ICLR

Image Analysis & Biometrics Lab

ILLUSION

Unveiling Truth With a Comprehensive Multimodal, Multilingual Deepfake Dataset

Kartik Thakral[1], Rishabh Ranjan[1], Akanksha Singh[1,2], Akshat Jain[1], Mayank Vatsa[1], Richa Singh[1]

[1]IIT Jodhpur [2]IISER Bhopal

MOTIVATION

With the increasing prevalence of deepfakes on the internet enabling misinformation, fraud, and social engineering, there is an urgent need for robust detection methods to safeguard digital trust and security.

Images

Videos

Audio

DATA BREACH • DIGITAL PRIVACY • 3 min read •

Thousands Of AI-Generated and Deepfake Images Exposed in Unprotected Database Online

Alina BÎZGĂ
April 01, 2025

Promo Protect all your devices, without slowing them down.
Free 30-day trial

PBS NEWS WEEKEND

How AI is being used to create explicit deepfake images that harm children

Mar 22, 2025 5:35 PM EDT

Football and media personality Eddie McGuire used in deepfake financial advertisement

By Melissa Brown | Scams and Fraud

Mon 31 Mar

AI and deepfakes blur reality in India elections

16 May 2024 Share Save

Meryl Sebastian
BBC News, Kochi

Deepfake Audio Of Philippine President Urging Military Action Against China Sparks Concerns

The audio features a deepfake voice of Marcos Jr, where he indicates to his military to intervene if China poses a threat to the Philippines.

Edited by: Anjali Thakur | World News | Apr 25, 2024 09:21 am IST

Unusual CEO Fraud via Deepfake Audio Steals US$243,000 From UK Company

September 05, 2019

INTRODUCTION

Problem Statement

The purpose of the dataset is to aid in the creation of multimodal deepfake detection algorithms that are robust to all forms of fake media and unified across all three modalities, are bias-free and imperceptible to human eyes.

