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

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

[1]TrueMedia.org
[2]University of Washington, Seattle
[3]Miraflow AI
[4]Georgetown University, Washington D.C.
[5]Chung-Ang University, Seoul
[6]Yonsei University, Seoul

Abstract

Existing academic deepfake detection benchmarks suggest that state-of-the-art detection models can identify deepfakes with high accuracy. To evaluate detector performance on deepfakes currently circulating on social media platforms, we present Deepfake-Eval-2024, an in-the-wild detection benchmark. When open-source detection models are evaluated on Deepfake-Eval-2024, AUC drops by 50% for video, 48% for audio, and 45% for image models as compared to academic datasets. When open-source models are finetuned on a subset of Deepfake-Eval-2024, performance improves, but significant gaps still remain. We also evaluate commercially available deepfake detectors on Deepfake-Eval-2024 and find that their performance is superior to open-source models but still need significant improvement to reach high accuracy on real-world media. To date, Deepfake-Eval-2024 is the first in-the-wild dataset with all three media modalities and the largest and most diverse in-the-wild audio deepfake detection dataset. Deepfake-Eval-2024 contains 44 hours of videos, 56.5 hours of audio, and 1,975 images, encompassing the latest manipulation technologies, diverse media content, 88 different website sources, and 52 different languages.

1 Introduction

Deepfakes are digital fabrications in which realistic likenesses of people are synthetically generated or altered by a deep learning model to say or do something that never occurred [1]. Although there are positive uses of AI-generated likenesses [2], deepfakes also pose a growing threat to society. Deepfakes can be used to fabricate messages from politicians[3], create non-consensual pornographic content, spread misinformation, and damage reputations, harming lives, businesses, and nations [4].

Advances in generative AI models have precipitated a surge of highly-realistic deepfakes. There has been a 4-fold increase in the number of deepfakes detected in fraud from 2023 to 2024 [5]. In 2023 alone, there were an estimated 500,000 deepfakes shared on social media [5]. Further, the realism of AI generated content continues to increase with the release of new methods and models such as large vision models like Open-AI's Sora [6], and latent diffusion based models like Midjourney version 6.1 [7]. Recent research has shown that humans are no longer able to accurately identify if content is AI generated or not [8].

Preprint. Under review.

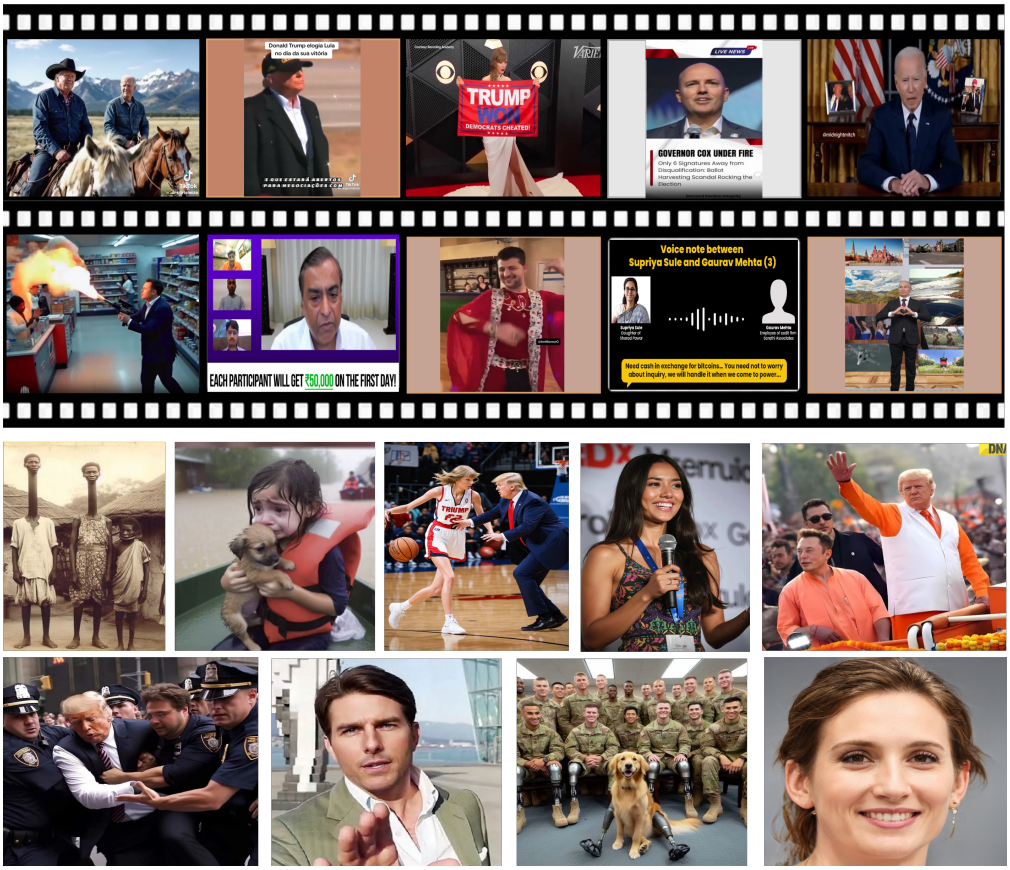

Figure 1: Examples of deepfake videos and audio (rows 1–2) and images (rows 3–4) from our Deepfake-Eval-2024, demonstrating a diversity of content styles and generation techniques, including lipsync, faceswap, and diffusion. Images have been resized for presentation.

