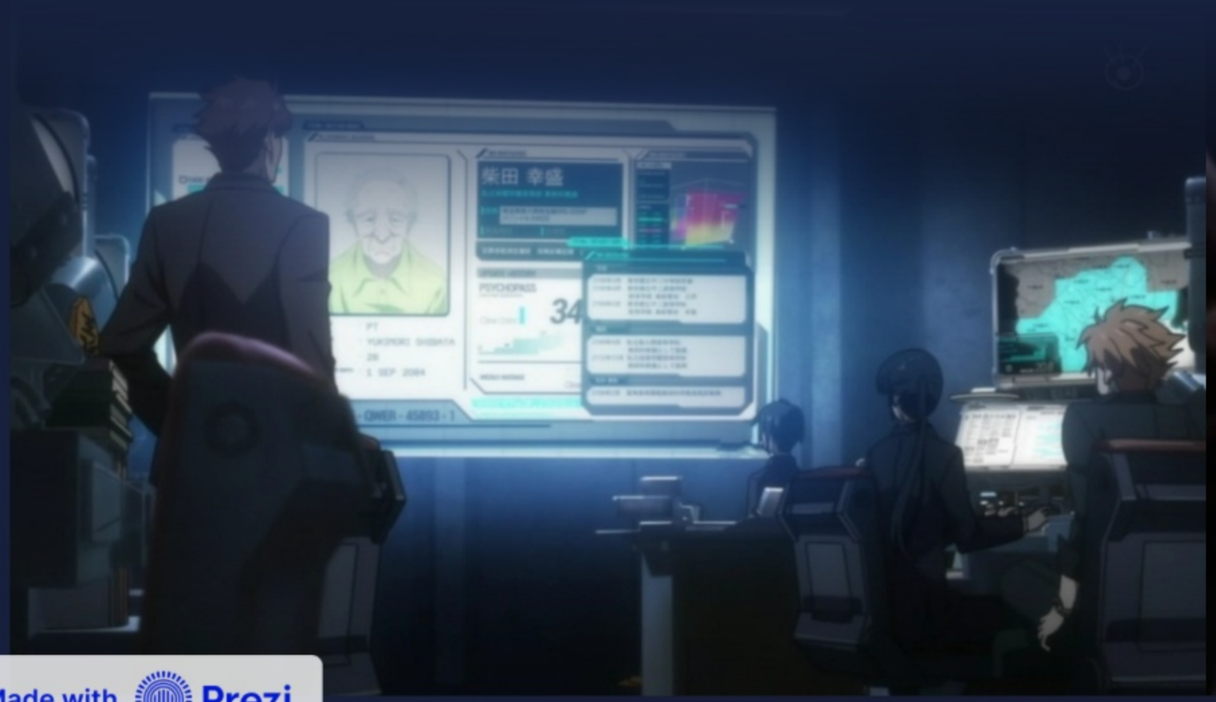
Deepfake Video Detection using Vision Temporal Transformer and Multi-Head Attention

Leveraging AI technologies for accurate detection and classification of manipulated videos to combat misinformation.

Deepfake Video Detection using Vision Temporal Transformer and Multi-Head Attention

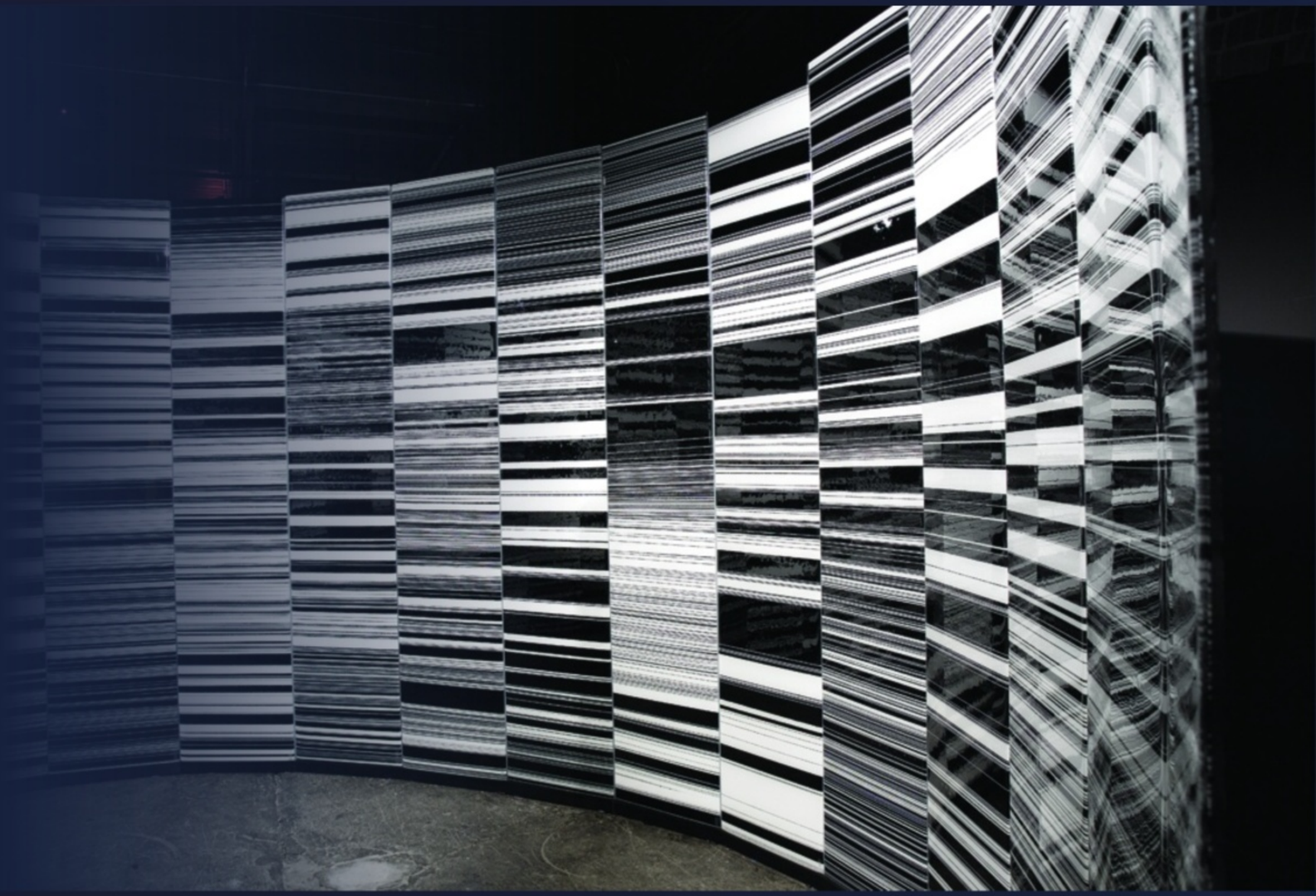
Deepfake Video Detection using Vision Temporal Transformer and Multi-Head Attention

This presentation introduces an advanced AI-based system designed to detect deepfake videos. By leveraging the Vision Temporal Transformer and Multi-Head Attention mechanisms, the system analyzes facial features, motion patterns, and the temporal relationships between frames, enhancing the accuracy of deepfake detection.



What is Deepfake?

Deepfake videos utilize advanced deep learning technology to create or alter video content, making them appear authentic. By manipulating facial features, expressions, and movements, these videos can achieve a level of realism that makes them difficult to distinguish from genuine footage.