Research Gaps

Unimodal

Most SOTAs are unimodal

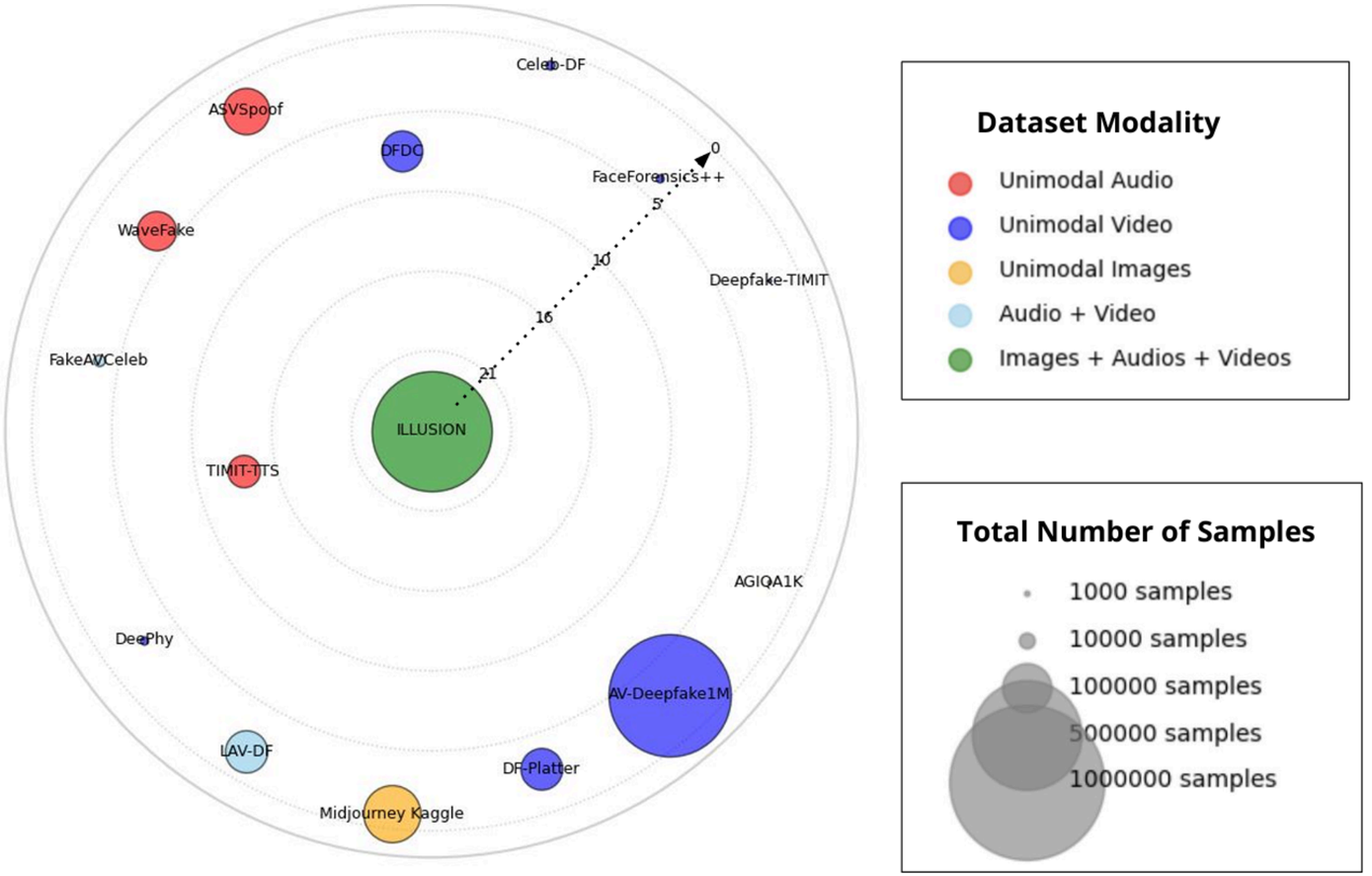
Variation

Exhaustive list of models, video length, quality of sync

Bias

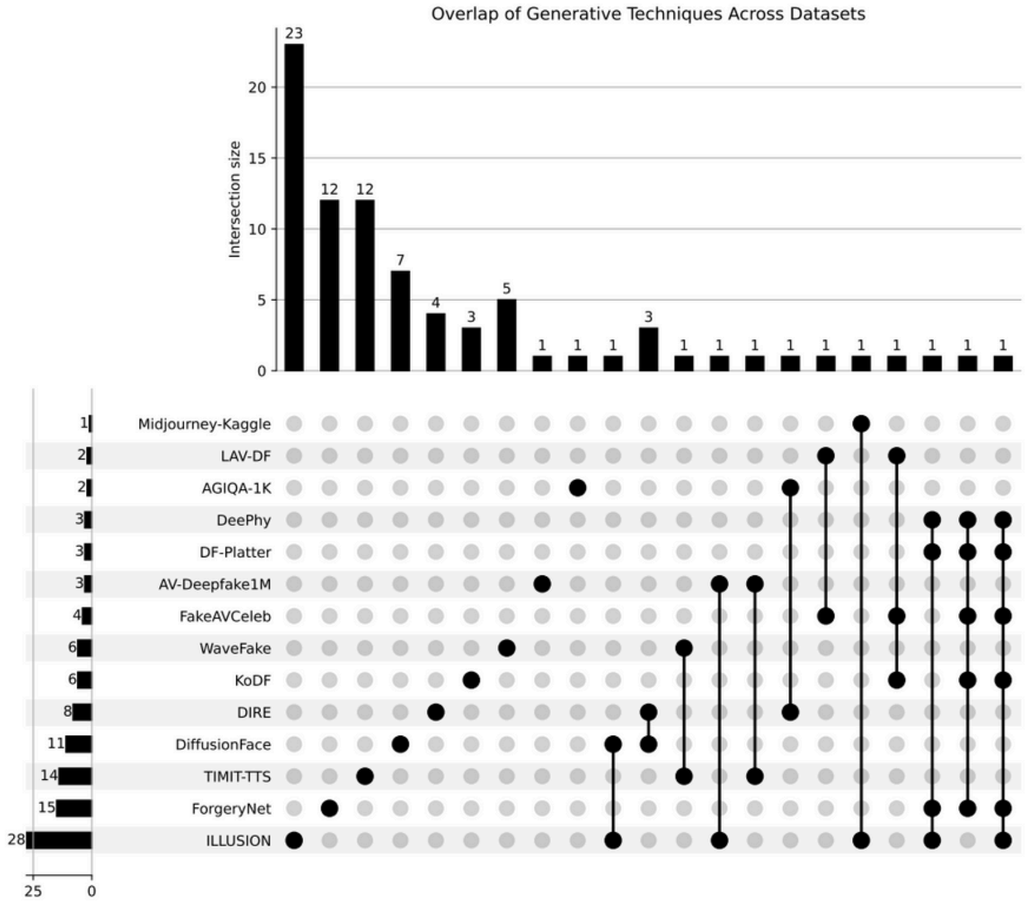
Sex and skin-type biases

RELATED WORKS

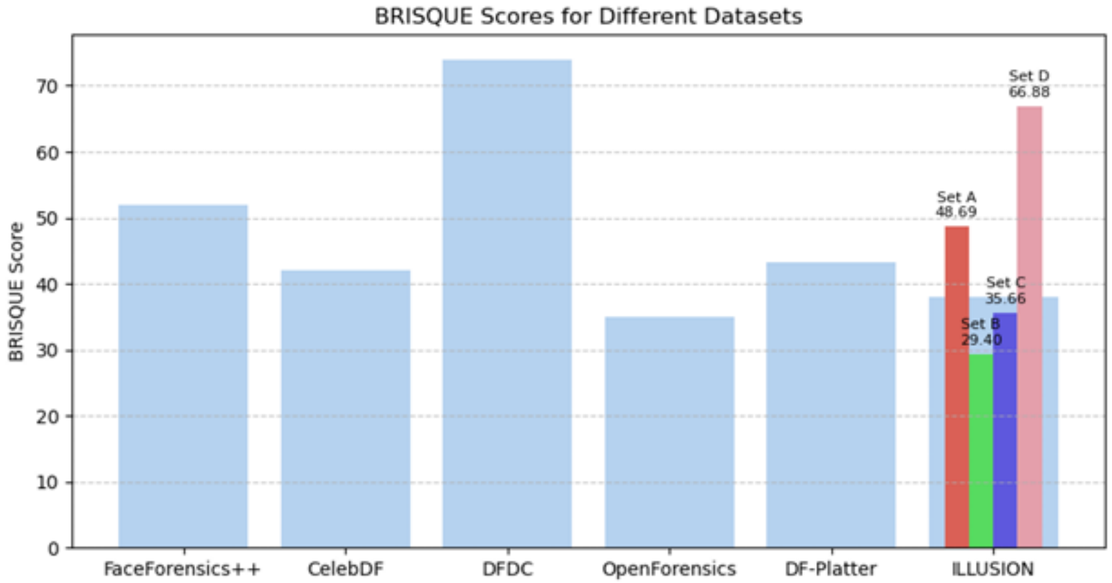


Comparison of the proposed dataset with existing datases based on modalities, size, and manipulations