In today's ecosystem of increasing deepfake prevalence and realism, the need for accurate AI generated content detectors has become paramount. There is a movement to embed cryptographic provenance information in watermarks [9]. However, such initiatives do not offer a full solution for detecting content created by organizations and individuals who choose not to comply. Thus, technologies that accurately detect AI-generated content directly are essential.

Dozens of deepfake detection models have been developed to classify AI-generated audio, images, and videos [10], but the accuracy of these detection models on contemporary deepfakes being circulated on social media was previously unknown. Published deepfake detection models have reported high accuracy on academic datasets, but the majority of these datasets are outdated or generated using a limited number of techniques. In addition, these datasets fail to represent the diversity of real-world media. For example, most audio datasets do not contain background noise or music, and most video datasets show only single scenes, without common manipulations such as video splicing or news-media overlays. Deepfake-Eval-2024 addresses these gaps in existing deepfake datasets. Figure 1 illustrates examples of deepfake videos and images from our dataset, highlighting the diversity of content and the variety of creation techniques.

We achieve high dataset diversity through the data collection approach. Each item in the dataset is a piece of media that a social media or TrueMedia.org user thought might AI manipulated. As a result, the dataset is smaller than synthetic datasets, but much more diverse.

In order to evaluate the state of deepfake detection on contemporary real-world deepfakes, we present Deepfake-Eval-2024 as a benchmark for deepfake video, audio, and image detection models. We summarize our contributions with the following:

- We present a challenging multi-modal in-the-wild deepfake evaluation dataset comprised of contemporary data collected in 2024.
- To our knowledge, this is the first in-the-wild dataset with inclusion of all three media modalities and the largest and most diverse audio in-the-wild deepfake detection dataset.
- We evaluate the state-of-the art of deepfake detection on contemporary in-the-wild data using open-source and commercial deepfake detection models.
- We provide recommendations for the the field of deepfake detection based on our findings.

2 Related Work

Many deepfake detection datasets have been released during the past five years in an effort to advance deepfake detection and generation techniques. Supplementary Tables 8, 9, 10 provide a detailed survey of existing popular datasets compared to Deepfake-Eval-2024. Large-scale deepfake detection challenges such as ASVspoof [11] and the Deepfake Detection Challenge (DFDC) [12] have driven the creation of larger datasets; both include more than 300 hours of data, while even newer datasets such as AV-Deepfake1M gather close to 2,000 hours [13] and some image datasets include over one million data points [14]. However, most of these datasets were created prior to newer content generation methods [6, 7] and do not include in-the-wild-data representative of prevalent online deepfake content.

Deepfake datasets have not kept up with the fast moving field of AI content generation. The most recent in-the-wild video datasets are from 2020 [15] and 2021 [16] (Supp. Table 8). The only other significant in-the-wild audio deepfake dataset was published in 2022 [17] (Supp. Table 9), and has fewer hours of audio than Deepfake-Eval-2024. Further, to the best of our knowledge, there are no other in-the-wild deepfake image-focused datasets (Supp. Table 10).

Synthetically generated datasets are limited by a lack of diversity in generative techniques, content styles, and languages. For example, many deepfake video datasets consist of real videos of paid individuals sitting in specific positions, and fake videos derived from applying a handful of AI-manipulation techniques to the curated videos (e.g. [18, 12, 19]). Deepfake-Eval-2024 includes images of people in a wide variety of settings and positions, demonstrating a wide range of actions (Figure 1). The content in most audio datasets lack language diversity, usually only including clips in English [20, 21, 11, 17]. The maximum number of languages in an existing major audio dataset is two (Supp. Table 9). In comparison, Deepfake-Eval-2024 has 42 languages in our audio dataset, and 52 different languages in combined video and audio datasets (Figure 3).

3 Dataset

Deepfake-Eval-2024 is composed of 44 hours of videos and 56.5 hours of audio and 1,975 images. (Tables 1, 2, 3 contain complete summary statistics.) The data includes real, AI-generated, and AI-manipulated content. Audio data includes audio from videos, in addition to audio-only media. The majority of video data has corresponding labeled audio.

