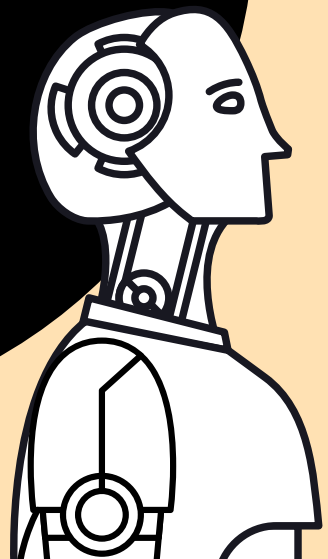
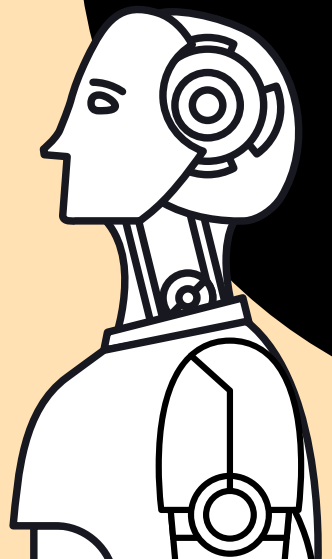
SEED | HACKATHON

BEYOND PIXELS: MULTI-MODAL DEEPFAKE DETECTION FROM COMPRESSED BITSTREAMS AND AUDIO-VISUAL BEHAVIOR WITH CONTINUOUS ADVERSARIAL ADAPTATION



PRESENTATION BY
TEAM **MUGEN**

SAYAN ROY
HARSH RAJ GUPTA
ANGSHUMAN ROY



Problem Statement

Real-Time Deepfake Detection at Social Media Scale

Process millions of 5-second video clips per hour, detect heavily recompressed deepfakes across visual and audio channels, flag partial manipulations with precision, and maintain false positives below 1% through calibrated thresholds and production-grade infrastructure.

Operational Realities

Internet videos arrive multi-compressed through various codec pipelines. Deepfake artifacts hide in the codec domain—motion vectors, DCT coefficients, and quantization parameters. Our system must survive 10× recompression cycles while detecting manipulations across visual content, audio spoofing, and audio-visual synchronization mismatches.

The attack surface is broad: sophisticated GANs, neural voice cloning, face-swap with real audio, and hybrid manipulations demanding multi-modal detection.

Success Criteria

Speed

Clip-level decision ≤ 250ms p95 latency on GPU inference path

Scale

Horizontal scaling via NVDEC hardware decode
+ DeepStream batching + Triton serving

Safety

False positive rate ≤ 1% through temperature scaling and cohort-based thresholds

Training Data Strategy

Visual Video Datasets

FaceForensics++ C23
(H.264 medium compression): 1,000 original videos with multiple manipulation methods—our stable pretraining foundation with controlled compression artifacts for ablation studies.

DFDC:
Diversity of identities, lighting conditions, and realistic "video-fake with real-audio" scenarios that mirror production attack vectors.

Celeb-DF v2:
590 high-quality celebrity deepfakes
Our generalization benchmark, forcing the model beyond training distribution patterns.