COMPARISON WITH EXISTING DATASETS

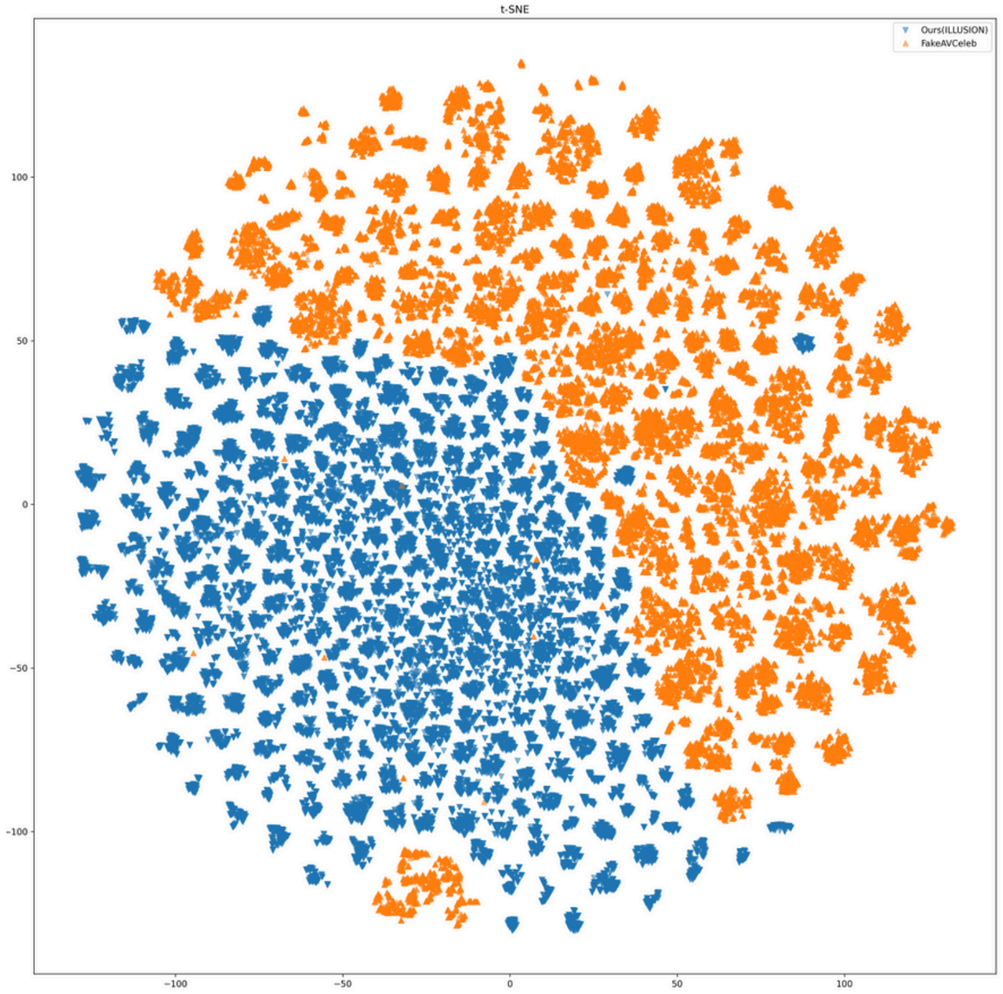


Overlap of generative techniques across datasets (UpSet Plot)



Comparison of visual quality of datasets

COMPARISON WITH EXISTING DATASETS



Feature-level comparison between proposed dataset and FakeAVCeleb (PCA and t-SNE).

Dataset	Jaccard Index
AGIQA-1K	0.00
AV-Deepfake1M	0.03
DF-Platter	0.07
Midjourney-Kaggle	0.04
LAV-DF	0.00
DeePhy	0.07
TIMIT-TTS	0.00
FakeAVCeleb	0.03
ForgeryNet	0.05
KoDF	0.03
DiffusionFace	0.03
DIRE	0.00
WaveFake	0.00

Uniqueness of the proposed dataset in comparison to existing datasets using Jaccard index.

DATASET STATISTICS

28
techniques

4
sets

139740
real samples

27244
fake audio

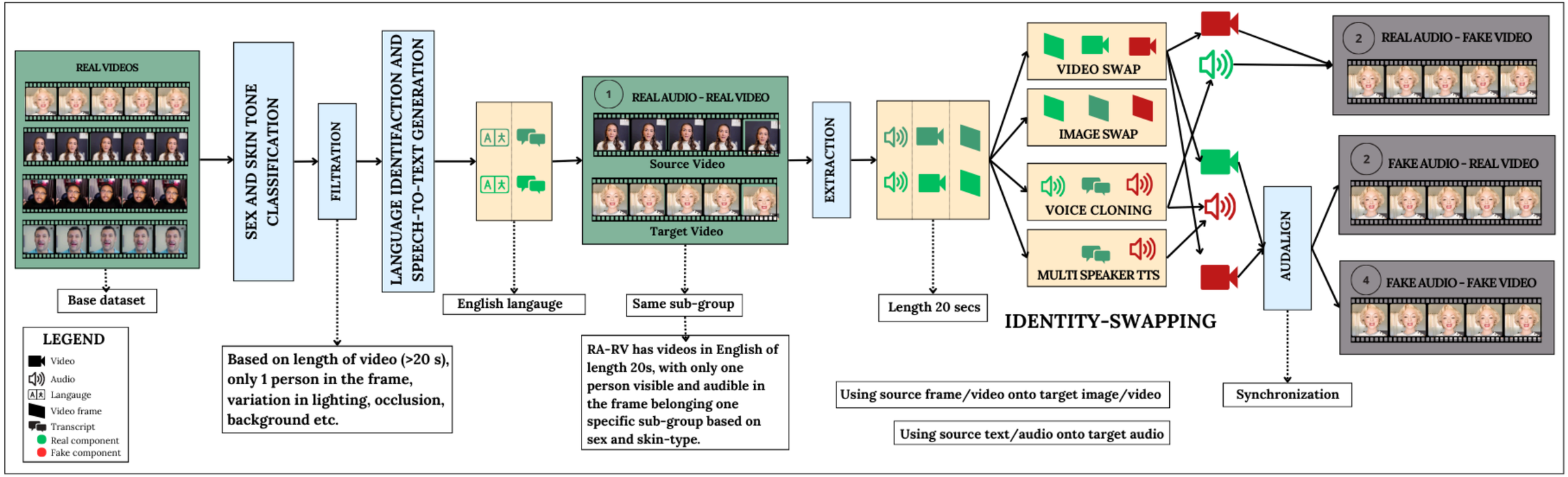
299454
fake videos

905548
fake images

1371986
total samples

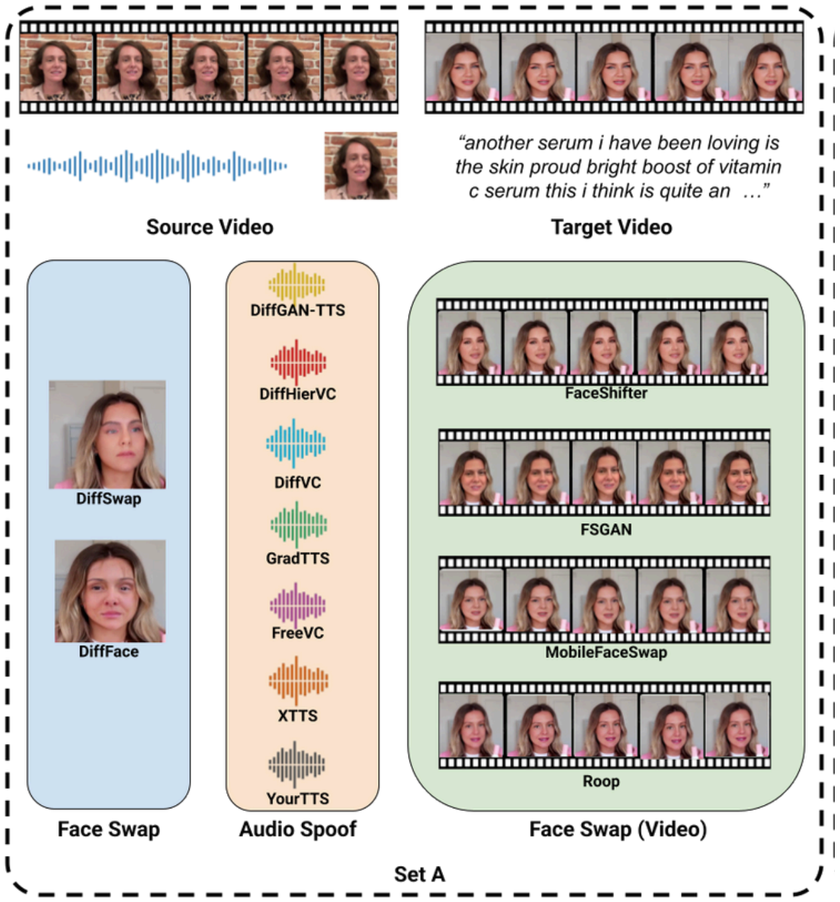
Collated samples of techniques used.