3.1 Data Collection

All data was collected through the TrueMedia.org deepfake detection platform in 2024. TrueMedia.org was a non-profit application originally used primarily by journalists and fact-checkers starting in April 2024, and available to the general public starting in September 2024. Users provided a social media link or directly uploaded content to be checked for AI-manipulation. We also created a bot

Table 1: **Deepfake-Eval-2024 Video Summary Statistics**

Category	Total Duration (hrs)	Count	Avg. Duration (s)	Avg. FPS	Mode Resolution (W×H)
Real	28.9	1,072	96.94	30.92	1,280×720
Fake	16.2	964	60.47	29.09	576×720
All	45.1	2,036	79.68	30.05	576×720

Table 2: **Deepfake-Eval-2024 Audio Summary Statistics**

Category	Total Duration (hrs)	Count	Avg. Duration (s)	Avg. Sampling Rate (kHz)
Real	36.6	1,110	124.80	44.83
Fake	19.9	710	101.51	44.40
All	56.5	1,820	115.46	44.66

Table 3: **Deepfake-Eval-2024 Image Summary Statistics**

Category	Count	Mode Resolution
Real	1,208	1,200×1,200
Fake	767	1,024×1,024
All	1,975	1,024×1,024

on X (previously known as Twitter) that allowed users to add content to our platform by tagging the bot. We also uploaded posts from X that had been flagged by X Community Notes as potentially manipulated media. The top five most common data sources of Deepfake-Eval-2024 are X, direct upload, TikTok, Instagram, and Youtube (Figure 2, Supp. Figure 4).

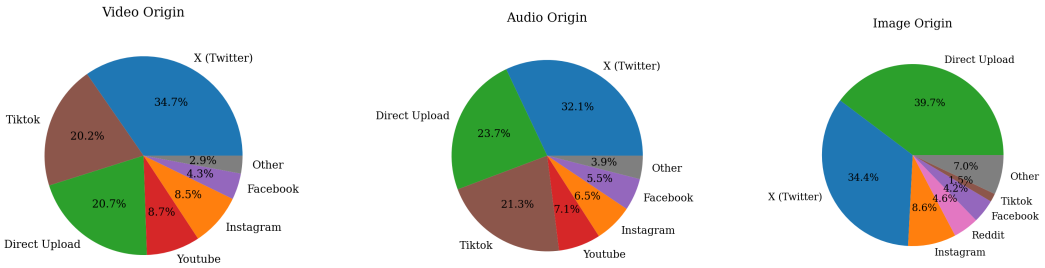

Figure 2: Origins of data in Deepfake-Eval-2024 separated by modality. In total, media was shared from 88 different web-domain names. Direct upload indicates that the media was uploaded directly to the TrueMedia.org deepfake detection platform, instead of through a social media or website link.

This data collection method ensures that Deepfake-Eval-2024 is a **challenging** dataset. Users often brought media to TrueMedia.org when it could not be easily identified as real or fake by a human. Thus, we estimate that Deepfake-Eval-2024 has a greater proportion of challenging examples in both real and fake categories than prior datasets.

Our data collection approach also provided for increased diversity with respect to generative models, ethnicity, language, and content. Deepfake-Eval-2024 is a sample of currently circulating AI-generated content. Thus, we estimate that our dataset includes AI-generated and manipulated content from every type of contemporary model commonly used to generate deepfakes. Further, TrueMedia.org users came from all over the world, resulting in increased ethnic and linguistic diversity in our dataset. Our dataset is 78.7% English and includes a total of 52 different languages (Figure 3). The content of the media itself has larger variations than the clean standardized content typically found in academic datasets. The diversity in data origins (Figure 2) results in our dataset containing a wide variety of different media styles, including videos of political speeches and self-shot content-creators, images of large crowds and close-up portraits, and audio clips of debating politicians and background music.

3.2 Data Filtering

We removed duplicate data using a combination of manual review and hash functions. We included cases where two pieces of media have minor, non-visible variations, and thus appear to be the same (e.g., different cropping of the same video). In order to tailor our datasets to evaluate deepfake detection models, we removed images and videos that do not contain photorealistic faces. This resulted in the removal of cartoons, art, and scenes without humans. We use GPT-4o (version

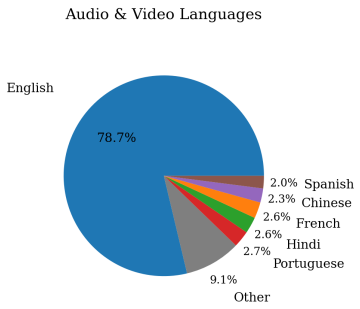

Figure 3: Language distribution in Deepfake-Eval-2024 audio and video data. The dataset contains a total of 52 different languages (42 languages in Deepfake-Eval-2024-audio and 49 languages in Deepfake-Eval-2024-video). Languages were identified using speech recognition model Whisper, which is known to perform well on language identification tasks [22].

