LLM-Augmented Audio-Visual Lip-Sync Deepfake Detection

Pilot Notebook (Design & Planning)

Step 1: Pilot Dataset Selection

- Goal: Assemble a pilot list of 20–50 clips (train/val).
- Deliverable: Table with dataset names, real/fake counts, purpose.

Step 2: Preprocessing Policy

- Face detector choice
- Lip ROI size
- FPS standardization
- Crop stability rules

Step 3: Feature Plan

- SyncNet scores
- AV-HuBERT embeddings
- Whisper transcripts
- MFA alignment outputs

Step 4: LLM Scoring Specification

- Prompt templates
- Output fields
- Decision rubric

Step 5: Fusion & Evaluation Metrics

- Feature fusion approach
- Metrics: AUC, EER
- IoU for temporal localization

Step 6: Error Taxonomy

- Desync types
- Dubbing artifacts
- Rhythm anomalies

Step 7: Privacy & Ethics Note

- Face/voice handling
- Storage policy
- Sharing constraints

Step 8: Minimal Demo Checklist

- What constitutes a successful first pipeline run
- Evidence to capture

Pilot Dataset Selection (Step 1)

Goal: Assemble a pilot list of 20–50 clips (train/val) covering all datasets.

Dataset	Real Clips	Fake Clips	Total	Purpose (train/val)
FakeAVCeleb	5	5	10	Training
AV-Deepfake1M	5	5	10	Validation
LAV-DF	3	3	6	Training (temporal loc)
AVLips	3	3	6	Stress-test
LRS2 / LRS3	4	–	4	Pretraining/finetuning (real only)
Total	20	16	36	---

Notes:

- Core training/validation clips: FakeAVCeleb & AV-Deepfake1M
- Temporal and stress-test clips: LAV-DF & AVLips
- Real-only clips for VSR pretraining: LRS2 / LRS3
- This pilot set is small enough to test the pipeline quickly but representative of all datasets.

Total pilot clips = 36, which is within the recommended 20–50 range for quick testing.

Preprocessing Policy (Step 2)

Goal: Prepare video clips for consistent feature extraction.

1. **Face detector:**
    - Use **RetinaFace** for accurate face and landmark detection.
    - Alternative for pilot: OpenCV Haar cascade for fast checks.
2. **Lip ROI size:**
    - Crop mouth region including small margin.
    - Resize to **128x128 pixels** for model input.
    - Maintain aspect ratio if possible.
3. **FPS standardization:**
    - Convert all clips to **25 FPS**.
    - Ensures consistent temporal resolution across datasets.
4. **Crop stability rules:**
    - Center crop on lips using landmarks for each frame.
    - Smooth bounding box positions across frames to avoid jitter.
    - Skip frames with occluded faces or very small bounding boxes.

    Consistent preprocessing ensures reliable feature extraction across all clips.

Feature Plan (Step 3)

Goal: Define features to extract from each pilot clip.

1. **SyncNet Scores (Audio-Visual Sync):**
    - Compute per-frame sync score between audio and lip movements.
    - Helps detect misaligned lip-sync forgeries.
2. **AV-HuBERT Embeddings / Lip-Reading Features:**
    - Extract visual speech embeddings from cropped lip region.
    - Frame-level embeddings for temporal modeling.
    - Optional: generate lip-reading transcripts.
3. **Whisper Transcripts (Audio ASR):**
    - Convert audio to text using Whisper.
    - Provides textual content for LLM scoring and semantic comparison.
4. **Montreal Forced Aligner (MFA) Alignment Outputs:**
    - Align phonemes with audio timestamps.
    - Compare phoneme timing with visual visemes to detect desync.

Notes:

- These features will be **combined in later stages** for classifier input.
- SyncNet + AV-HuBERT → core A/V features
- Whisper + MFA → textual/phoneme timing features

All features will be saved as numpy arrays (.npy) for easy loading into the classifier.

LLM Scoring Specification (Step 4)

Goal: Use LLM to evaluate semantic and prosody consistency.

