Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024

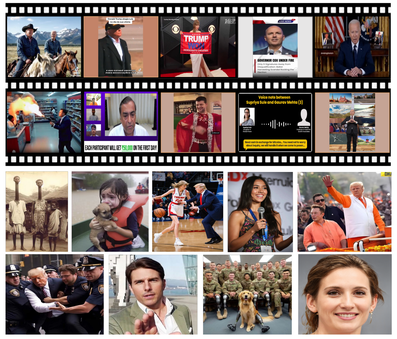
Nuria Alina Chandra[1] Hannah Lee[1] Ryan Murtfeldt[1,2] Lin Qiu[1,2] Arnab Karmakar[1,2] Emmanuel Tanumihardja[1,2] Kevin Farhat[3]
Ben Caffee[1,2] Changyeon Lee[4] Jongwook Choi[5] Sejin Paik[6] Aerin Kim[1,7] Oren Etzioni[1,2]

[1]TrueMedia.org [2]University of Washington [3]Allen Institute for Artificial Intelligence [4]University of Maryland [5]Chung-Ang University [6]Georgetown University [7]Miraflow AI

Summary

- We introduce **Deepfake-Eval-2024**, a a challenging benchmark consisting of in-the-wild deepfakes from social media and TrueMedia.org.
- **Deepfake detection** is one method to address the surge of realistic deepfakes used for misinformation, creation of non-consensual imagery, and reputational damage.
- Academic benchmarks are outdated and not representative of real world examples
- Commercial and open-source models show significant limitations on in-the-wild data

Dataset Curation

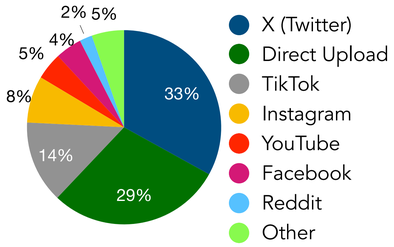


- Data was collected through:
  - The TrueMedia.org platform: non-profit application used by journalists and fact-checkers in 2024
  - Posts flagged by X Community Notes and other community moderation platforms

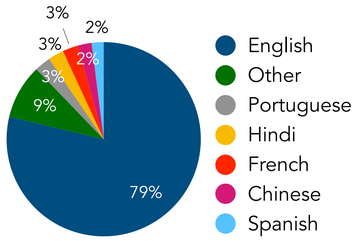
Dataset Composition

Modality	Size	Avg Resolution
Video	45.1 hrs	30.05 FPS, 576 x 720 px
Audio	56.5 hrs	44.66 kHz
Image	1,975 images	1024 x 1024 px

Data Origin



Languages

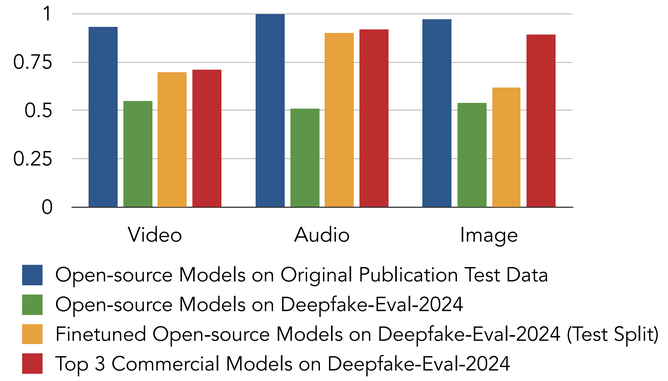


Experiments

- Evaluated 3 open-source models per modality
- Evaluated 22 commercially available detection models from Hive, Reality Defender, Pindrop, AI or Not, Hiya, Fraunhofer, Sensity AI
- We finetune all open-source models on Deepfake-Eval-2024 to determine if performance can improve and identify limitations

Results

Average AUC for Open-Source Models



- AUC drops on average by **41% for video, 48% for audio,** and **45% for image** open-source models on Deepfake-Eval-2024 compared to previous benchmarks
- Top commercial models exceed the performance of both original and finetuned open-source models
- Models:
  - Are 21.3% worse on diffusion-generated videos
  - Perform worse at classifying audio with non-English languages, silences, and background noise
  - Have lower accuracy on images with text overlays