GENERATION PIPELINE: SET A



Creation of each class label of Set A: Identity-swaps

DATASET DESCRIPTION: SET A



Structure:

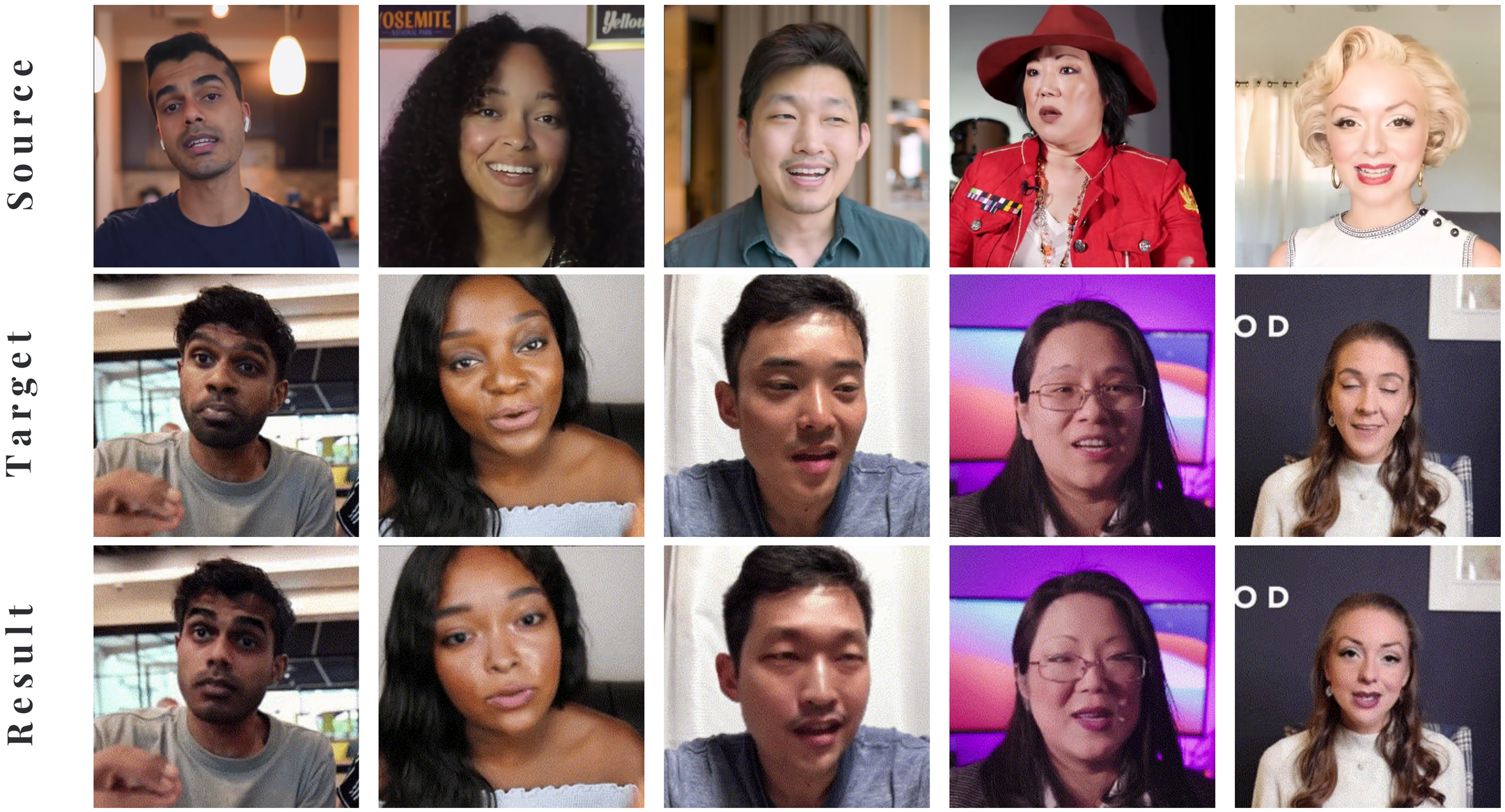
- Skin-tone based:
  1. Bin 1 Light (Fitzpatrick I and II)
  2. Bin 2 Light-Medium (Fitzpatrick III)
  3. Bin 3 Medium-Dark (Fitzpatrick IV)
  4. Bin 4 Dark (Fitzpatrick V and VI)

- Sex based: Male & Female

Classes:

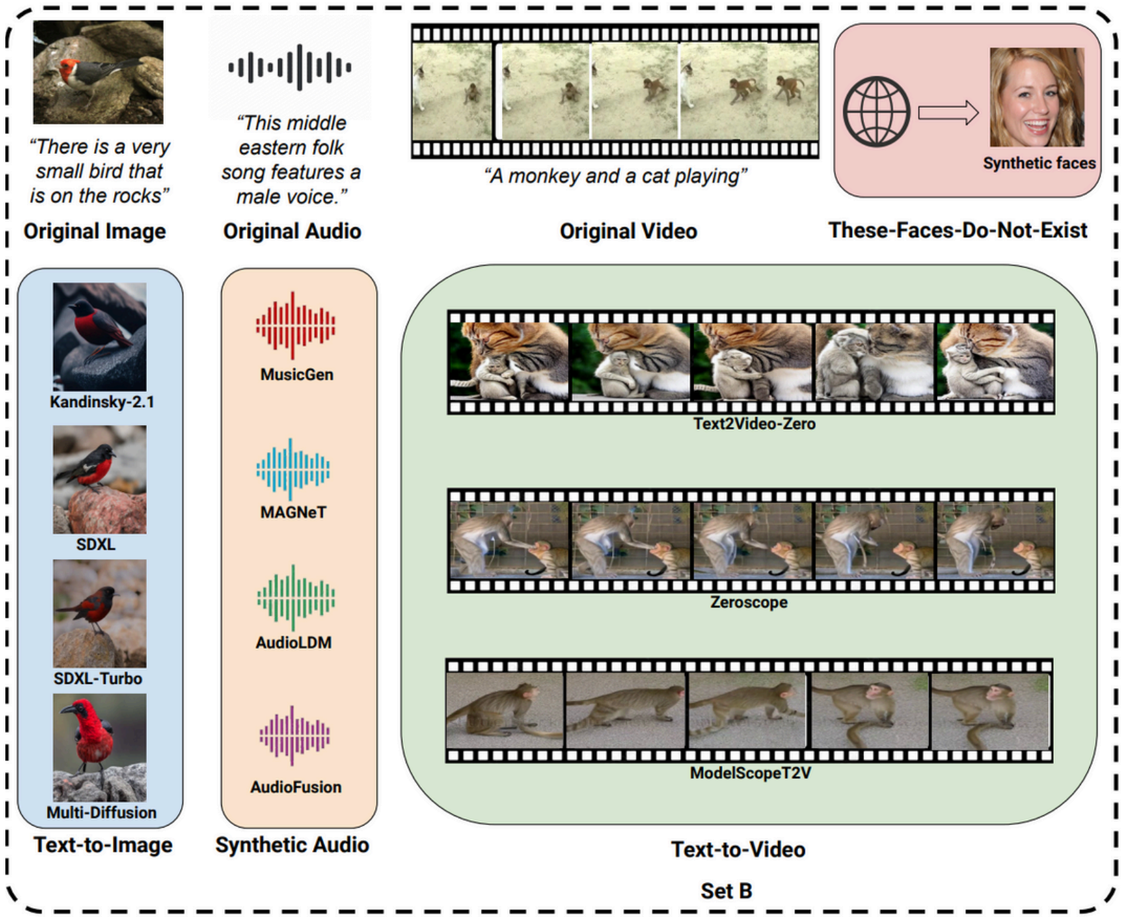
1. Real Audio - Real Video
2. Real Audio - Fake Video
3. Fake Audio - Real Video
4. Fake Audio - Fake Video

