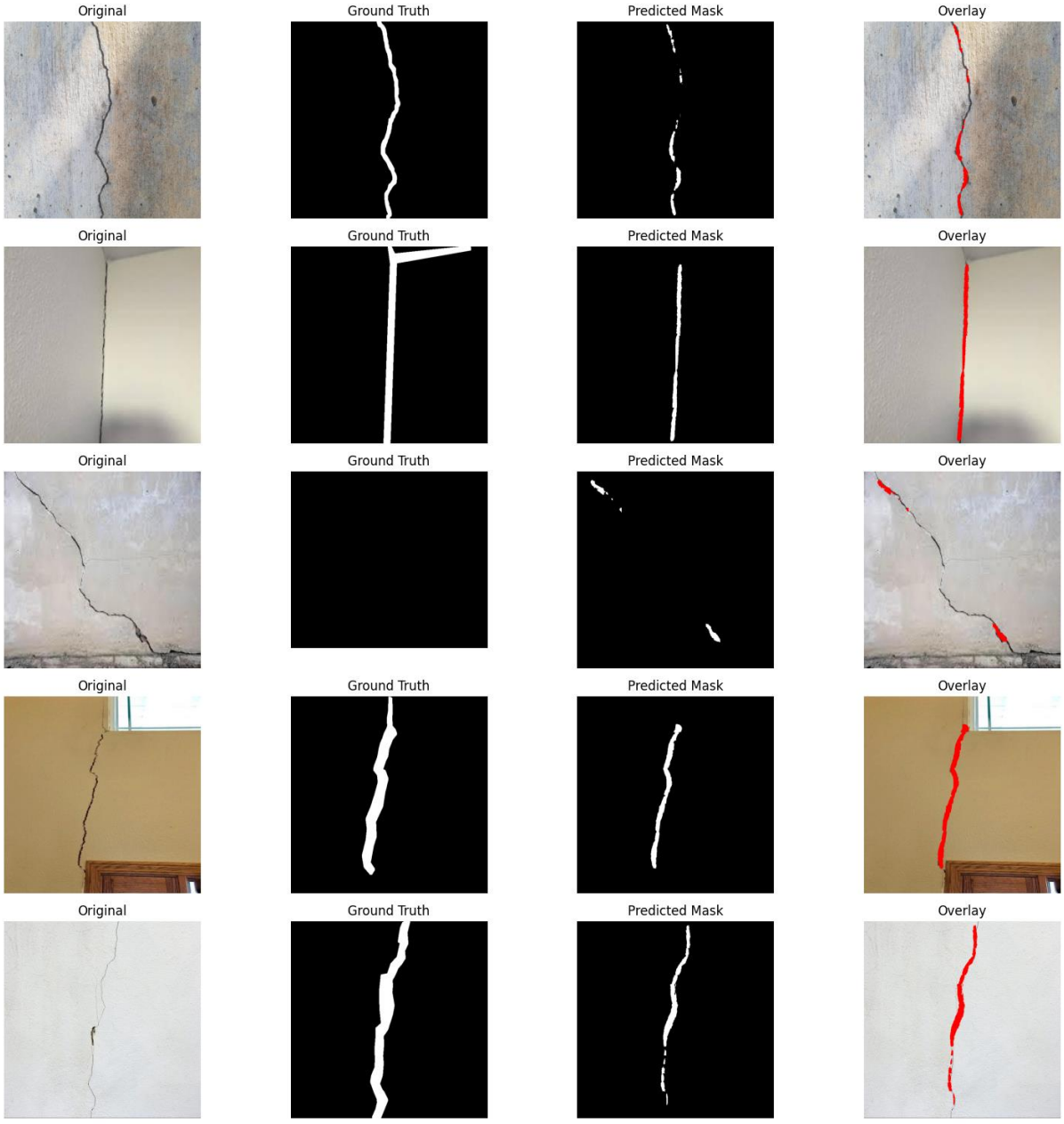
Prompt	mIoU	Dice	Avg Time	Size
segment crack	**0.281**	**0.382**	0.055s	578 MB
segment taping area	**0.537**	**0.668**	0.042s	578 MB