Threat to Digital Media Trust

Deepfake technology reduces trust in digital videos.
When fake videos look real, people start to doubt what they see online. This makes it harder to trust real and authentic content

Spread of Misinformation

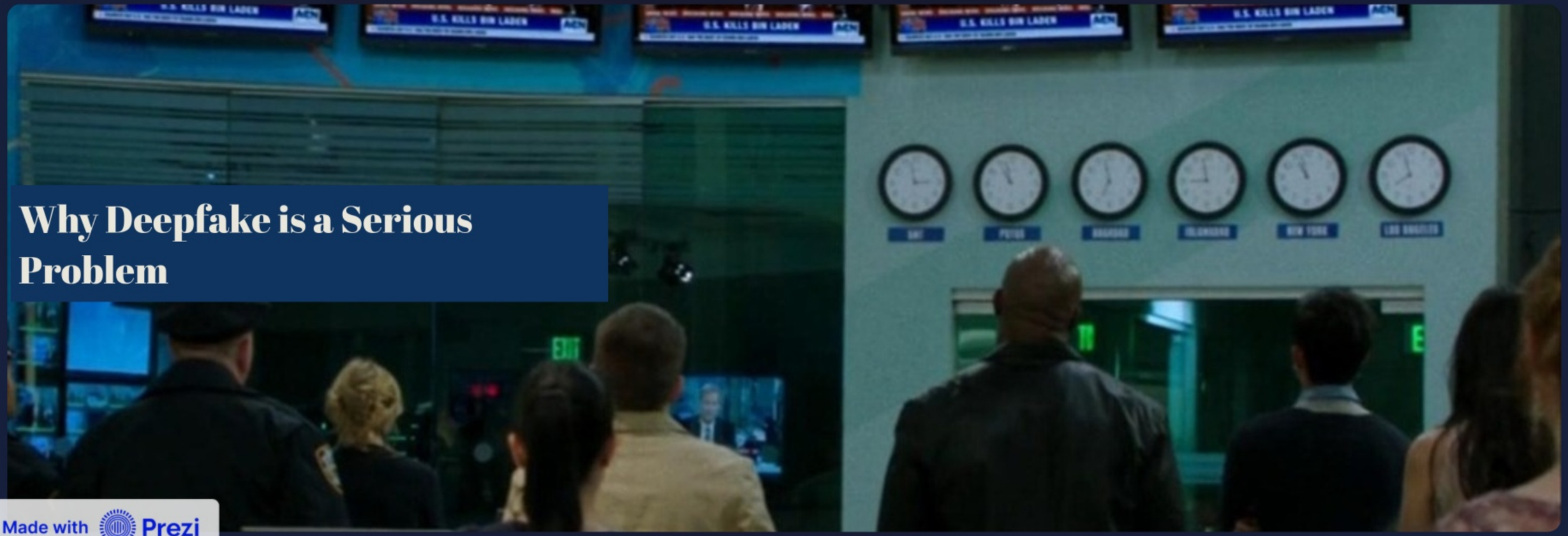
Misinformation spread through deepfake videos can have severe consequences, influencing public opinion and decision-making, often resulting in social unrest or misinformation campaigns.

Impersonation of Individuals

Deepfake videos can impersonate real individuals, leading to identity theft, reputational damage, and even financial loss for those targeted, raising serious ethical and legal implications.

Difficulty of Detection

The sophistication of deepfake technology makes it extremely challenging for viewers and even advanced detection systems to differentiate between real and manipulated content, complicating efforts to combat this issue.



Why Deepfake is a Serious Problem

Threat to Digital Media Trust

Deepfake technology reduces trust in digital videos.
When fake videos look real, people start to doubt what they see online.
This makes it harder to trust real and authentic content

Spread of Misinformation

Misinformation spread through deepfake videos can have severe consequences, influencing public opinion and decision-making, often resulting in social unrest or misinformation campaigns.

Impersonation of Individuals

Deepfake videos can impersonate real individuals, leading to identity theft, reputational damage, and even financial loss for those targeted, raising serious ethical and legal implications.

Difficulty of Detection

The sophistication of deepfake technology makes it extremely challenging for viewers and even advanced detection systems to differentiate between real and manipulated content, complicating efforts to combat this issue.

Limitations of Existing Detection Methods

Frame-by-Frame Analysis

Traditional detection methods primarily focus on analyzing each frame of a video in isolation. This frame-by-frame analysis does not consider the motion dynamics or relationships between consecutive frames, which are crucial for identifying deepfakes effectively.

Neglecting Temporal Relationships

By not considering the temporal relationships between frames, traditional methods miss significant clues that could indicate manipulation. Temporal inconsistencies can manifest as unnatural movements or abrupt changes in facial expressions, which are critical for detection.

Overlooking Motion Patterns

Many detection methods miss important motion details. Deepfake videos often contain unnatural movements between frames.
These issues cannot be seen when each frame is analyzed alone

Frame-by-Frame Analysis

Traditional detection methods primarily focus on analyzing each frame of a video in isolation. This frame-by-frame analysis does not consider the motion dynamics or relationships between consecutive frames, which are crucial for identifying deepfakes effectively.

Neglecting Temporal Relationships

By not considering the temporal relationships between frames, traditional methods miss significant clues that could indicate manipulation. Temporal inconsistencies can manifest as unnatural movements or abrupt changes in facial expressions, which are critical for detection.

Overlooking Motion Patterns

Many detection methods miss important motion details.
Deepfake videos often contain unnatural movements between frames.
These issues cannot be seen when each frame is analyzed alone

Project Goal

Project Goal

This project aims to build a reliable deepfake detection system.
It combines visual features, time-based analysis, and motion information.
This helps the system accurately distinguish between real and fake videos

Proposed Framework Overview

A structured multi-stage framework designed to enhance deepfake detection accuracy.

Video Frame Extraction

Decomposing videos into individual frames to enable detailed analysis.

Motion-Based Frame Selection

Applying Optical Flow to identify and retain the most informative frames.

Face Detection and Preprocessing

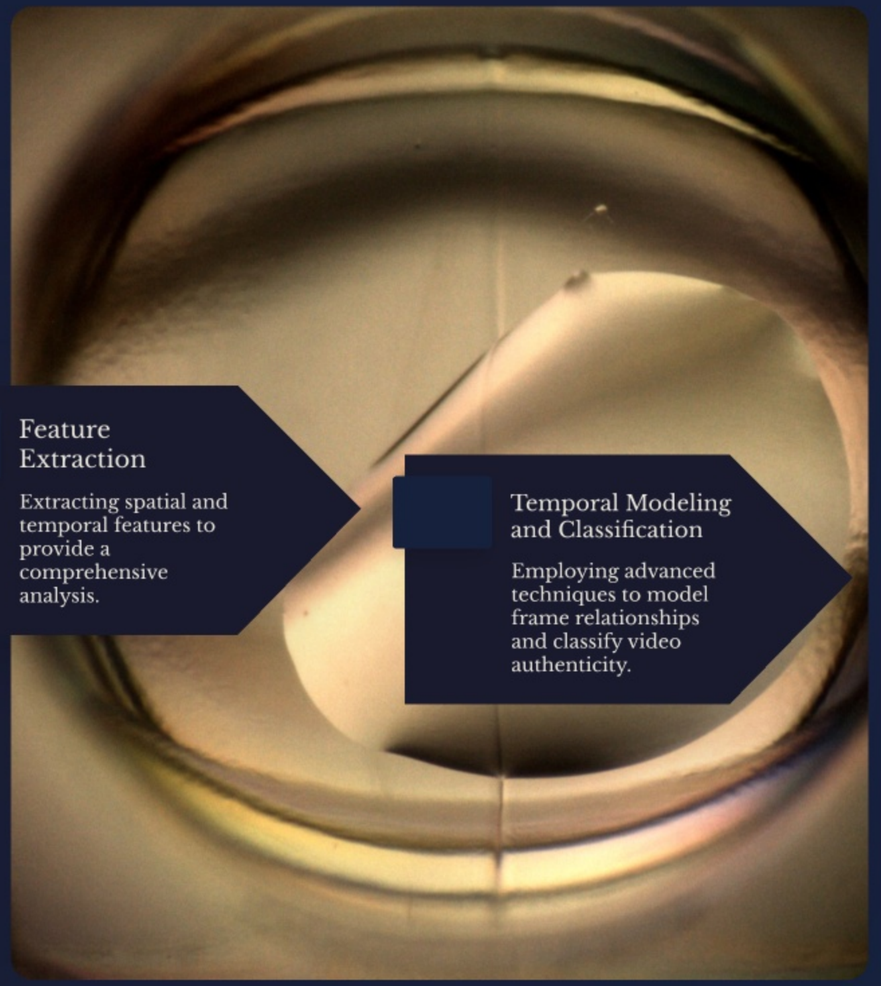
Utilizing a YOLO-based model for precise facial region detection and normalization.

Feature Extraction

Extracting spatial and temporal features to provide a comprehensive analysis.

Temporal Modeling and Classification

Employing advanced techniques to model frame relationships and classify video authenticity.



Video Frame Extraction

Decomposing videos into individual frames to enable detailed analysis.

Motion-Based Frame Selection

Applying Optical Flow to identify and retain the most informative frames.

Face Detection and Preprocessing

Utilizing a YOLO-based model for precise facial region detection and normalization.

Feature Extraction

Extracting spatial and temporal features to provide a comprehensive analysis.