Different Components of Set A: Identity-swaps



Illustrative example of Set A: Identity-swaps creation

DATASET DESCRIPTION: SET B



Used real dataset:

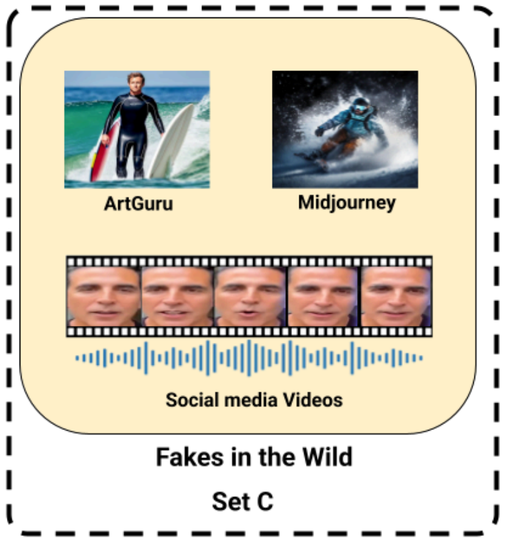
- Audio: MusicCaps
- Images: COCO
- Video: MSR-VTT

Structure for This Person Does Not Exist

- Skin-tone based: Bin 1, Bin 2, Bin 3 & Bin 4
- Sex based: Male & Female

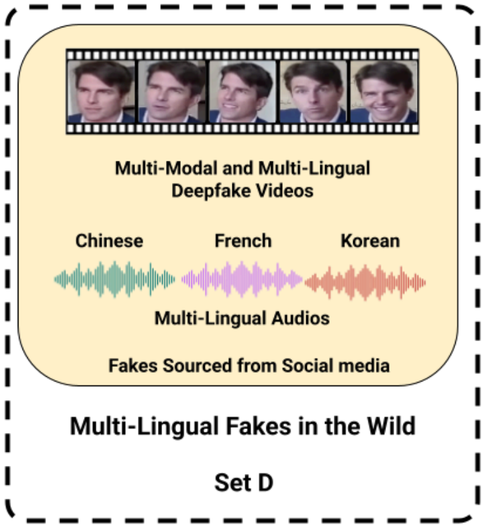
Different Components of Set-B: AI Generated Content

DATASET DESCRIPTION: SET C & SET D



Sample Generated from:

- Images: Midjourney & ArtGuru
- Videos: Real-World Deepfakes from Internet



Languages used:

26 languages, including French, German, Italian, Chinese, Korean, Arabic, Japanese, Tamil, Kannada, Oriya, Hindi, Sanskrit, Latin, Punjabi, and Gujarati.

Different Components of Set-C & D: Unseen Test Sets

RESEARCH QUESTIONS & PROTOCOLS

RQ1

How effective are the detection systems in detecting **multi-modal identity-swaps**?

RQ2

How effective are the detection systems in identifying **synthetically generated media**?

RQ3

How robust and reliable are the current state-of-the-art detection algorithms when **deployed in real-world scenarios**?

RQ4

Is it feasible to detect identity swaps and synthetic media in a **zero-day attack** setting?

RQ5

Is it possible to successfully **trace back the source** of a given deepfake?

