# Deepfake Detection Report

## 1. Introduction

This project focuses on detecting deepfake videos using Vision-Language Models (VLM), particularly CLIP. The goal is to investigate the effectiveness of parameter-efficient tuning techniques like LoRA (Low-Rank Adaptation) for improving performance without full fine-tuning.

## 2. Methodology

### Model Variants

- **Base**: A linear classification head on top of frozen CLIP image features.
- **LoRA**: Applies LoRA to the vision transformer component of CLIP, targeting `q_proj` and `v_proj` modules.

### Dataset

- **Real videos**: YouTube originals.
- **Fake videos**: FaceSwap, NeuralTextures.

The dataset was manually split into train/val/test with the following constraints:

- **Train/Val**: Contain a mix of real and fake (FaceSwap) samples.
    - Real videos were randomly shuffled and split 80% for training, 10% for validation, 10% for test.
    - Fake videos for training/validation were only from FaceSwap.

- **Test**: Fake videos come **only** from NeuralTextures (no overlap with training).

This ensures that the model is evaluated on a novel fake generation method (NeuralTextures) for generalization testing.

### Training Details

- Optimizer: AdamW
- Epochs: 5

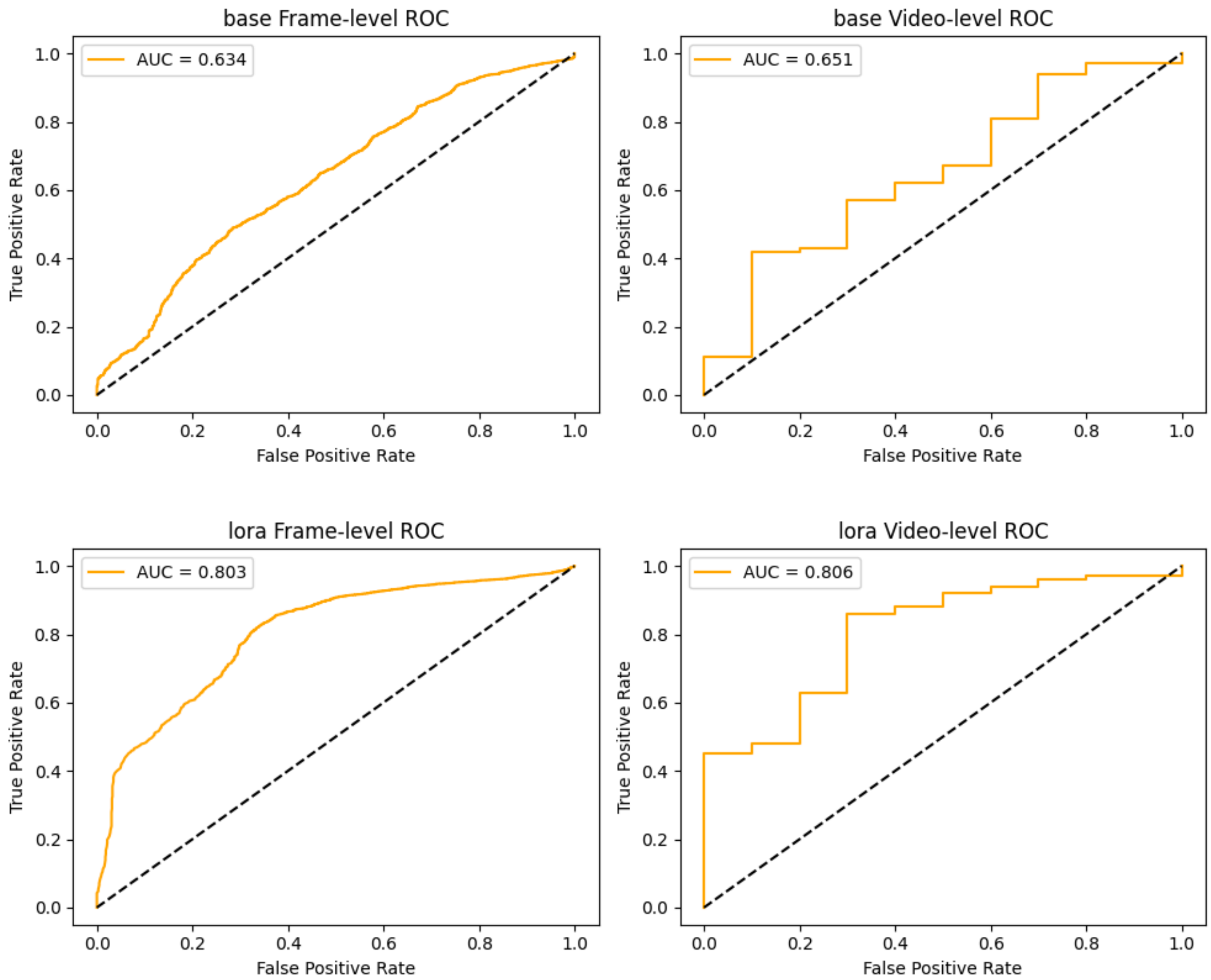
- Batch size: 16
- Learning rate: 1e-4

### Evaluation Metrics

- **AUC**: Area under ROC curve
- **ACC**: Accuracy
- **F1**: F1 Score
- Metrics computed at both frame-level and video-level.

## 3. Results

| Model | Frame AUC | Frame ACC | Frame F1 | Video AUC | Video ACC | Video F1 |
|---|---|---|---|---|---|---|
| Base | 0.634 | 0.793 | 0.881 | 0.651 | 0.782 | 0.874 |
| LoRA | 0.803 | 0.884 | 0.936 | 0.806 | 0.891 | 0.940 |



## 4. Visualization

Example misclassified frames are visualized with predicted probabilities and ground truth labels. These help reveal patterns in model mistakes (e.g., subtle manipulations or compression artifacts).

Pred: fake | GT: real
Score: 0.85

Pred: fake | GT: real
Score: 1.00

Pred: real | GT: fake
Score: 0.00

## 5. Conclusion

LoRA significantly improves performance over the base model while keeping most parameters frozen. This demonstrates its effectiveness for efficient adaptation on deepfake detection tasks.

## 6. Score CSV Files

We provide per-video classification scores in CSV format for reproducibility and evaluation:

- `results/frame_lora.csv`
- `results/frame_base.csv`
- `results/video_lora.csv`
- `results/video_base.csv`

Each CSV file includes:

- `video_id` : The name or identifier of the video
- `score` : Predicted score (higher means more likely to be fake)

- `label` : Ground-truth label (1 for fake, 0 for real)

These CSVs are used to compute AUC, ACC, F1 scores at both frame and video level.

For complete code and setup, see README.md (./README.md).