1. **Prompt Templates:**
    - Ask LLM to rate plausibility of spoken content based on Whisper transcripts and optional AV-HuBERT lip-reading transcripts.
    - Example prompt:

    ```
    Given the transcript and lip-reading content, rate how plausible the speech
    Transcript: <Whisper transcript>
    Lip-reading text (optional): <AV-HuBERT transcript>
    Rate consistency from 0 (fake) to 1 (real).
    ```

2. **Output Fields:**
    - **Score**: numeric value 0–1
    - **Optional reasoning text** for explainability
3. **Decision Rubric:**
    - Score > 0.7 → likely real
    - Score < 0.3 → likely fake
    - Score 0.3–0.7 → uncertain, combine with SyncNet/AV-HuBERT for final decision

Example: Score = 0.8 → likely real; Score = 0.2 → likely fake.

Fusion & Evaluation Metrics (Step 5)

Goal: Combine features and define performance metrics.

1. **Feature Fusion Approach:**
    - Concatenate SyncNet scores, AV-HuBERT embeddings, phoneme/text alignment features, and LLM score.
    - Feed combined feature vector into a light classifier (e.g., XGBoost or LightGBM).

   - Alternative: weighted combination or late fusion of separate predictions.

2. **Evaluation Metrics:**
   - **Video-level:**
     - AUC (Area Under ROC Curve)
     - EER (Equal Error Rate)
   - **Segment/frame-level (temporal localization):**
     - IoU (Intersection over Union) between predicted fake segments and ground truth

Notes:

- Video-level metrics check overall detection performance.
- IoU is used only on datasets with **per-frame ground truth** (like LAV-DF and AVLips).
- Fusion strategy ensures all features contribute to final real/fake decision.

Segment-level IoU will be calculated for datasets with per-frame annotations (LAV-DF, AVLips).

Error Taxonomy (Step 6)

Goal: Categorize types of errors or anomalies in lip-sync deepfakes.

1. **Desynchronization Errors:**
   - Lip movements do not match audio (delays, misaligned phonemes/visemes).
2. **Dubbing / Overlay Artifacts:**
   - Audio replaced or edited independently (background noise mismatch, re-recorded speech).
3. **Rhythm / Prosody Anomalies:**
   - Timing, pitch, or intonation inconsistent with natural speech.
4. **Visual Artifacts (Optional for Pilot):**
   - Blurry or jittery lips, occlusions (hand, microphone, hair), sudden head movement.

Notes:

- This taxonomy will guide **analysis of model failures**.
- Helps in refining preprocessing, features, or LLM prompts if errors are frequent.

Error Type	Example Description
Desynchronization	Phonemes not aligned with visemes
Dubbing Artifact	Re-recorded speech over original video
Rhythm / Prosody	Flat or unnatural timing

Privacy & Ethics Note (Step 7)

Goal: Ensure responsible handling of face and voice data.

1. **Face / Voice Handling:**
    - Use data only for research and model development.
    - Avoid storing unnecessary personal information.
    - Crop faces/lips to minimize identity exposure.
2. **Storage Policy:**
    - Store videos, cropped faces, and features in secure, restricted drives.
    - Limit access to team members.
    - Encrypt or anonymize data where possible.
3. **Sharing Constraints:**
    - Do not publicly share raw videos with real people.
    - Share only derived features (embeddings, SyncNet scores) if needed.
    - Comply with dataset licenses (FakeAVCeleb, LAV-DF, etc.).

    All datasets (FakeAVCeleb, LAV-DF, etc.) will be used in accordance with their respective licenses.

Minimal Demo Checklist (Step 8)

Goal: Define success criteria for the first pilot pipeline run and ensure all steps produce verifiable outputs.

1. **Successful Pipeline Run:**
    - Detect and crop face/lips correctly for all pilot clips.
    - Extract all planned features:
        - SyncNet scores
        - AV-HuBERT embeddings
        - Whisper transcripts
        - MFA alignment outputs
    - Generate LLM score for semantic/prosody plausibility.

- Fuse features and output video-level and segment-level predictions (real/fake).

2. **Evidence to Capture:**
    - Save sample cropped lip frames for verification of preprocessing.
    - Export feature vectors for a few clips to check feature extraction.
    - Record video-level predictions for each clip.
    - Optional: visualize SyncNet scores or per-frame predictions to inspect results.
3. **Notes:**
    - This checklist ensures the pipeline runs successfully on the pilot dataset.
    - All steps—from face/lip detection, feature extraction, LLM scoring, to fusion and predictions—should produce outputs.
    - Evidence captured (frames, features, predictions) will serve as proof for debugging and verification.
    - Optional: include a small flowchart showing the pipeline (Preprocessing → Feature Extraction → LLM → Fusion → Prediction) for clarity.