Protocol 1

Multi-modal Deepfake Detection

Protocol 3

Generalization on Real-World Deepfake Media

Protocol 2

Zero-shot/ Zero-day Generalization

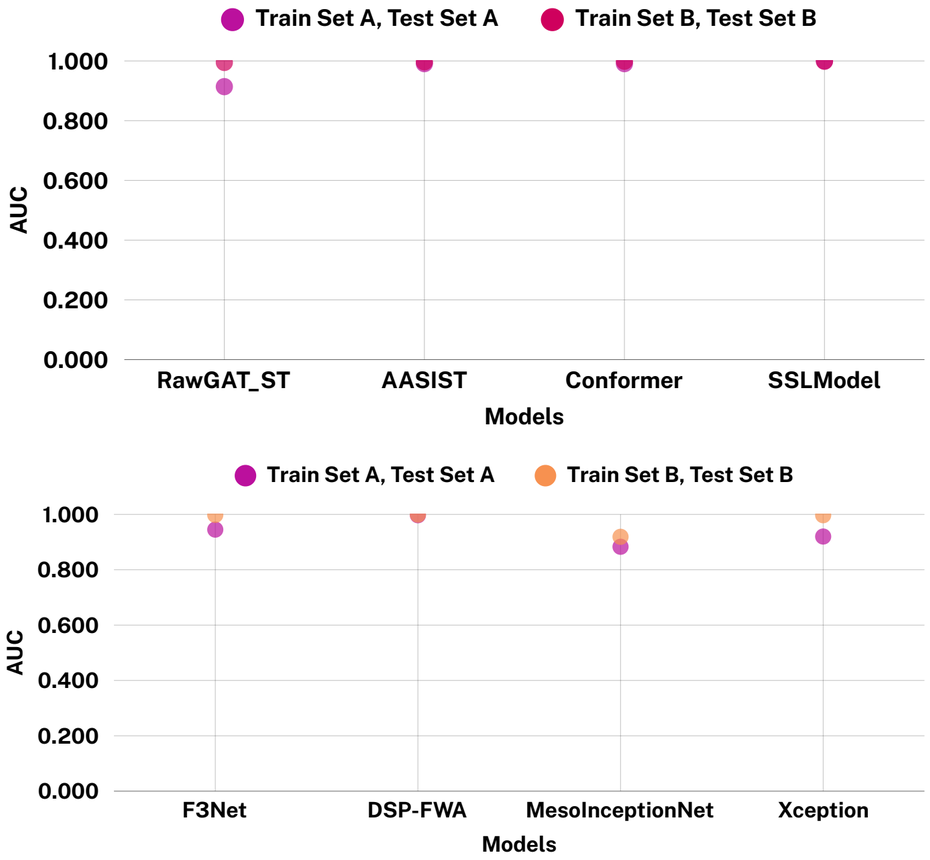
Protocol 4

Performance on Model Attribution

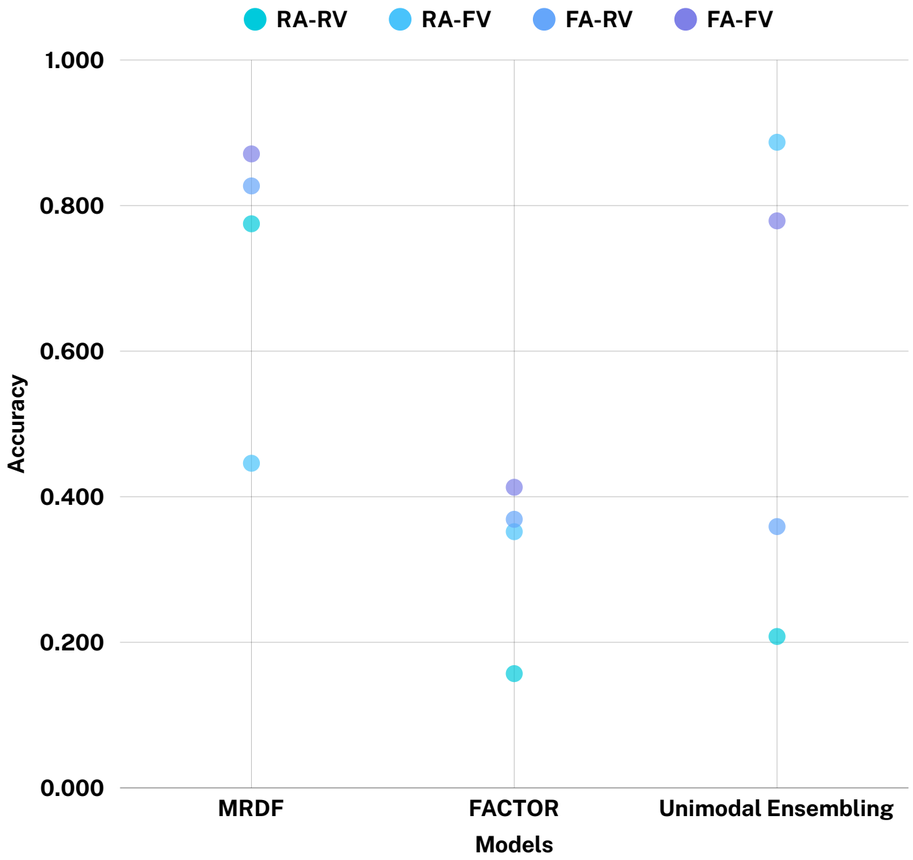
EXPERIMENT 1

Results & Discussions

1. All architectures perform well for both audio and visual unimodal detection.
2. DSP-FWA and SSLModel perform best in visual and audio models, respectively.
3. MRDF outperforms FACTOR across all classes.



Unimodal audio detection (top) & unimodal visual detection (bottom) performance



Multimodal detection performance when trained and tested on Set A

EXPERIMENT 2

Results & Discussions

1. Models trained on Set A and tested on Set B show random performance for both audio and visual models, and vice versa is also true.
2. The inference is that the artifacts of identity swaps and synthetic media are completely different.

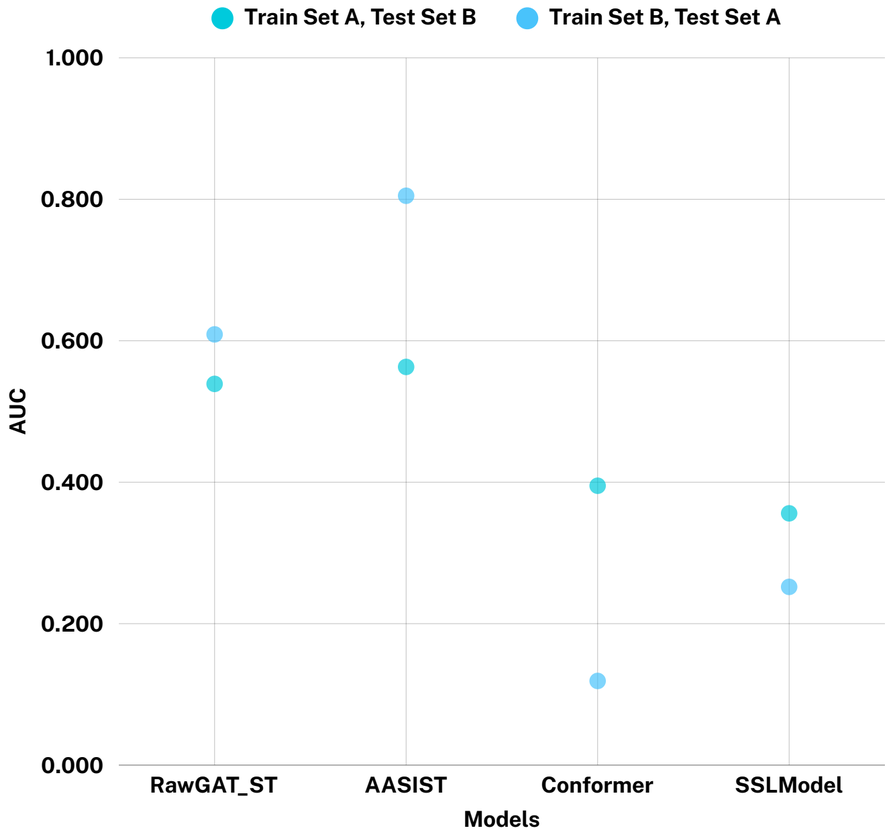
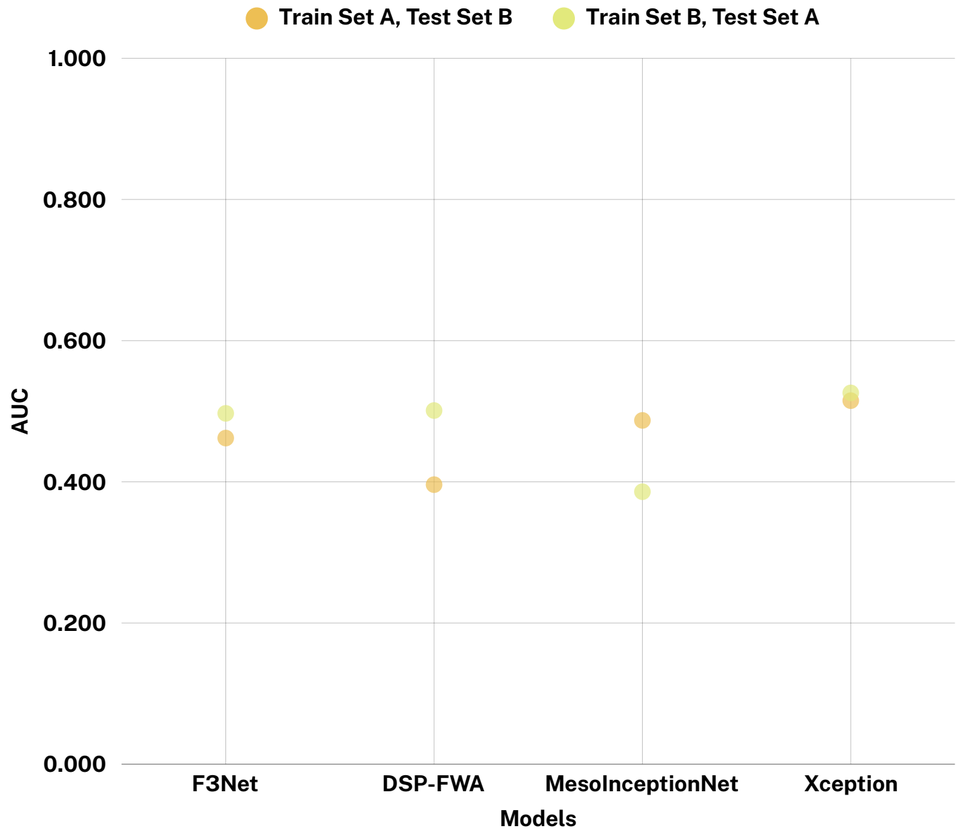