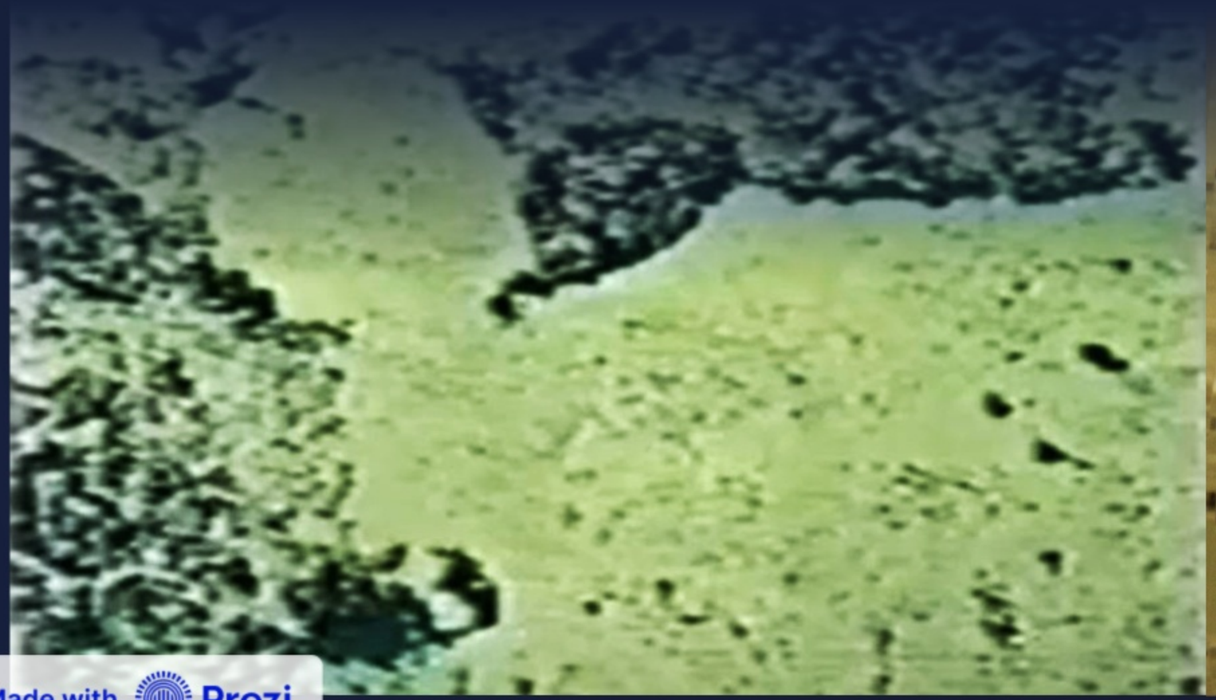
Temporal Modeling and Classification

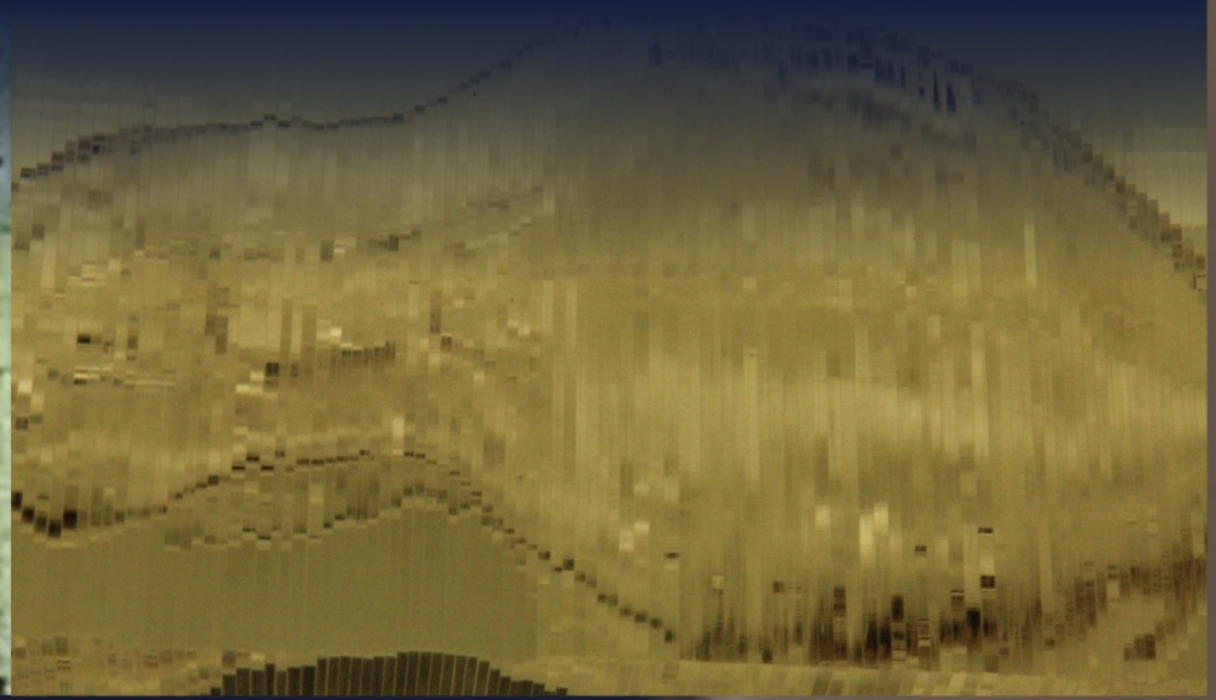
Employing advanced techniques to model frame relationships and classify video authenticity.

Motion-Based Frame Selection

Motion-Based Frame Selection

In the motion-based frame selection process, videos are decomposed into individual frames, enabling detailed analysis of motion dynamics. Optical Flow techniques are employed to evaluate the movement between these frames, ensuring that only the most relevant frames are retained for further processing, thereby optimizing computational efficiency and enhancing detection accuracy.





Face Detection and Preprocessing

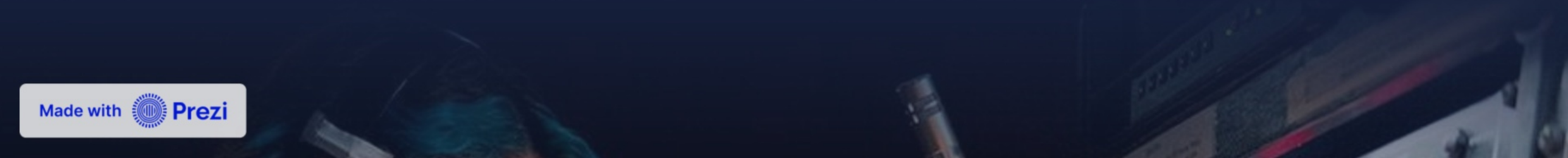
Utilizing a YOLO-based face detection model enables precise localization of facial regions within the selected frames. The detected faces are then cropped, resized, and normalized to maintain consistency and enhance the quality of input data for the subsequent detection model, ensuring accurate analysis of potential manipulations.

Feature Extraction

Residual Frame Generation

In the deepfake detection process, residual frames are crucial for identifying discrepancies between consecutive video frames. By calculating the difference between frames, this technique effectively brings to light any subtle anomalies that may indicate manipulation, helping to improve detection accuracy.

Residual Frame Generation



Residual Frame Generation

In the deepfake detection process, residual frames are crucial for identifying discrepancies between consecutive video frames. By calculating the difference between frames, this technique effectively brings to light any subtle anomalies that may indicate manipulation, helping to improve detection accuracy.

Feature Extraction

Spatial Features from Facial Frames

Spatial features are extracted by analyzing the facial frames, capturing essential characteristics such as facial landmarks, textures, and expressions that are crucial for identifying authenticity in videos. These features help in delineating the subtle differences between real and manipulated visuals.