2024-08-06) to identify images with photorealistic faces. We note that GPT-4o's responses have high precision, but also a high false negative rate. To account for this, we manually review all images marked as non-photorealistic faces by GPT-4o. To identify videos with photorealistic faces we first use the dlib face detection library [23] to determine whether each frame contains a face or not. Videos where no faces are detected are then reviewed manually to check for missed faces.

We remove data that labelers were unable to categorize. Labeled videos are still included in the Deepfake-Eval-2024-video dataset even when the label of their corresponding audio could not be determined. Likewise, labeled audio from videos is included in Deepfake-Eval-2024-audio even when the corresponding video is unlabeled.

3.3 Data Labeling

We label media as fake if it was AI generated or manipulated. We choose to define our labels this way despite the challenge of differentiating AI-manipulated content from traditional forms of manipulation so that this dataset can be used to benchmark deepfake detection models, the majority of which are trained to differentiate between AI-generated and real content.

The labeling team consisted of seven people: Three experienced AI generated content labelers, and four machine learning research interns. The team met regularly to discuss the taxonomy, verification process, and edge cases. Our verification process consisted of locating original sources using reverse image search or searching the web using quotes or situation descriptions, then confirming the trustworthiness of the source, and scanning media for characteristics of AI-generated media (Appendix B). See Appendix C for our detailed verification process. The team labeled each piece of media as 'fake', 'real', or 'unknown' when the appropriate label was not clearly discernible from authenticated sources or media characteristics. Media labeled 'unknown' were excluded.

We relied on articles published by professional fact-checking organizations such as Snopes[1] and AFP Fact Check[2] to confirm if AI manipulation was used. Additionally, we utilized community moderation platforms like X Community Notes to locate and review primary sources. We conducted source context verification by comparing social media posts with original materials such as full-length videos to verify if media had been manipulated. We further scrutinized the media for evidence of AI manipulation using specific forensic markers. For example, we relied on the synchronization of mouth movements with vocal sounds as a primary measure of authenticity in video and audio media. For videos and images, face-swaps were classified as fake unless created prior to 2023, which is supported by research that reported a 704% increase in AI-generated face swaps during that year [24]. Other common indicators of AI manipulation, including anatomical implausibilities, sociocultural implausibilities, and stylistic artifacts, were identified based on the framework created

[1]Snopes is one of the oldest and most widely trusted fact-checking websites, operating since 1994. The site employs professional journalists who investigate viral claims and urban legends. https://www.snopes.com/about/

[2]AFP Fact Check is the fact-checking division of Agence France-Presse, a major world news agency. They operate in multiple languages and verify content across various social media platforms. https://factcheck.afp.com/about-afp

Table 4: **Inter-rater Disagreement Statistics Across Modalities**

Modality	N Checked	Total Disagreement	Real vs Fake	Real vs Unknown	Fake vs Unknown
Video	243	6.6%	2.1%	3.7%	0.8%
Audio	342	7.9%	0.6%	3.5%	3.8%
Images	269	9%	0%	3%	6%

by Kamali et al. [25]. When the label of audio media could not be determined from sources, due to the challenge of manually detecting AI-generated audio (and differentiating it from non-AI generated voice impersonators), media were marked as fake if and only if there were both audible traits indicating that it was fake (e.g. sociocultural implausibilities), in addition to at least two commercial audio detectors predicting the media as fake. We note that this labeling approach results in audio labels that are more likely to be correlated with existing detectors, which may result in a less challenging audio dataset. For all modalities, when labelers were uncertain about the presence or absence of forensic markers, they labeled the media as 'unknown'. See Appendix B for complete taxonomy and media examples.

To assess the consistency and quality of our annotations, the labeling team-lead checked a random sample of 10% of the data for each of the three modalities (video, audio, and images). Annotations created by the team-lead were excluded to avoid self-assessment bias. For videos, we find a 6.6% disagreement between labelers, with the largest discrepancy between real and unknown at 3.7%. For audio, we find a 7.9% disagreement between labelers, with the largest discrepancies between real and unknown at 3.5% and between fake and unknown at 3.8%. And for images, we find a 9% disagreement between labelers, with the largest discrepancy between fake and unknown at 6%. (See Table 4 for complete disagreement breakdown.)

Common human errors include: differentiating between dubbed videos (where the video has not been AI-manipulated, and thus is real), and lipsynced videos (where the video has been AI-manipulated to make the mouth match new words and thus should be marked as fake); determining if audio sound is synthetic; and judging whether an image is fake vs unknown based on anatomical or sociocultural implausibilities.

4 Experiments

To evaluate the state-of-the-art of deepfake detection on real world in-the-wild deepfakes, we test an array of deepfake detection models on Deepfake-Eval-2024. We select standard models that encompass the primary deepfake detection model architectures associated with each modality. Models were also selected based on the availability of pretrained model weights and runnable training code. All models were chosen prior to experimentation, and were not omitted on the basis of performance.

