Cross-Manipulation Deepfake Detection with Vision-Language Foundation Models

1 Background and motivation

Deepfakes produced by ever-evolving generative pipelines are notoriously hard to detect when the forgery type at deployment differs from the type seen during training. Recent studies show that Vision-Language Models (VLMs) such as CLIP already encode semantic regularities that can be "re-programmed" for universal deepfake recognition with only a handful of additional parameters arXiv. You will explore that capability under a strict **cross-type** setting.

2 Dataset and official split

You will download a pre-extracted frame package (link provided separately).
The archive contains three classes, all sourced from the FaceForensics++ C40 subset by the following Link: https://www.dropbox.com/t/2Amyu4D5TulaIofv

After decompressing this zipped file, you will have:

- Real_youtube: Real videos
- NeuralTextures: Fake videos
- FaceSwap: Fake videos

Each class holds automatically extracted RGB frames (PNG/JPG) sampled at 1 fps; filenames keep the original video IDs. **Note that you can only train on "Real_youtube" and "FaceSwap" and test on "NeuralTextures".**

3 Task requirements

- **Model backbone** – choose a publicly available VLM (CLIP, BLIP-2, LLaVA …).
- **Adaptation strategy** – you may add LoRA/Adapter layers, prompt tokens or visual re-programming masks; full fine-tuning of backbone weights is **forbidden**.
- **Training data** – use only Real + FaceSwap frames.
- **Evaluation** – run the frozen detector on every NeuralTextures frame, aggregate to video level if you wish, and report: **AUC, EER, F1, Accuracy, ROC curve**.
- **Analysis** – discuss why the chosen PEFT (Parameter-Efficient Fine-Tuning) scheme helps (or fails) to generalise; illustrate at least three misclassified NeuralTextures examples with saliency or prompt visualisation.
- *Either frame-level or video-level reasoning is acceptable; your mark depends on generalisation quality and insight, not on temporal modelling sophistication.*

4 Recommended reading (all links verified)

- **Adapting Vision-Language Models for Universal Deepfake Detection** – retains CLIP's text tower and shows prompt tuning advantages
- **Standing on the Shoulders of Giants: Reprogramming VLM for General Deepfake Detection** – input-perturbation LoRA-free re-programming
- **Facial Feature-Guided Adaptation for Foundation Models** – side-decoder with facial-component guidance for video forensics
- **Can ChatGPT Detect DeepFakes?** – contrasts multimodal LLMs with classical detectors, useful for discussion

5 Deliverables (submit via a public GitHub repo)

- **Source code** – tidy structure; one-command reproduction script; requirements.txt.
- **README** – dataset unpack instructions, training & inference commands, expected run-time.
- **Report** – PDF ≤ 6 pages, covering methodology, experiments, discussion, references.
- **Pre-trained weights** – upload to GitHub-large-files or release section; link in README.
- **Results JSON/CSV** – per-video scores and overall metrics for auditing.

6 Grading scheme (100 points)

Category	Max pts	A-level description
Reproducibility & code quality	15	Clean repo; scripted data loading; seeds fixed; runs on a single RTX 3060 or Colab T4 without manual patching.
Correct split adherence	10	No NeuralTextures frame leaks (checked by hash); split lists respected; any leak → 0 pts in this row **and** Experimental Results row.
Model design & PEFT implementation	25	Well-motivated choice of VLM; PEFT obeys parameter-efficiency rule (< 5 % of total parameters trainable); code modular and documented.
Experimental results	25	All requested metrics; AUC on NeuralTextures ≥ 0.80 earns ≥ 18 pts; ROC curve plotted and discussed.

Insightful analysis	15	Error taxonomy, comparison with at least one baseline (e.g., frozen CLIP linear probe), discussion referencing R1–R4.
Report writing	8	Logical flow, concise tables/figures, correct citation style, ≤ 6 pages.
Bonus: interpretability / extra unseen fake type	2	E.g., Grad-CAM visualisation or additional zero-shot test on Face2Face.

7 Integrity rules

- Any form of *NeuralTextures* exposure during training or prompt crafting constitutes cheating.
- Generated content (e.g., synthetic frames) is allowed only inside the FaceSwap + Real training pool.
- Peer discussion is welcome; sharing code or weights across teams is not.