Temporal Features from Residual Frames

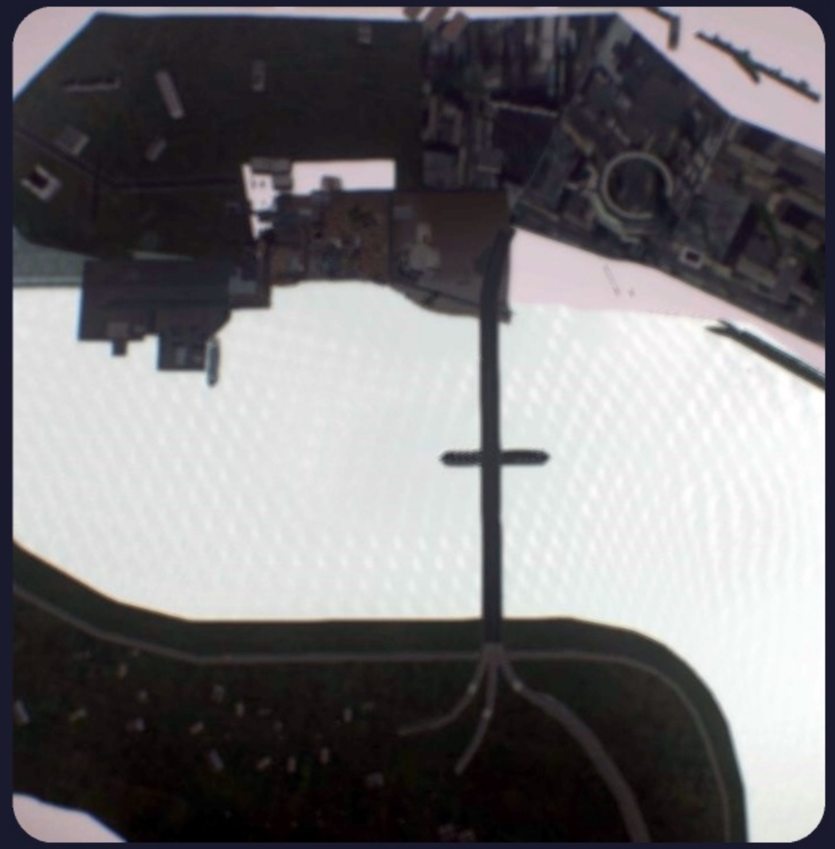
Temporal features derive from the differences between residual frames, revealing the motion information across time. These features identify inconsistencies and unnatural movements that may indicate manipulation, enhancing the detection process by leveraging the dynamic aspects of video content.

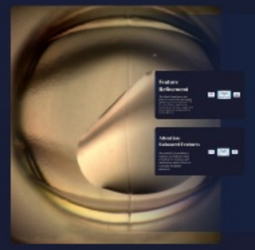
Spatial Features from Facial Frames

Spatial features are extracted by analyzing the facial frames, capturing essential characteristics such as facial landmarks, textures, and expressions that are crucial for identifying authenticity in videos. These features help in delineating the subtle differences between real and manipulated visuals.

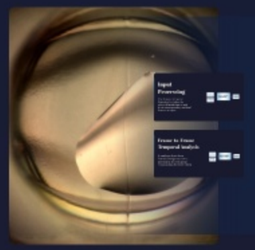
Temporal Features from Residual Frames

Temporal features derive from the differences between residual frames, revealing the motion information across time. These features identify inconsistencies and unnatural movements that may indicate manipulation, enhancing the detection process by leveraging the dynamic aspects of video content.





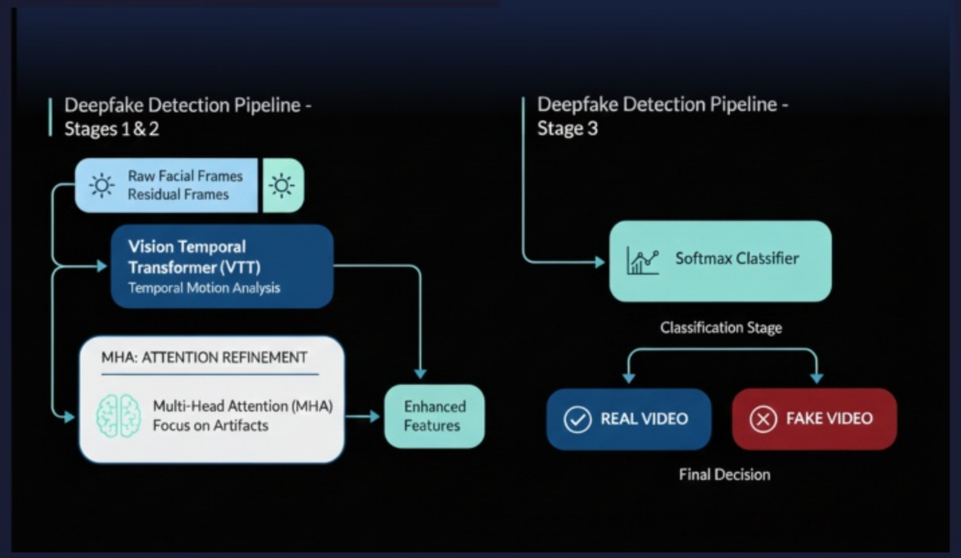
Multi-Head Attention (MHA)

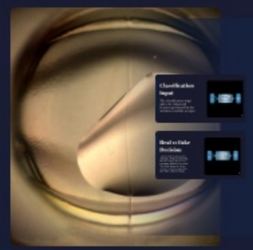


Vision Temporal Transformer (VTT)

System Architecture: Feature Refinement and Final Classification.

The system operates as an integrated pipeline that starts by extracting temporal motion patterns via the Vision Temporal Transformer (VTT), then refines these features using Multi-Head Attention (MHA) to pinpoint specific deepfake artifacts, and finally passes the enhanced data through a Softmax classifier to deliver a definitive "Real" or "Fake" verdict.





Classification Stage

(MHA)

(VTT)

System Architecture: Feature Refinement and Final Classification.

The system operates as an integrated pipeline that starts by extracting temporal motion patterns via the Vision Temporal Transformer (VTT), then refines these features using Multi-Head Attention (MHA) to pinpoint specific deepfake artifacts, and finally passes the enhanced data through a Softmax classifier to deliver a definitive "Real" or "Fake" verdict.

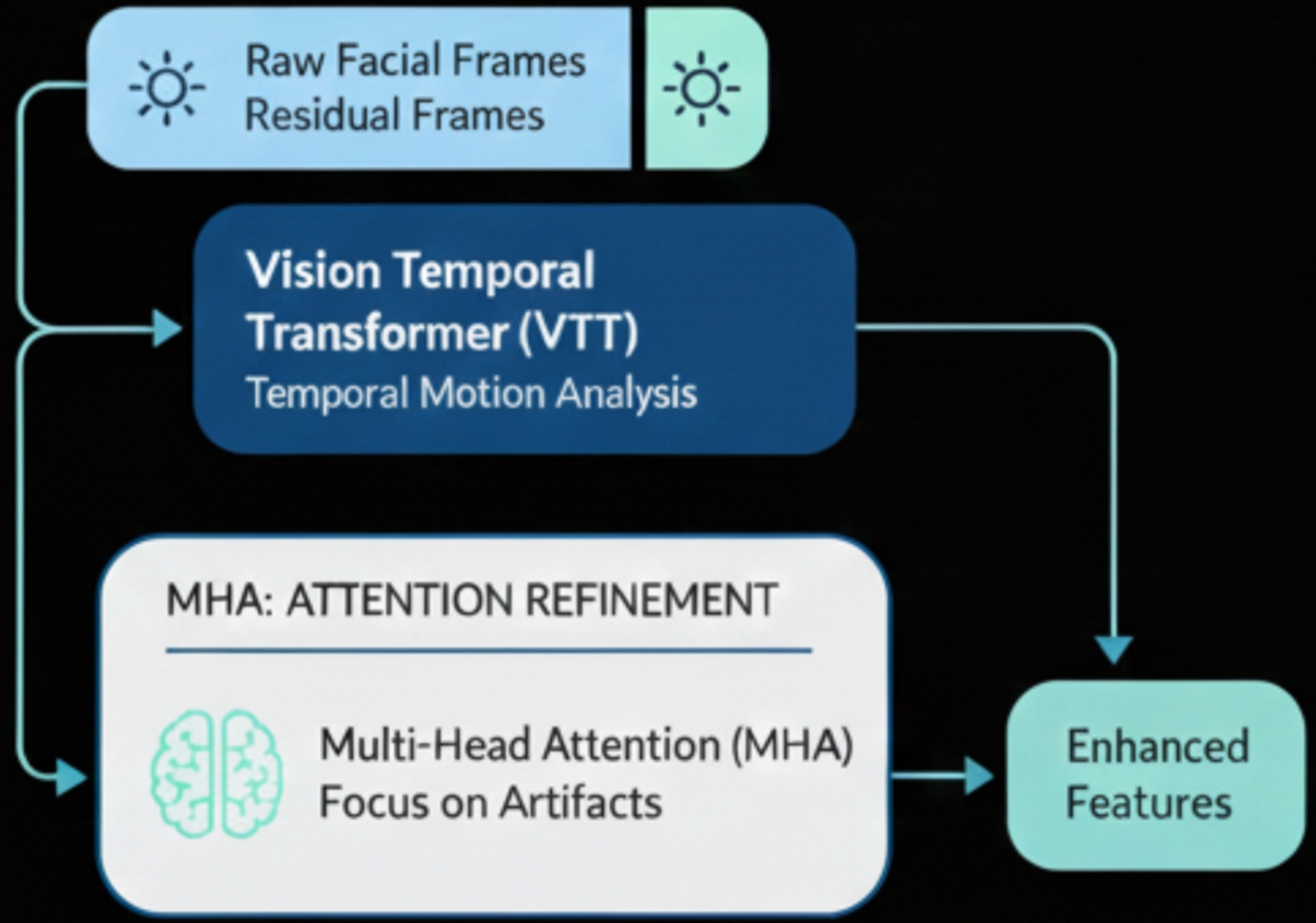
Deepfake Detection Pipeline - Stages 1 & 2