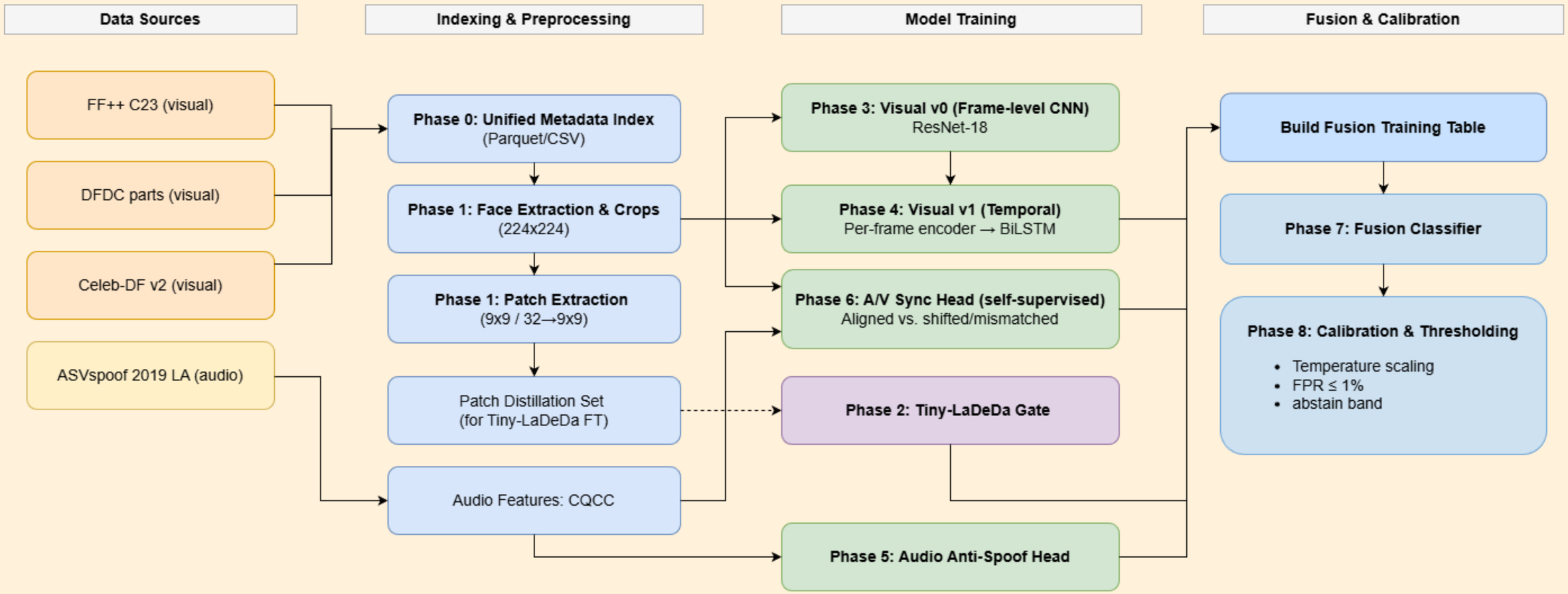
Audio Anti-Spoof

ASVspoof 2019 (LA):
Logical access track with bonafide speech versus synthesized attacks (TTS/voice conversion). We extract CQCC acoustic features as anti-spoof baselines, critical for detecting neural voice cloning.

Patch-Level Gate

Tiny-LaDeDa + WildRF:
9×9 spatial patch scoring with a distilled tiny model: 375× fewer FLOPs than full inference. Perfect for early-stage triage, quickly passing clean content and escalating suspicious regions.

End-to-End System Training



Training Phases

Phase 1: Data Pipeline

1

Indexing and extraction using PyAV and OpenCV.
Built combined clip dataset with face crops and audio chunks from FF++, DFDC, and Celeb-DF.

Phase 2: Patch Gate

2

Integrated Tiny-LaDeDa for ultra-fast triage.
Trained 9×9 patch classifier on WildRF augmented data.

Phase 3: Visual Models

3

Visual v0: Frame encoder pretrained on FF++ C23 and DFDC with compression augmentation.
Visual v1: Temporal head with BiLSTM over frame sequences, Fine-tuned on Celeb-DF for generalization.

Phase 4: Audio & Sync

4

Audio anti-spoof trained on ASVspoof 2019 LA track with CQCC features.
A/V sync head trained with SyncNet-style contrastive learning on aligned and mismatched lip-audio pairs.

Phase 5: Fusion & Calibration

5

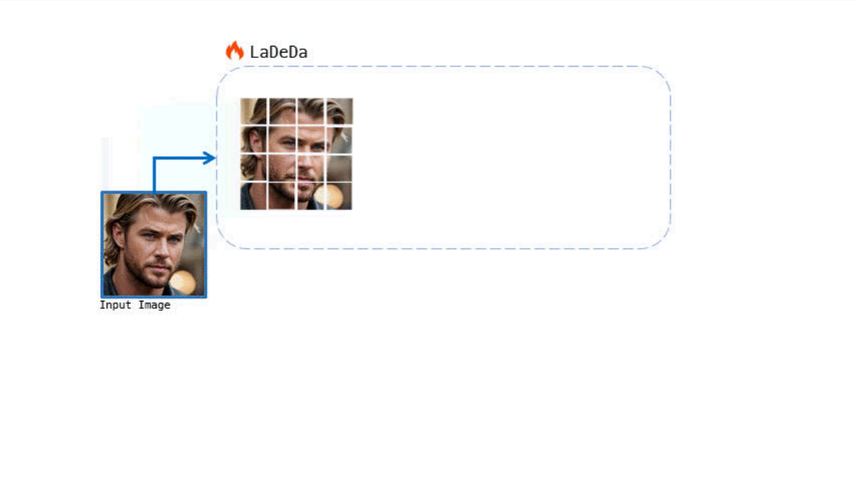
Trained fusion MLP on multi-head outputs.
Applied temperature scaling on validation cohort stratified by codec/bitrate to achieve <1% FPR with abstain band.