Chart: Cross train-test set audio detection performance

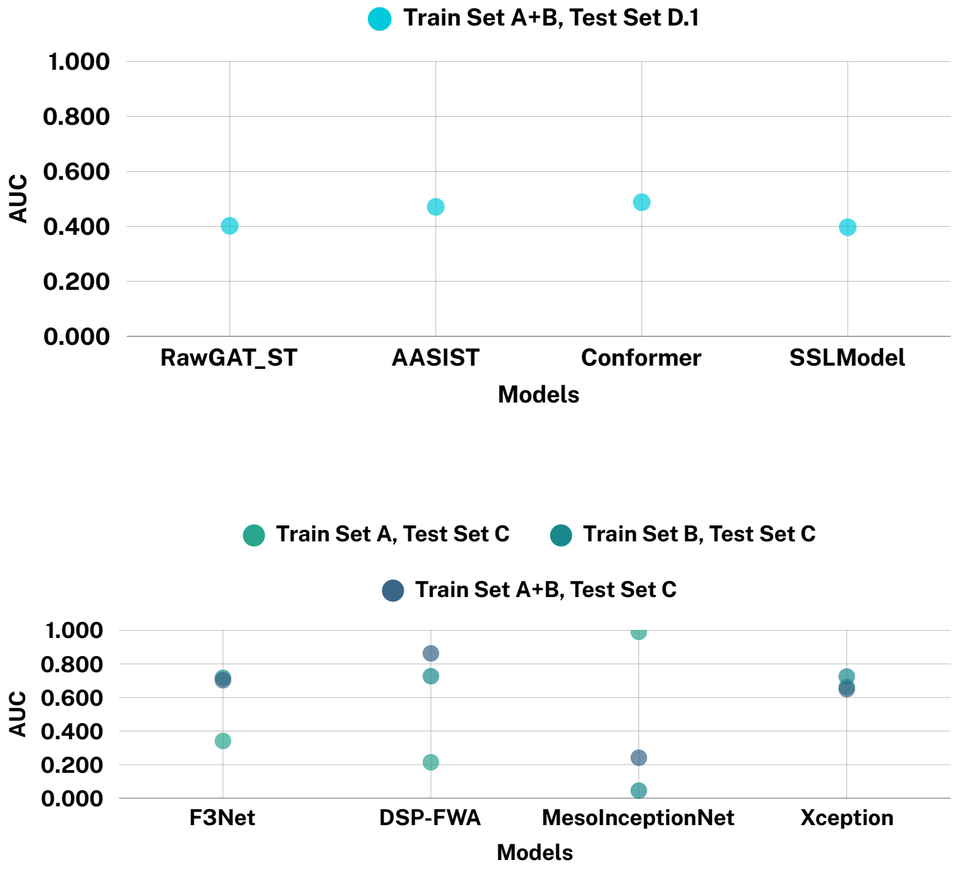


Cross train-test set visual detection performanceA

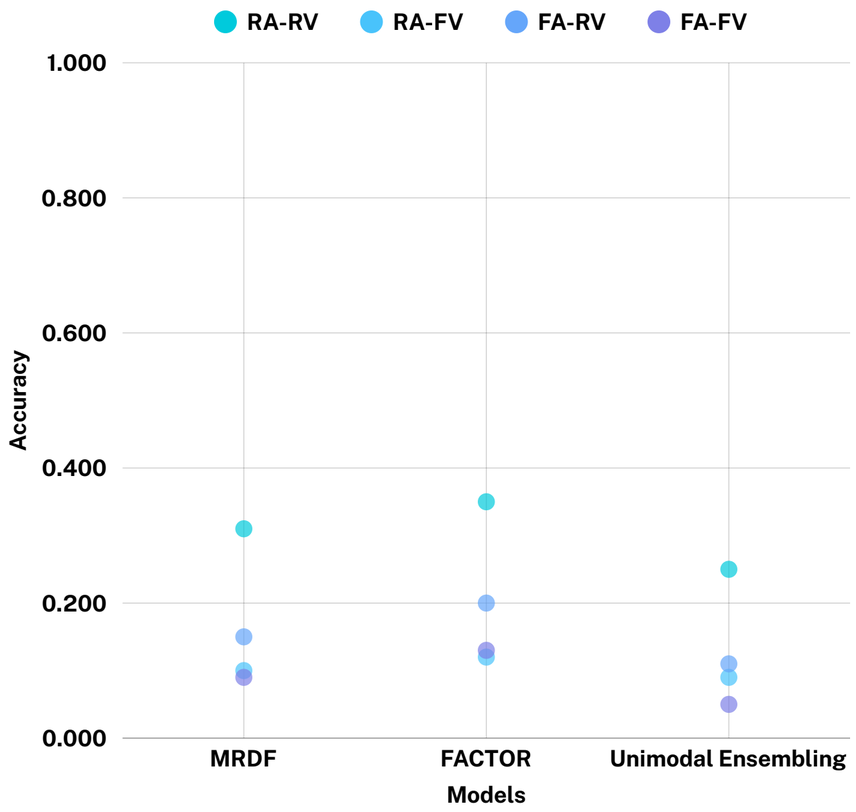
EXPERIMENT 3

Results & Discussions

1. Models perform better when trained on Set B and tested on Set C than when trained on Set A because Set C has more synthetic samples.
2. There is slight increase in performance when trained on both Set A and Set B.
3. Audio and multimodal models perform poorly when trained on both Sets A and B and tested on Set D.1. and D.2 respectively. Experiment highlights the current models struggle to generalise over unseen and real-world samples.