Deepfake Detection Pipeline - Stage 3

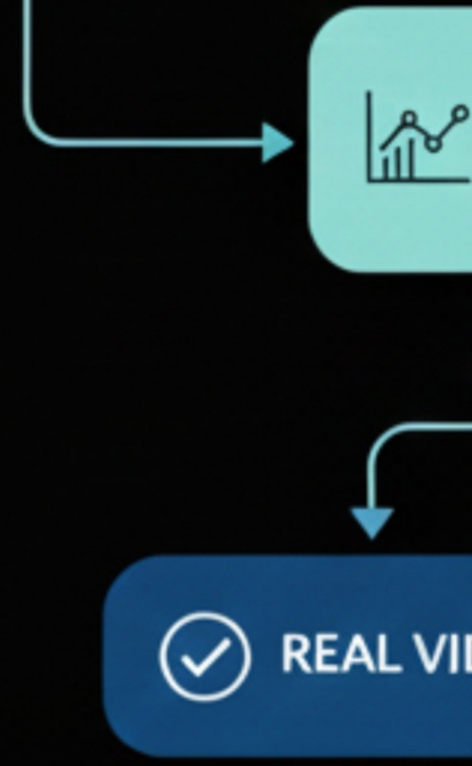
System Architecture: Feature Refinement and Final Classification.

The system operates as an integrated pipeline that starts by extracting temporal motion patterns via the Vision Temporal Transformer (VTT), then refines these features using Multi-Head Attention (MHA) to pinpoint specific deepfake artifacts, and finally passes the enhanced data through a Softmax classifier to deliver a definitive "Real" or "Fake" verdict.

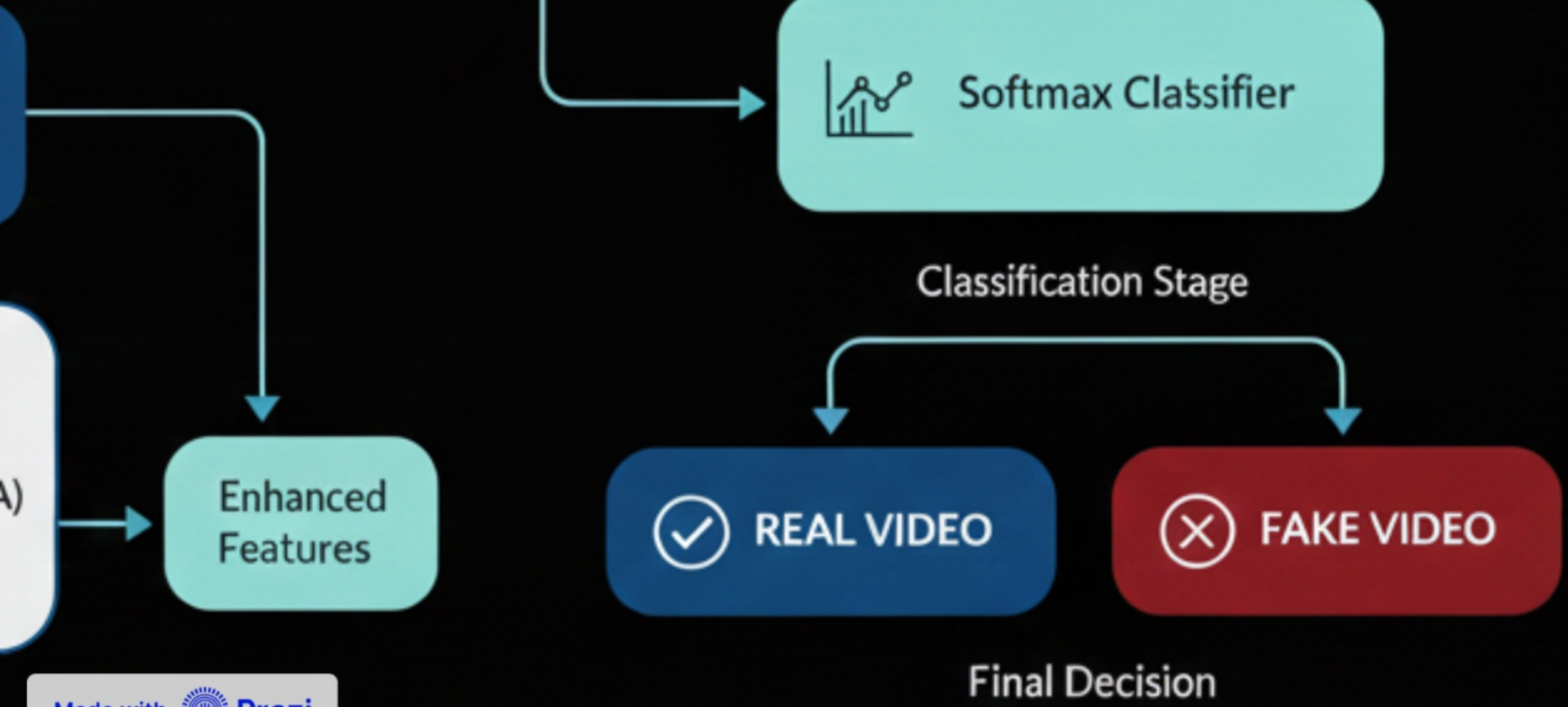
Deepfake Detection Pipeline - Stages 1 & 2



Deepfake Dete... Stage 3



Deepfake Detection Pipeline - Stage 3

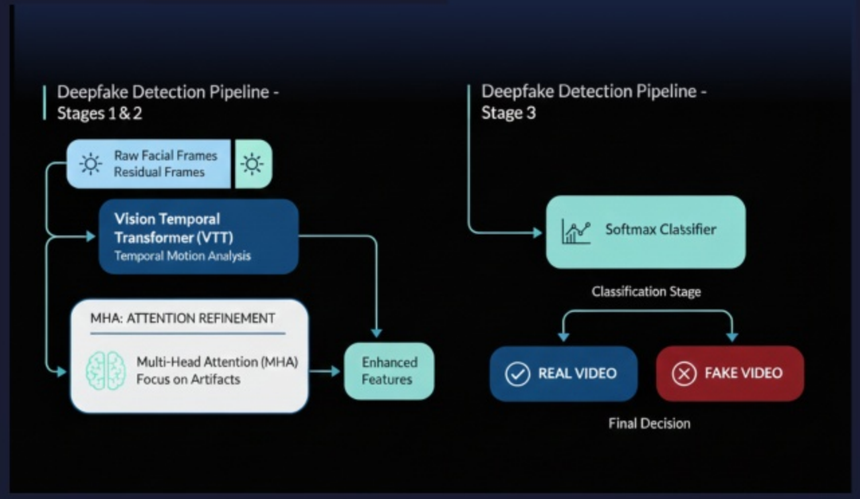


Multi-Head Attention (MHA)