Phase 6: Production Deploy

6

Integrated NVDEC, DeepStream, and Triton Inference Server.
Deployed Kafka event pipeline with FastAPI endpoints and monitoring dashboard.

Visual Detection Stack: Three-Tier Approach



Training on C23 and DFDC compression distributions plus codec noise augmentation ensures robustness. The temporal head is critical as it detects inter-frame inconsistencies that survive aggressive recompression where spatial artifacts degrade.

Tiny-LaDeDa Patch Gate

Ultra-cheap triage using 9×9 spatial patches. A distilled 4-layer CNN classifier scores each patch, pools to face and clip scores. Passes clean content instantly and escalates suspicious regions to full detection.

Visual v0: Frame Encoder

ResNet-18 trained on face crops from FF++, DFDC, and Celeb-DF with JPEG compression augmentation. Detects single-frame manipulation artifacts.

Visual v1: Temporal Head

Per-frame embeddings fed through BiLSTM across 16 sampled frames per 5-second clip. Outputs clip-level and per-frame scores, catching temporal warping and flicker that persists through 10× recompression.

Audio & Audio-Visual Consistency Detection

Audio Anti-Spoof Head

Trained on ASVspoof 2019 Logical Access track, distinguishing bonafide speech from text-to-speech and voice conversion attacks. We extract Constant Q Cepstral Coefficients (CQCC) as input features for a lightweight CNN architecture.

This head is essential for catching neural voice cloning deepfakes where the visual content is manipulated but audio is synthesized to match lip movements.

A/V Synchronization Head

Self-supervised contrastive learning inspired by SyncNet architectures. Positive pairs: aligned lip movements and corresponding audio. Negative pairs: time-shifted or identity-mismatched audio. Outputs a "sync score" using contrastive loss or binary cross-entropy.

Critical for DFDC-style attacks where real audio is paired with fake video. Misalignment detection flags these hybrid manipulations that single-modality detectors miss.

Fusion, Partial Labels & Calibration Strategy

1

Feature Fusion

Inputs: visual clip score, patch gate score, audio spoof score, A/V sync score, plus lightweight codec metadata (bitrate, codec type).
Fusion model: compact MLP

2

Label Space

Four-class detection: Real, Video-Fake/Audio-Real, Video-Real/Audio-Fake, Both-Fake.
Enables granular partial manipulation flagging for moderator review.

3

Temperature Scaling

Calibration on real-heavy validation cohort stratified by codec and bitrate.
Optimizes decision thresholds to guarantee FPR ≤ 1% with optional abstain band for borderline cases.

Temperature scaling transforms raw model outputs into calibrated confidence scores. This is critical for production deployment of uncalibrated models may exhibit high accuracy but poor probability estimates, leading to unreliable thresholding and excessive false positives.

End-to-End System Design

1

Ingest & Decode

NVDEC hardware acceleration +
DeepStream batched decode for H.264/H.265/AV1 streams

2

Clip Builder

PyAV extracts 5-sec clips, samples frames, crops faces, and chunks audio

3

Multi-Head Detection

Patch gate (Tiny-LaDeDa), visual(frame+temporal), audio anti-spoof, A/V sync

4

Fusion & Calibration

Temperature-scaled ensemble with FP rate control

5

Serving Layer

Triton Inference Server, Kafka event streams, FastAPI endpoints

Each component is designed for horizontal scalability. Hardware decode offloads compute engines, allowing visual/audio models to focus purely on detection inference. Kafka provides backpressure management and event replay across the pipeline stages.

FastAPI Service & Moderator UI

API Endpoints

/score_clip

Upload file or URL for single clip analysis.
Returns multi-head scores, final decision, and confidence interval.

/score_stream

WebSocket or RTSP proxy for real-time stream monitoring with frame-by-frame scoring.

/metrics

Prometheus-compatible metrics endpoint for throughput, latency, GPU utilization, and error rates.

/explanations

Per-head score breakdown with patch heatmaps and frame-level temporal plots for transparency.