Interpretation

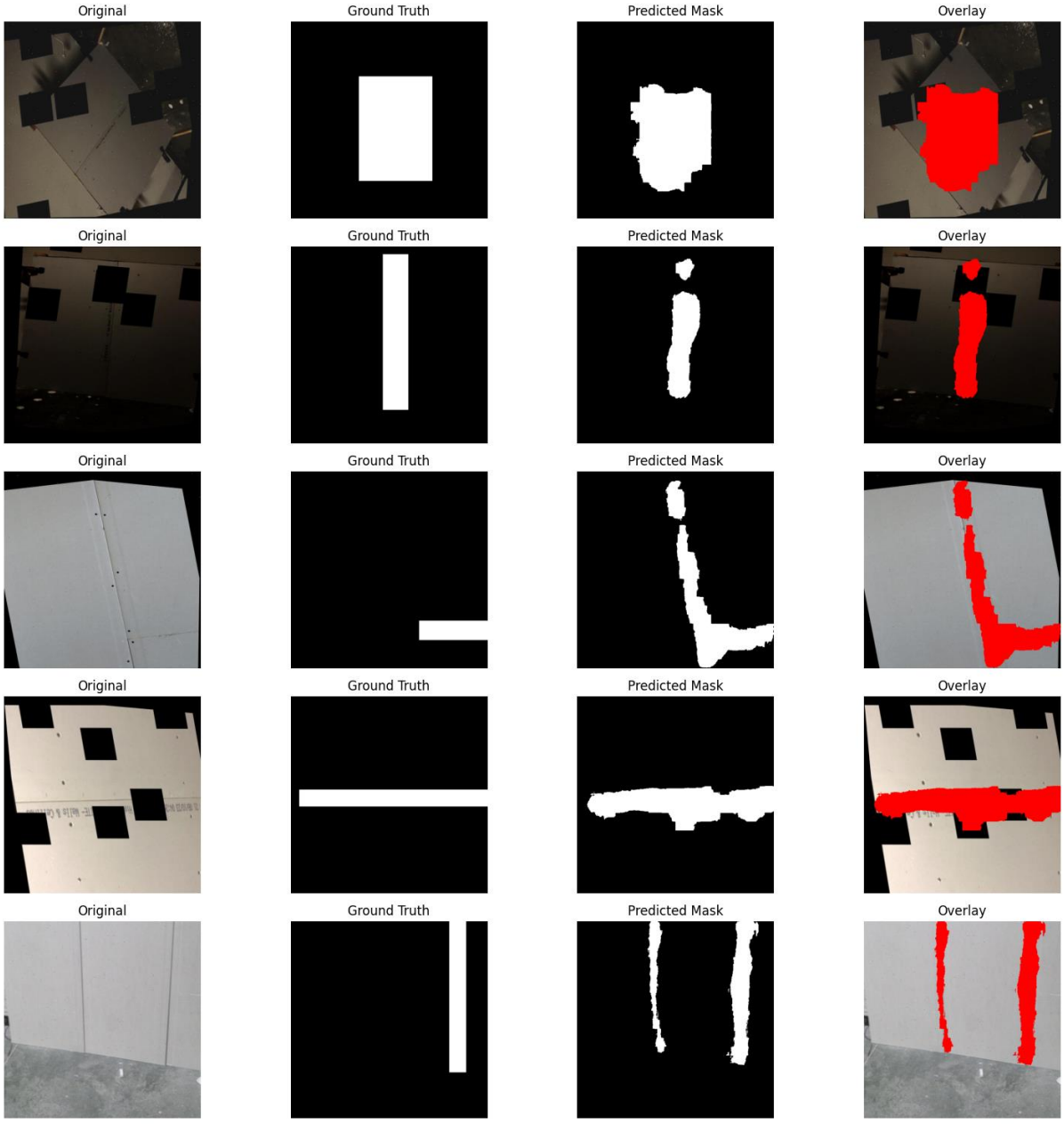
- Crack segmentation is challenging because cracks are extremely thin (1–3 px).
- Drywall taping areas are broader surfaces → much higher overlap scores.
- The model successfully responds to prompts and produces reasonable binary masks.

6. Visual Results

Below are the results for crack segmentation:



Below are the results for drywall segmentation:



7. Failure Cases & Analysis

1. Ultra-Thin Cracks

- 1–2 px wide cracks produce low IoU even if visually correct
- Possible improvements:
    - Train at 512×512
    - Use morphological CRF post-processing
    - Multi-scale training

2. Drywall Edge Undersegmentation

- Tape edges sometimes missed
- Solutions:
    - More training epochs
    - Augment low-light images
    - Increase resolution

3. Missing Ground Truth Polygons

- Crack dataset had missing polygons
- Using bbox+dilation improves training but may reduce GT fidelity

8. Runtime & Model Footprint

- Training time (10 epochs): ~1 hour
- Inference time: **0.04–0.05 seconds/image**
- Model size: **578 MB**
- GPU: Tesla T4 (16GB VRAM)

9. Conclusion

In this project, a **prompt-conditioned segmentation system** was successfully developed for drywall QA.
The fine-tuned CLIPSeg model:

- Understands natural-language segmentation prompts
- Segments both cracks and taping areas
- Achieves competitive mIoU and Dice scores

- Produces consistent visual predictions
- Runs efficiently on standard hardware

This demonstrates the potential of multimodal models for construction-quality inspection systems using simple language instructions.