4.1 Evaluating Open-Source Deepfake Detection Models

For each modality, we evaluate three different open-source deepfake detection models on the modality-appropriate Deepfake-Eval-2024 data. For image detection we include a single layer perceptron with a CLIP-backbone (UFD [26]), a model based on diffusion inversion (DistilDIRE [27]), and a convolutional neural network (NPR [28]). For audio detection we include a spectro-temporal graph attention network (AASIST [29]), a convolutional neural network applied to raw waveforms (RawNet2 [30]), and a model with a self-supervised component (P3 from Wang et al. [**?**]). We choose video models that have a generative convolutional vision transformer (GenConViT [31]), a temporal convolutional network (FTCN [32]), and a model that evaluates style latent vectors (Styleflow [33]). We use the code and preprocessing approaches described in the original publications. To adapt to open-source audio models with a limit of four seconds, we split Deepfake-Eval-2024 audio files into four-second segments and report performance on these segments.

We compare the performance of open-source models on our dataset to the performance of each model on the test datasets reported in its original publication (Table 5). To account for different reporting metrics used across publication, we recompute predictions on the originally published test datasets to provide a full array of evaluation metrics. Where multiple test datasets were reported in the original

publication, we compute results on as many of the datasets as possible and report average metrics across these test datasets.

4.2 Finetuning on Deepfake-Eval-2024

In order to determine if models can improve on real-world deepfake detection performance by training on more representative data, we finetune all open-source models on 60% of Deepfake-Eval-2024, and evaluate the performance on the remaining 40% of the data (Table 6). This split mirrors real-world scenarios where models must generalize from limited training data to detect unseen deepfake techniques. We finetune each model following the original authors' recommended training procedures and hyper-parameters where available, using early stopping to avoid overfitting.

4.3 Evaluating Commercial Deepfake Detectors

In addition, we evaluate commercially available deepfake detection models from companies that partnered with TrueMedia.org: Hive, Reality Defender, Pindrop, AI or Not, Hiya, Fraunhofer, and Sensity AI. In total, we evaluate 22 different commercial models (six commercial video models, eight commercial audio models, and eight commercial image models). For each modality, we evaluate commercial models on a subset of Deepfake-Eval-2024. We were unable to evaluate all commercial models on the entire dataset due to the high per-query cost of many commercial vendors. For each modality, we report the results of the best performing model on the entirety of Deepfake-Eval-2024 (Table 7). We anonymize the names of the commercial models to comply with contractual agreements. The commercial models are evaluated using their latest available versions as of December 2024. All vendors were blind to the test data.

Some open-source and commercial models fail to run on all media files due to model constraints (e.g. media length limits, or requirements for a face to be detected in a certain number of frames). When a model fails to produce a prediction, we exclude this file when calculating the metrics for the associated model.

5 Results

Table 5: **Open-Source Model Performance Across Modalities**

		Deepfake-Eval-2024				Original Publication Test Data			
Modality	**Model**	**AUC**	**Prec.**	**Recall**	**F1**	**AUC**	**Prec.**	**Recall**	**F1**
Video	GenConViT [31]	0.37	0.60	0.50	0.54	0.96	0.93	0.99	0.96
	FTCN [32]	0.50	0.51	0.67	0.41	0.87	0.91	1.00	0.95
	Styleflow [33]	0.51	0.54	0.43	0.48	0.95	0.96	0.89	0.77
Audio	AASIST [29]	0.43	0.31	0.51	0.39	1.00	1.00	0.95	0.97
	RawNet2 [30]	0.53	0.66	0.39	0.49	0.99	0.60	0.99	0.74
	P3 [34]	0.58	0.36	1.00	0.53	1.00	1.00	0.96	0.98
Image	UFD [26]	0.56	0.63	0.999	0.77	0.94	0.95	0.67	0.75
	DistilDIRE [27]	0.52	0.64	0.87	0.74	0.99	0.99	0.98	0.98
	NPR [28]	0.53	0.69	0.29	0.41	0.98	0.95	0.94	0.94

Original publication test data includes the following datasets for each model. Where multiple datasets are specified, the reported metrics are averages over these datasets. Genconvit: [35], [12], [36], [37]; FTCN: [36]; Styleflow: [36], [38], [39], [40]; AASIST, RawNet2, and P3 were all evaluated on the LA eval set of ASVspoof2019 [41]; UFD: [42] and subsets of LAION-400M [43] and AI generated images from latent diffusion models [44], Glide [45], and DALL·E mini [46] provided by the original publication [26]; DistilDIRE: ImageNet and AI generated images from Stable Diffusion v1 [44] and ADM [47] as specified in the original publication [27]; NPR: [48], [49], and the dataset from [26].

All off-the-shelf open-source models perform poorly on Deepfake-Eval-2024 (Table 5), with a max AUC of 0.58 across modalities and models. Further, many off-the-shelf models have an AUC lower than 0.5, worse than random guessing, suggesting that these models have perhaps learned to predict deepfakes based on spurious correlations that do not exist in real-world data.