Moderator Dashboard

- **Clip Player:** 5-second looping video with synchronized audio playback.
- **Per-Head Visualization:** Bar charts for Visual, Patch Gate, Audio, and Sync scores with color-coded confidence.
- **Final Decision:** Clear verdict with calibrated probability and manipulation type tag.
- **Partial Manipulation Tags:** "Video-Fake / Audio-Real" granular labels for review prioritization.
- **Abstain Flags:** Borderline cases requiring human review before policy enforcement.
- **Metadata Chips:** Codec type, bitrate, resolution, and recompression estimates.

OpenAPI/Swagger auto-documentation provides interactive testing for judges. JWT authentication secures admin actions and sensitive operations.

Scaling to Millions of Clips Per Hour

Throughput Infrastructure

- **NVDEC Hardware Decode**
  GPU-accelerated decode independent of CUDA compute cores. Supports H.264, H.265, and AV1. Handles multi-hundreds of FPS per GPU for H.264 streams.
- **DeepStream Batching**
  Multi-stream batching with zero-copy memory pipelines. Efficiently aggregates frames across concurrent video streams for batch inference.
- **Triton Inference Server**
  Dynamic batching across visual, audio, sync, and fusion models. HTTP and gRPC endpoints with model versioning and A/B testing support.
- **Kafka Event Streaming**
  Topic-based architecture: ingest → scored → policy enforcement. Provides backpressure handling, event replay, and pipeline decoupling.

350+	64
FPS per GPU	Batch Size
NVDEC H.264 decode throughput	Optimal Triton inference batching