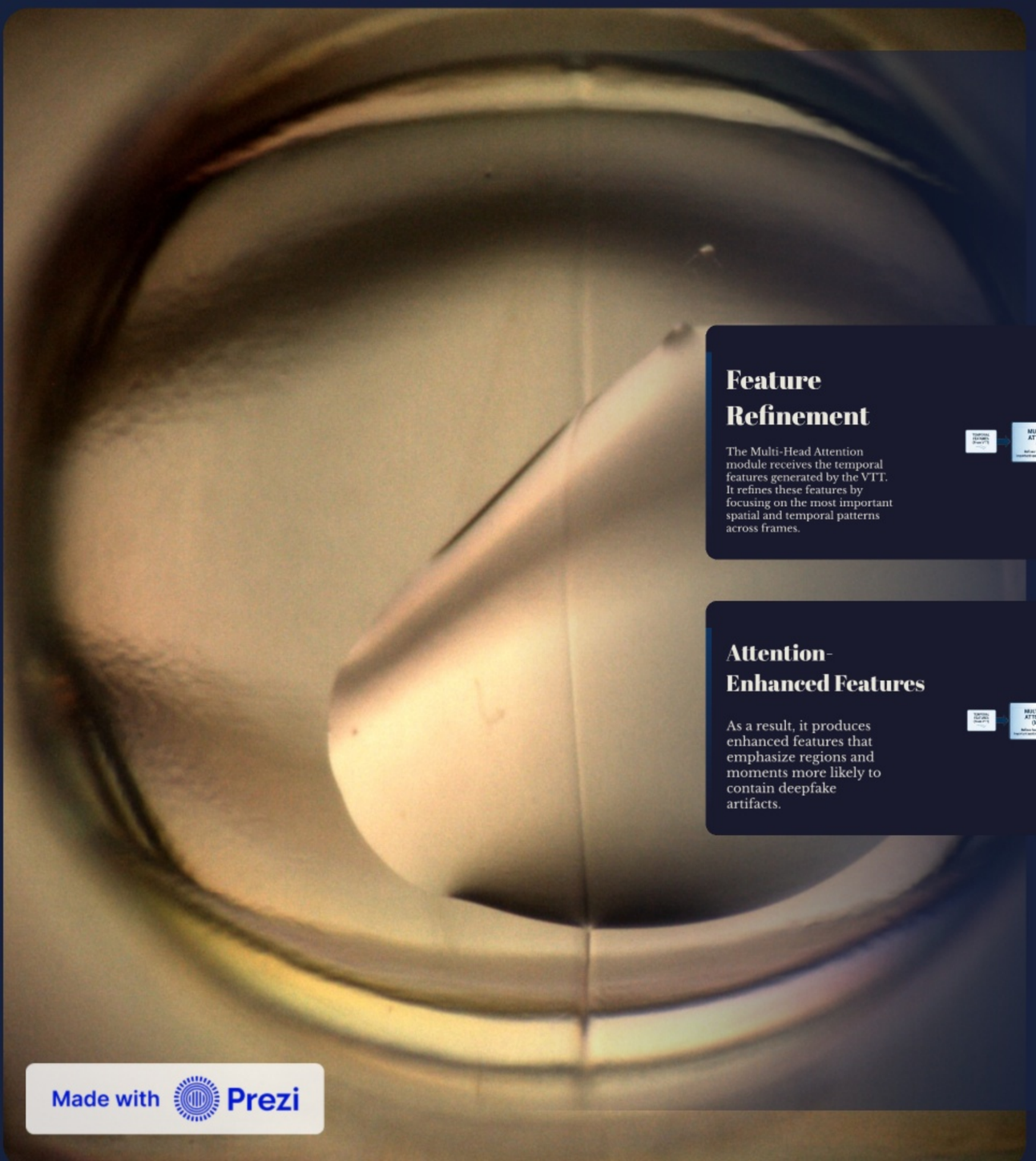
Vision Temporal Transformer (VTT)

System Architecture: Feature Refinement and Final Classification.

The system operates as an integrated pipeline that starts by extracting temporal motion patterns via the Vision Temporal Transformer (VTT), then refines these features using Multi-Head Attention (MHA) to pinpoint specific deepfake artifacts, and finally passes the enhanced data through a Softmax classifier to deliver a definitive "Real" or "Fake" verdict.



Classification Stage



Feature Refinement

The Multi-Head Attention module receives the temporal features generated by the VTT. It refines these features by focusing on the most important spatial and temporal patterns across frames.

Attention-Enhanced Features

As a result, it produces enhanced features that emphasize regions and moments more likely to contain deepfake artifacts.

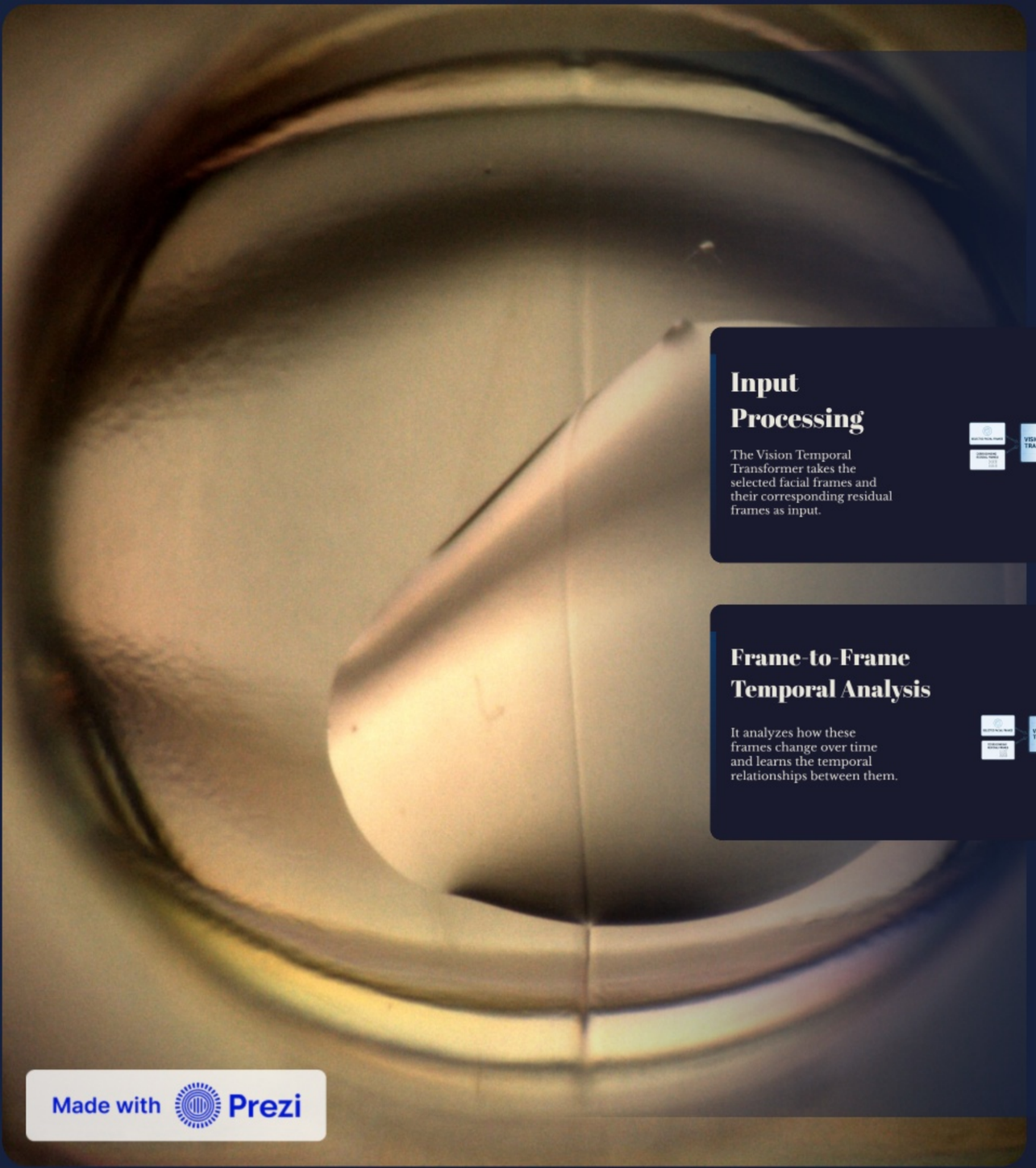
Multi-Head Attention (MHA)

Feature Refinement

The Multi-Head Attention module receives the temporal features generated by the VTT. It refines these features by focusing on the most important spatial and temporal patterns across frames.

Attention-Enhanced Features

As a result, it produces enhanced features that emphasize regions and moments more likely to contain deepfake artifacts.



Input Processing

The Vision Temporal Transformer takes the selected facial frames and their corresponding residual frames as input.

Frame-to-Frame Temporal Analysis

It analyzes how these frames change over time and learns the temporal relationships between them.

Vision Temporal Transformer (VTT)

Input Processing

The Vision Temporal Transformer takes the selected facial frames and their corresponding residual frames as input.

Frame-to-Frame Temporal Analysis

It analyzes how these frames change over time and learns the temporal relationships between them.