The poor performance of open-source models on Deepfake-Eval-2024 offers a stark contrast to the exceptional performance of these models on the datasets that they were originally tested on (5). We observe an average drop in AUC of 50% for video, 48% for audio, and 45% for image models when evaluated on Deepfake-Eval-2024, as compared to the academic datasets that the models were originally tested on. This drastic difference in performance suggests that the academic deepfake detection datasets which the models were trained to perform well are not representative of the threat of contemporary deepfakes, underscoring the importance of up-to-date, challenging, in-the-wild deepfake datasets like Deepfake-Eval-2024.

Table 6: **Open-Source Model Finetuning Results**

Modality	Model	Accuracy	AUC	Precision	Recall	F1
Video	GenConViT [31]	0.75	0.82	0.78	0.65	0.71
	FTCN [32]	0.65	0.71	0.64	0.61	0.62
	Styleflow [33]	0.53	0.56	0.52	0.66	0.58
Audio	AASIST [29]	0.84	0.91	0.80	0.76	0.78
	RawNet2 [30]	0.82	0.88	0.82	0.91	0.86
	P3 [34]	0.86	0.92	0.80	0.82	0.81
Image	UFD [26]	0.63	0.56	0.63	1.00	0.77
	DistilDIRE [27]	0.61	0.56	0.64	0.87	0.73
	NPR [28]	0.69	0.73	0.74	0.78	0.76

When open-source deepfake detection models are finetuned on a subset of Deepfake-Eval-2024, performance improves (Table 6, Supplementary Table 11). However, the degree of improvement varies across models, suggesting that some model architectures may be less suited to adapt to the challenges of real-world deepfake detection. For example, the simple single layer UFD model learns to predict all data as fake after finetuning, and Styleflow and DistilDIRE also show limited improvement in AUC. However, this limited improvement could also be attributed to the relatively small finetuning set size. Although in most cases performance improves after finetuning, there is still significant room for improvement, with the peak accuracy reaching 0.75 for videos, 0.86 for audio, and 0.69 for images. These results suggest that in addition to more representative training datasets, new model paradigms may be needed for robust and reliable deepfake detection. For example, incorporating additional transfer learning and ensemble methods could enhance model performance by leveraging diverse feature representations.

Table 7: **Best Commercial Model Performance on Deepfake-Eval-2024**

Modality	Accuracy	AUC	Precision	Recall	F1
Video	0.78	0.79	0.77	0.77	0.77
Audio	0.89	0.93	0.89	0.84	0.87
Image	0.82	0.90	0.99	0.71	0.83

We find that commercial models exceed the performance of open-source models (Table 7). Commercial audio models appear to be the most accurate out of all three modalities, reaching 89% accuracy, whereas commercial video and image models had a max accuracy of 78% and 82% respectively. No commercial models that we evaluated had an accuracy of 90% or above, suggesting that even commercial models need significant improvement to address the challenge of real-world deepfakes with high accuracy. In addition, we note that open-source models finetuned on a subset of Deepfake-Eval-2024 approach the accuracy of commercial models (finetuned video model GenConVit has an accuracy of 75%, and finetuned audio P3 from [34] has an accuracy of 86%). This suggests that the competitive advantage of these commercial models may be derived primarily from training dataset curation.

In both open-source and commercial models, performance across modalities varies, with audio detectors consistently performing with higher accuracy than image and video detectors. This suggests that audio deepfakes may currently be more algorithmically distinguishable than video or image deepfakes, possibly due to the more constrained nature of audio manipulation techniques. However,

the strong performance of audio models may be confounded by the fact that commercial audio model detectors were used during labeling to help distinguish between real voice impersonators and ai-generated audio, thus potentially biasing the audio dataset towards a greater quantity of audio that is already easily detected by existing audio models.

6 Discussion and Conclusion

We present Deepfake-Eval-2024, an in-the-wild deepfake dataset which uniquely captures the challenging real-world threat of contemporary deepfakes through a collection of diverse and recent data gathered from TrueMedia.org users around the world. We evaluate the state-of-the-art of deepfake detection on Deepfake-Eval-2024 and find that both open-source models and commercial models need to improve in order to achieve high accuracy on real-world deepfake detection. Our results suggest that improvement through both novel model design and training dataset development is necessary.

We recommend that future research explores ensemble models and multi-modal detection strategies, as integrating diverse signals from audio, video, and images may provide a more comprehensive defense against increasingly sophisticated deepfake attacks. We hope that the release of Deepfake-Eval-2024 as the first large multi-modal in-the-wild deepfake dataset helps to catalyze multi-modal deepfake detection research.