Unseen audio detection (top) & unseen visual detection (bottom) performance

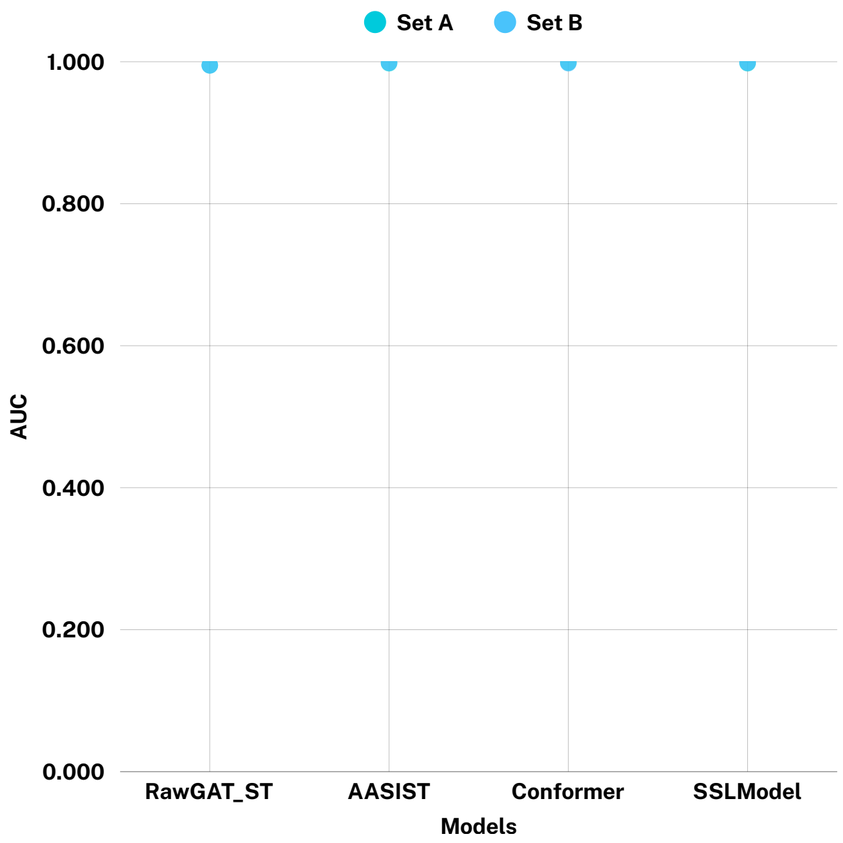


Unseen multimodal detection performance when tested on Set D.2

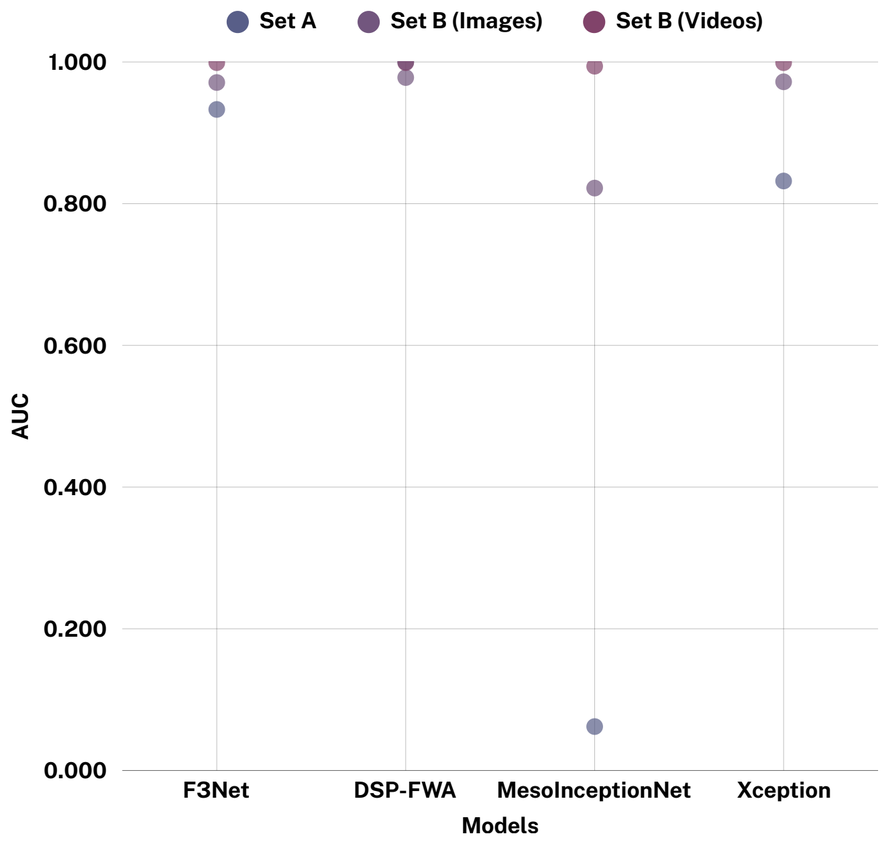
EXPERIMENT 4

Results & Discussions

1. Both audio and visual models perform well in the model attribution task for both Set A and Set B.
2. It can be inferred that generative models introduce unique signatures in their output, which can be differentiated easily.



Audio model attribution performance



Visual model attribution performance

CONCLUSION

With this paper, we introduce the ILLUSION dataset, a significant step towards a comprehensive, multimodal deepfake resource. Created using **28 state-of-the-art generative models**, ILLUSION provides diverse AI-generated content across **image, audio, and video** modalities, including both curated **real-world deepfakes** and **synthetic media**. This design enables models trained on ILLUSION to learn features that extend beyond synthetic artifacts, enhancing generalization across domains, particularly in **multilingual and noisy settings**. Results show that detection models trained on ILLUSION outperform those trained on existing datasets when evaluated on unseen generative techniques and real-world forgeries.

ICLR

Image Analysis & Biometrics Lab

Thank you!

The dataset is available at: https://www.iab-rubric.org/illusion-database.

Reach out to us at mvatsa@iitj.ac.in.