Classification Stage

Classification Input

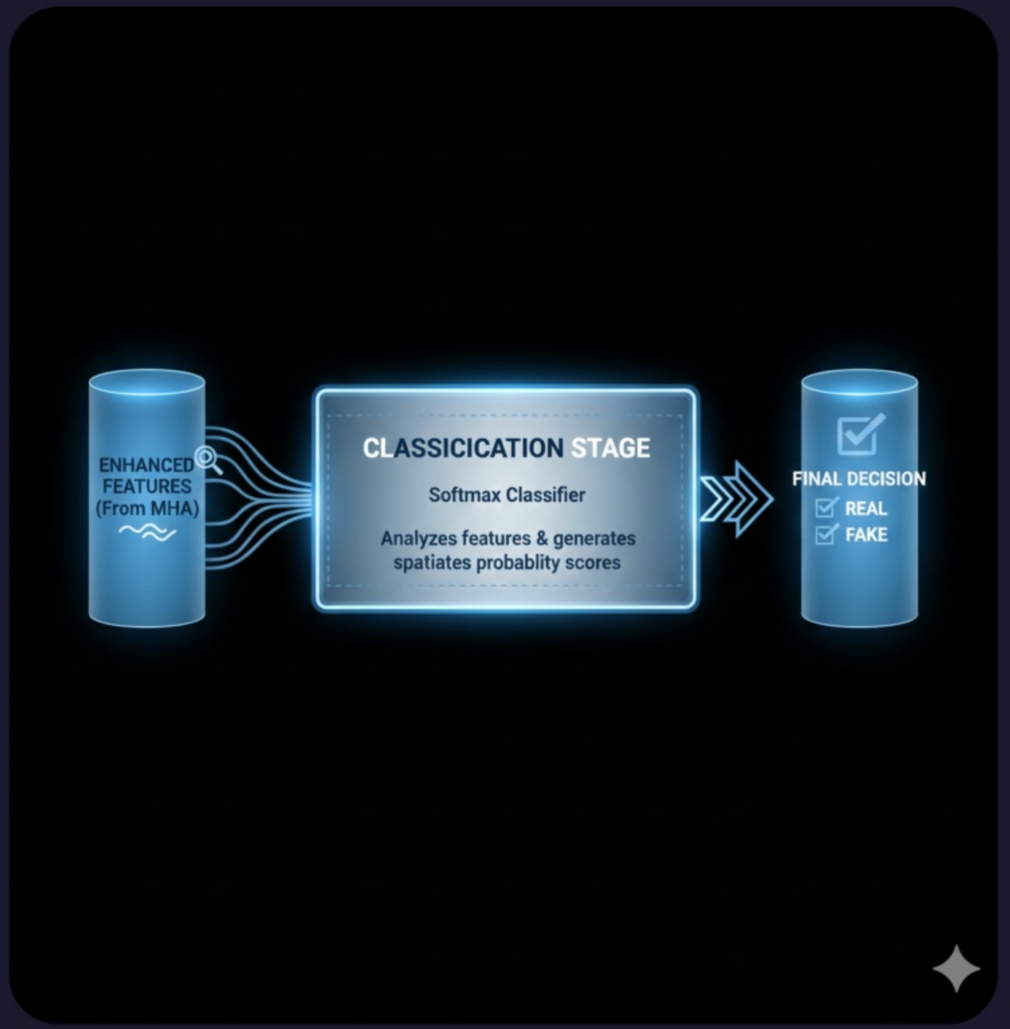
The classification stage takes the enhanced features produced by the attention module as input.

Real vs Fake Decision

These features are passed through a Softmax classifier, which analyzes them and generates probability scores. The final output is a clear decision indicating whether the video is Real or Fake.

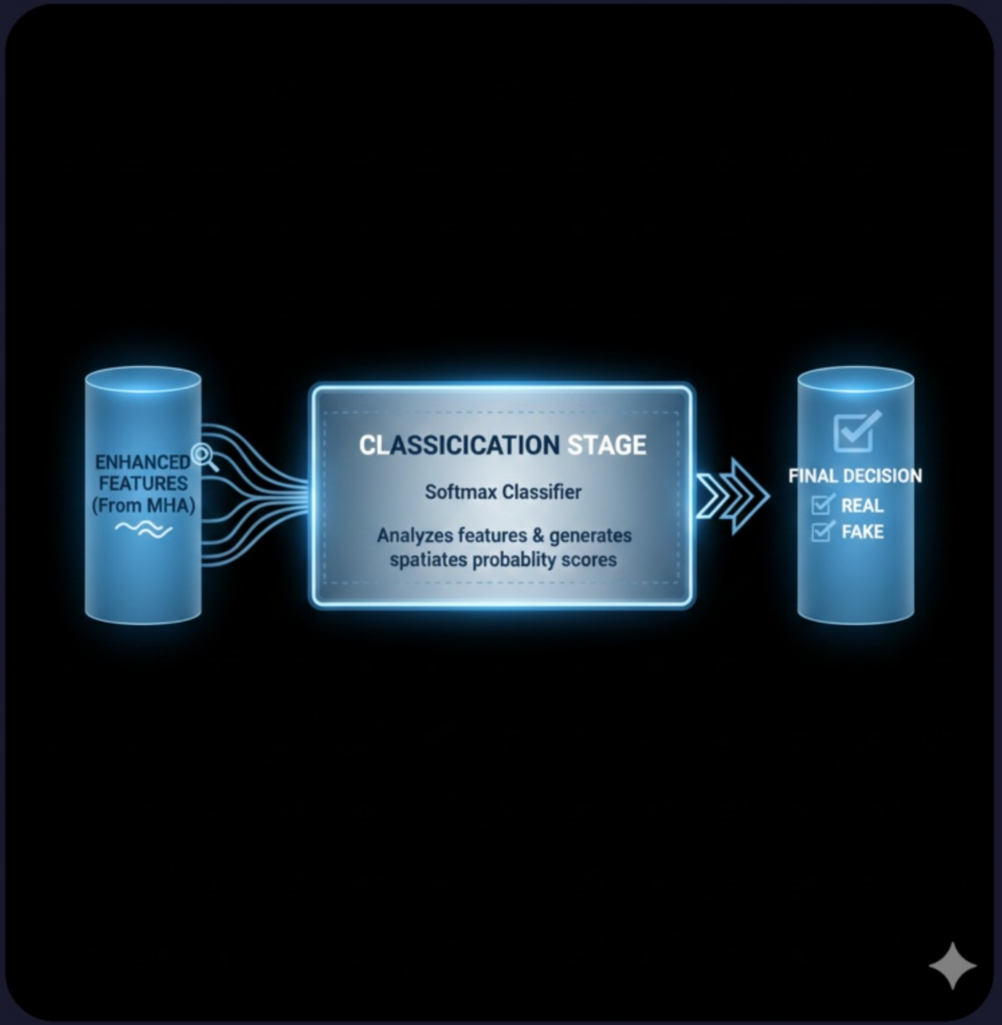
Classification Input

The classification stage takes the enhanced features produced by the attention module as input.



Real vs Fake Decision

These features are passed through a Softmax classifier, which analyzes them and generates probability scores. The final output is a clear decision indicating whether the video is Real or Fake.



Experimental Results

Model Evaluation Setup

System Implementation

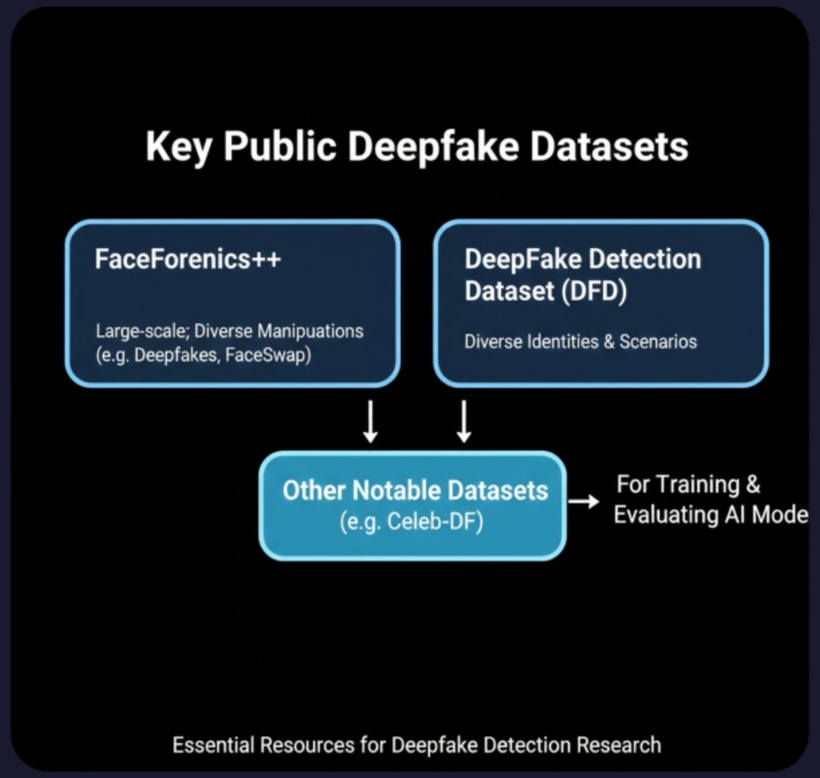
Dataset and Evaluation Process

The system is evaluated using real and manipulated video samples as input. These samples are collected from public datasets such as **FaceForensics**++ and the **DeepFake Detection** dataset.
The output of this setup is a trained and evaluated model that can reliably classify unseen videos.