Deepfake-Eval-2024 demonstrates the importance of challenging real-world data for the field of deepfake detection. Current academic deepfake detection benchmarking is not standardized and uses highly curated examples; our results show that most open-source academic models perform very well on their reported test sets. This makes it difficult to compare models and assess their abilities to detect fake media in the real world. We further show that open-source models lag significantly behind commercial detectors when evaluated on in-the-wild examples, but finetuning on even a small set of challenging data from Deepfake-Eval-2024 narrows this gap, underscoring the massive impact of real-world training data.

Our findings are a call to action for all deepfake detection practitioners to prioritize the production and sharing of challenging, real-world deepfake datasets. Although Deepfake-Eval-2024 is the largest and most diverse contemporary in-the-wild deepfake dataset today, much more work is needed to create deepfake datasets that are fully representative of deepfakes produced worldwide. Lastly, we emphasize that deepfakes are a constantly evolving threat and consistent vigilance is necessary to keep our models and datasets up-to-date and effective.

Acknowledgments and Disclosure of Funding

This work was made possible through by the incredible team at TrueMedia.org, including Alex Schokking, Art Min, Dawn Wright, Field Cady, James Allard, Kathy Thrailkill, Maryvel Dolotanora, Maxwell Bennett, Michael Bayne, Michael Langan, Molly Norris Walker, Paul Carduner, and Steve Geluso. We would like to thank Camp.org for the generous funding that supported this work.

References

[1] Christopher Whyte. Deepfake news: Ai-enabled disinformation as a multi-level public policy challenge. *Journal of Cyber Policy*, 5(2):199–217, 2020.

[2] Valdemar Danry, Joanne Leong, Pat Pataranutaporn, Pulkit Tandon, Yimeng Liu, Roy Shilkrot, Parinya Punpongsanon, Tsachy Weissman, Pattie Maes, and Misha Sra. Ai-generated characters: Putting deepfakes to good use. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22, New York, NY, USA, 2022. Association for Computing Machinery.

[3] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.

[4] Tina Brooks, G. Princess, Jesse Heatley, J. Jeremy, Scott Kim, M. Samantha, Sara Parks, Maureen Reardon, Harley Rohrbacher, Burak Sahin, S. Shani, S. James, T. Oliver, and V. Richard. Increasing threats of deepfake identities. Public-private analytic exchange program report, U.S. Department of Homeland Security, 2021.

[5] Sum and Substance Ltd. Identity fraud report 2024. Technical report, Sum and Substance Ltd., 2024.

[6] OpenAI. Sora system card, December 2024. Accessed January 9, 2025.

[7] Midjourney. Version 6.1, jul 2024. Accessed: 2024-01-09.

[8] Joel Frank, Franziska Herbert, Jonas Ricker, Lea Schönherr, Thorsten Eisenhofer, Asja Fischer, Markus Dürmuth, and Thorsten Holz. A representative study on human detection of artificially generated media across countries, 2023.

[9] C2PA. Coalition for content provenance and authenticity (c2pa), 2024. A Joint Development Foundation project formed through an alliance between Adobe, Arm, Intel, Microsoft and Truepic.

[10] Hannah Lee, Changyeon Lee, Kevin Farhat, Lin Qiu, Steve Geluso, Aerin Kim, and Oren Etzioni. The tug-of-war between deepfake generation and detection, 2024.

[11] Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, and Kong Aik Lee. Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2507–2522, 2023.

[12] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge dataset, 2020.

[13] Zhixi Cai, Shreya Ghosh, Aman Pankaj Adatia, Munawar Hayat, Abhinav Dhall, Tom Gedeon, and Kalin Stefanov. Av-deepfake1m: A large-scale llm-driven audio-visual deepfake dataset. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7414–7423, 2024.

[14] Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu. Forgerynet: A versatile benchmark for comprehensive forgery analysis. *arXiv preprint arXiv:2103.05630*, 2021.

[15] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2382–2390, 2020.

[16] Jiameng Pu, Neal Mangaokar, Lauren Kelly, Parantapa Bhattacharya, Kavya Sundaram, Mobin Javed, Bolun Wang, and Bimal Viswanath. Deepfake videos in the wild: Analysis and detection. In *Proceedings of The Web Conference 2021*, 2021.

[17] Nicolas M Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. Does audio deepfake detection generalize? *Interspeech*, 2022.

[18] Sarah Barrington, Matyas Bohacek, and Hany Farid. Deepspeak dataset v1.0, 2024.

[19] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection. In *CVPR*, 2020.

[20] Ricardo Reimao and Vassilios Tzerpos. For: A dataset for synthetic speech detection. *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–10, 2019.

[21] Hasam Khalid, Shahroz Tariq, and Simon S. Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset, 2021.

[22] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[23] Davis E. King. Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.*, 10:1755–1758, December 2009.

[24] iProov. Threat intelligence report 2024: The impact of generative AI on remote identity verification. Technical report, iProov, 2024.