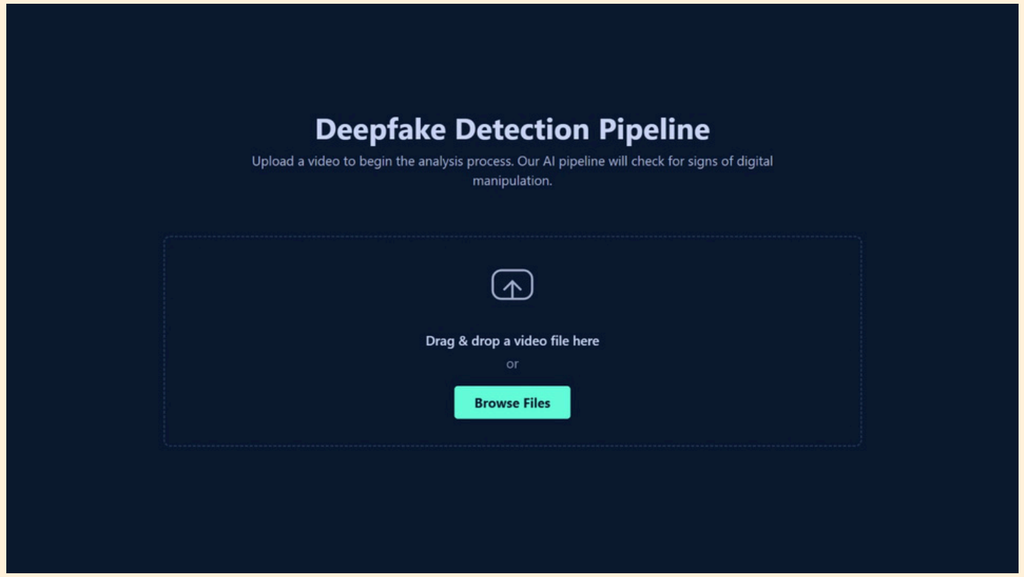
10K+

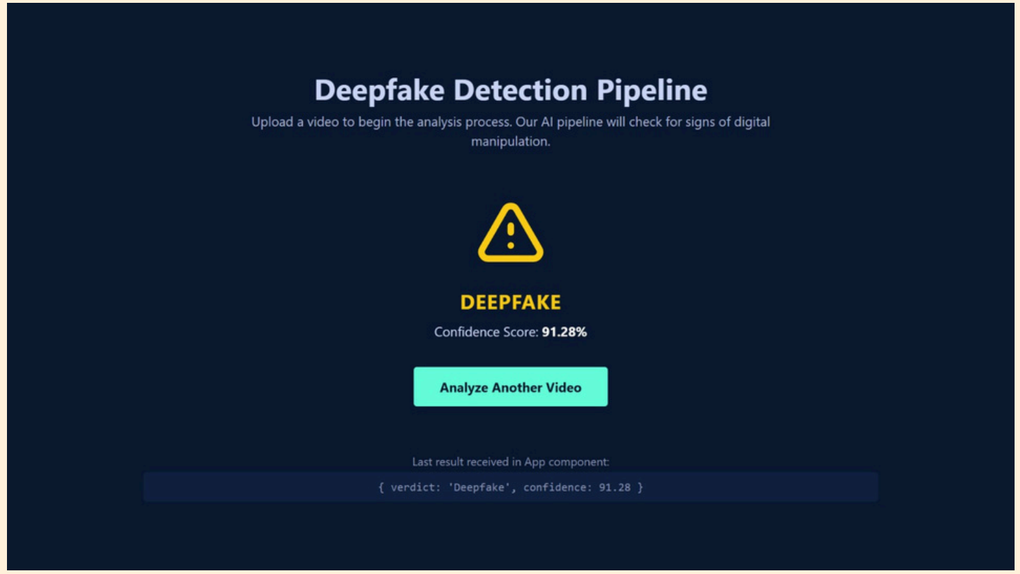
Clips/Hour

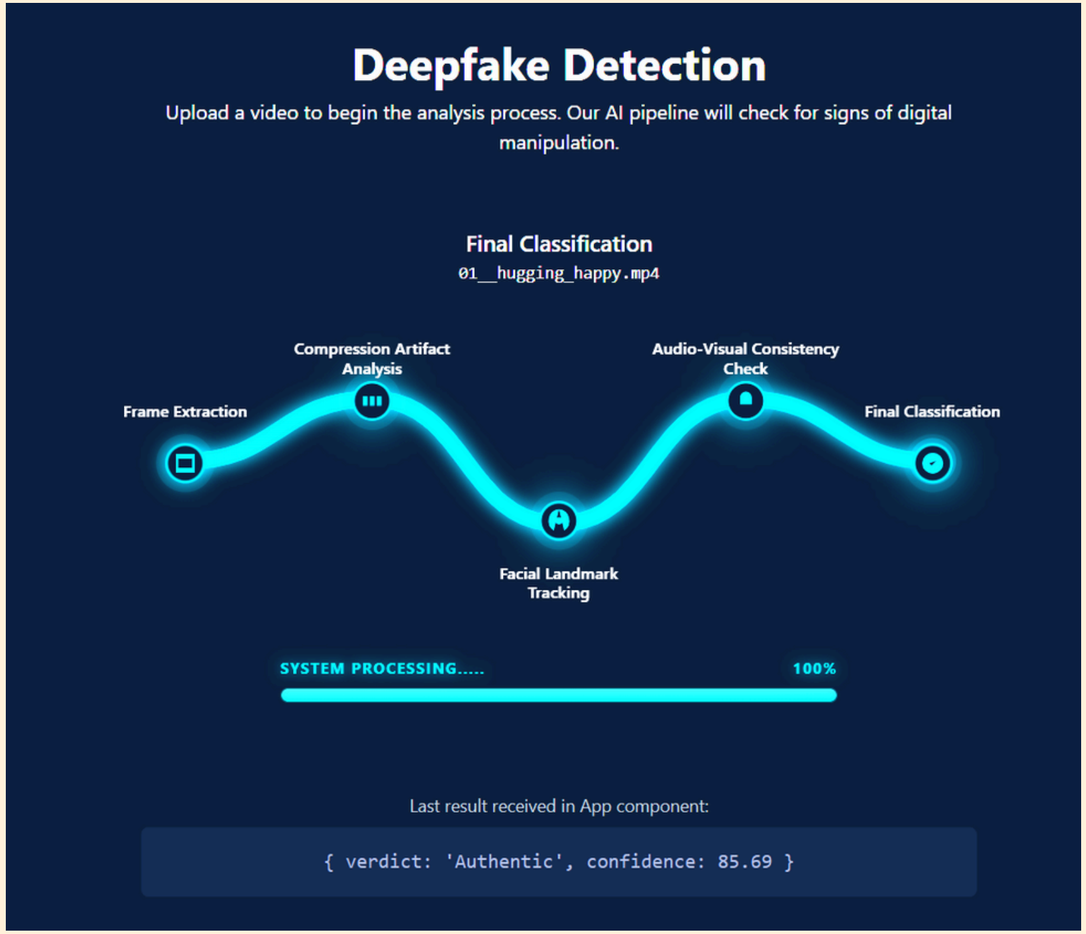
Per GPU with full pipeline

Horizontal scaling: Add GPU nodes behind load balancer. Linear throughput growth with independent decode and inference lanes.

User Interface







THANK YOU