Model Evaluation Setup

Dataset and Evaluation Process

The system is evaluated using real and manipulated video samples as input. These samples are collected from public datasets such as **FaceForensics**++ and the **DeepFake Detection** dataset.
The output of this setup is a trained and evaluated model that can reliably classify unseen videos.



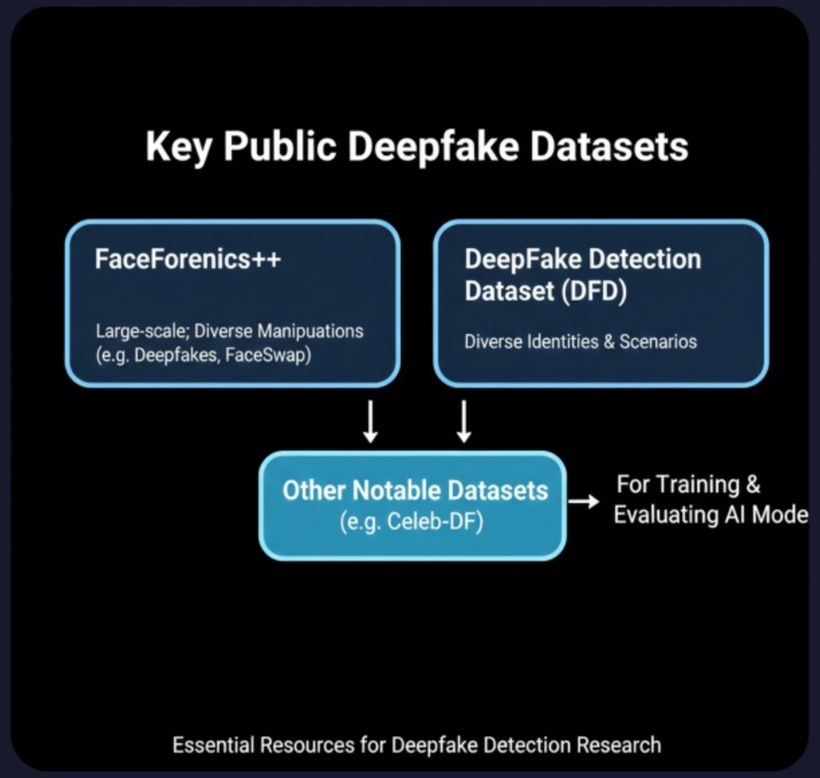
Detection Accuracy Results

The trained model is tested using unseen video samples.
The evaluation process measures accuracy and classification performance.
The output shows an overall accuracy of approximately 85%, indicating strong and balanced detection results.

Experimental Results

Detection Accuracy Results

The trained model is tested using unseen video samples.
The evaluation process measures accuracy and classification performance.
The output shows an overall accuracy of approximately 85%, indicating strong and balanced detection results.



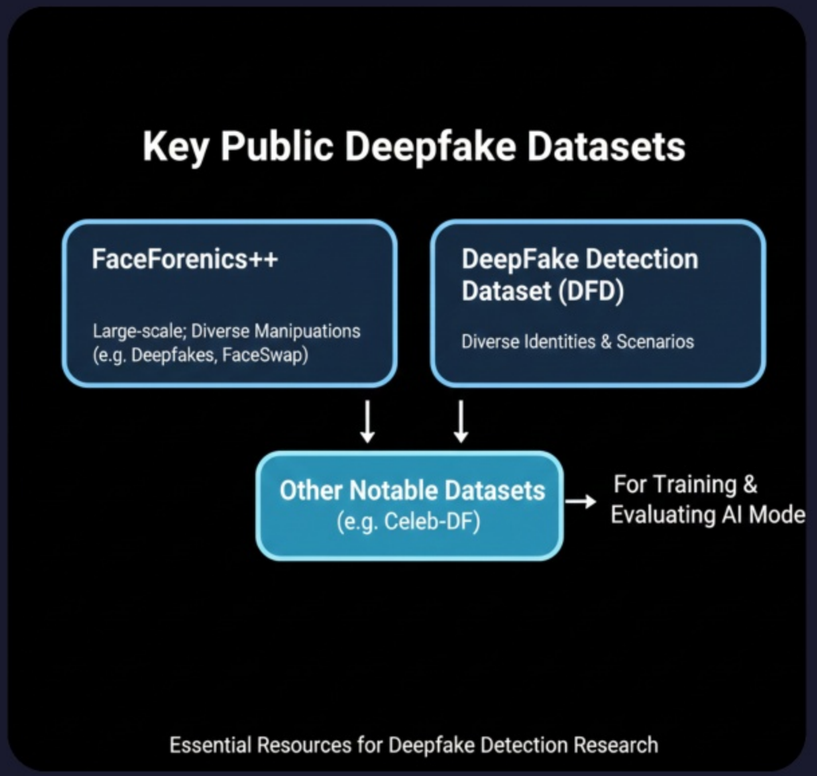
User Interaction

The system receives video input from users through a web or mobile application. The backend server processes the uploaded video using the proposed detection pipeline.
The output is the final detection result, which is returned to the user as Real or Fake with a confidence score.

System Implementation

User Interaction

The system receives video input from users through a web or mobile application.
The backend server processes the uploaded video using the proposed detection pipeline.
The output is the final detection result, which is returned to the user as Real or Fake with a confidence score.



Conclusion

By taking spatial, temporal, and motion-based features as input, the proposed system effectively analyzes deepfake videos.
Through the integration of VTT and Multi-Head Attention, the system produces reliable detection results.
The final output demonstrates improved accuracy in distinguishing real videos from manipulated ones.

Future Work

Future development will focus on improving the system's performance and usability. This includes training the model on larger and more diverse datasets to enhance generalization.
Optimization techniques will be applied to reduce processing time and improve efficiency.
The system will be extended to support real-time deepfake video detection, enabling instant analysis of live or streamed video content.
In addition, future work will include adding image-based deepfake detection, allowing the system to detect manipulated images as well as videos.
Further improvements may also focus on enhancing the user interface and providing more detailed output analysis.

Conclusion

By taking spatial, temporal, and motion-based features as input, the proposed system effectively analyzes deepfake videos.
Through the integration of VTT and Multi-Head Attention, the system produces reliable detection results.
The final output demonstrates improved accuracy in distinguishing real videos from manipulated ones.

Future

Future development will
This includes training the
generalization.
Optimization techniques
efficiency.
The system will be extend
instant analysis of live or
In addition, future work w
allowing the system to de
Further improvements m
providing more detailed

Conclusion

By taking spatial, temporal, and motion-based features as input, the proposed system effectively analyzes deepfake videos.
Through the integration of VTT and Multi-Head Attention, the system produces reliable detection results.
The final output demonstrates improved accuracy in distinguishing real videos from manipulated ones.

tion-based features
ively analyzes

d Multi-Head
ble detection results.
oved accuracy in
ipulated ones.

Future Work

Future development will focus on improving the system's performance and usability. This includes training the model on larger and more diverse datasets to enhance generalization.
Optimization techniques will be applied to reduce processing time and improve efficiency.
The system will be extended to support real-time deepfake video detection, enabling instant analysis of live or streamed video content.
In addition, future work will include adding image-based deepfake detection, allowing the system to detect manipulated images as well as videos.
Further improvements may also focus on enhancing the user interface and providing more detailed output analysis.

Future Work

Future development will focus on improving the system's performance and usability. This includes training the model on larger and more diverse datasets to enhance generalization.
Optimization techniques will be applied to reduce processing time and improve efficiency.
The system will be extended to support real-time deepfake video detection, enabling instant analysis of live or streamed video content.
In addition, future work will include adding image-based deepfake detection, allowing the system to detect manipulated images as well as videos.
Further improvements may also focus on enhancing the user interface and providing more detailed output analysis.

Deepfake Video Detection using Vision Temporal Transformer and Multi-Head Attention

Leveraging AI technologies for accurate detection and classification of manipulated videos to combat misinformation.

Take this with you.
Revisit anytime.

Missed something? Want to explore further?
Scan or click below to open this presentation.
Anytime, anywhere.

View presentation