[25] N. Kamali, K. Nakamura, A. Chatzimparmpas, J. Hullman, and M. Groh. How to distinguish AI-generated images from authentic photographs. *arXiv preprint arXiv:2406.08651*, 2024.

[26] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *CVPR*, 2023.

[27] Yewon Lim, Changyeon Lee, Aerin Kim, and Oren Etzioni. Distildire: A small, fast, cheap and lightweight diffusion synthesized deepfake detection. *arXiv preprint arXiv:2406.00856*, 2024.

[28] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection, 2023.

[29] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *arXiv preprint arXiv:2110.01200*, 2021.

[30] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. End-to-end anti-spoofing with rawnet2, 2021.

[31] Deressa Wodajo, Solomon Atnafu, and Zahid Akhtar. Deepfake video detection using generative convolutional vision transformer, 2023.

[32] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection, 2021.

[33] Jongwook Choi, Taehoon Kim, Yonghyun Jeong, Seungryul Baek, and Jongwon Choi. Exploiting style latent flows for generalizing deepfake video detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1133–1143, 2024.

[34] Xin Wang and Junichi Yamagishi. Can large-scale vocoded spoofed data improve speech spoofing countermeasure with a self-supervised front end?, 2023.

[35] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *International Conference on Computer Vision (ICCV)*, 2019.

[36] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *IEEE Conference on Computer Vision and Patten Recognition (CVPR)*, 2020.

[37] Pavel Korshunov and Sebastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection, 2018.

[38] Nick Dufour and Andrew Gully. Contributing data to deepfake detection research. Google AI Blog, 2019. [Accessed 30-07-2023].

[39] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019.

[40] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2889–2898, 2020.

[41] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Hector Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, Lauri Juvela, Paavo Alku, Yu-Huai Peng, Hsin-Te Hwang, Yu Tsao, Hsin-Min Wang, Sebastien Le Maguer, Markus Becker, Fergus Henderson, Rob Clark, Yu Zhang, Quan Wang, Ye Jia, Kai Onuma, Koji Mushika, Takashi Kaneda, Yuan Jiang, Li-Juan Liu, Yi-Chiao Wu, Wen-Chin Huang, Tomoki Toda, Kou Tanaka, Hirokazu Kameoka, Ingmar Steiner, Driss Matrouf, Jean-Francois Bonastre, Avashna Govender, Srikanth Ronanki, Jing-Xuan Zhang, and Zhen-Hua Ling. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech, 2020.

[42] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot...for now. In *CVPR*, 2020.

[43] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, 2021.

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[45] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2022.

[46] Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifullah, Tanishq Abraham, Phuc Le Khac, Luke Melas, and Ritobrata Ghosh. Dall·e mini, 7 2021.

[47] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[48] Chuangchuang Tan, Renshuai Tao, Huan Liu, and Yao Zhao. Gangen-detection: A dataset generated by gans for generalizable deepfake detection. `https://github.com/chuangchuangtan/GANGen-Detection`, 2024.

[49] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. *arXiv preprint arXiv:2303.09295*, 2023.

[50] Govind Mittal, Chinmay Hegde, and Nasir Memon. Gotcha: Real-time video deepfake detection via challenge-response, 2023.

[51] Kartik Narayan, Harsh Agarwal, Kartik Thakral, Surbhi Mittal, Mayank Vatsa, and Richa Singh. Df-platter: Multi-face heterogeneous deepfake dataset. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9739–9748, 2023.

[52] Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Li Yuan, Chengjie Wang, Shouhong Ding, et al. Df40: Toward next-generation deepfake detection. *arXiv preprint arXiv:2406.13495*, 2024.

[53] Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao. Audio deepfake detection: A survey, 2023.

[54] Joel Frank and Lea Schönherr. WaveFake: A Data Set to Facilitate Audio Deepfake Detection. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

[55] João C. Neves, Ruben Tolosana, Ruben Vera-Rodriguez, Vasco Lopes, Hugo Proença, and Julian Fierrez. Ganprintr: Improved fakes and evaluation of the state of the art in face manipulation detection. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):1038–1048, 2020.

[56] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *In Proceeding of IEEE Computer Vision and Pattern Recognition (CVPR 2020)*, Seattle, WA, 2020.

[57] Jordan J. Bird and Ahmad Lotfi. Cifake: Image classification and explainable identification of ai-generated synthetic images. *IEEE Access*, 12:15642–15650, 2023.

A Appendix: Supplementary figures

A.1 Related Work Supplementary Figures

Here we provide a detailed overview of popular deepfake detection datasets and compare them to Deepfake-Eval-2024.

A.1.1 Note on Overlap between Modality Datasets

Most video datasets only include manipulated or AI-generated frames from videos without accompanying real or fake audio [35, 36, 15, 19], while a few datasets provide audio-visual (AV) data [21, 13, 18]. For datasets with AV data, if it is possible to separate audio and video components and labels, we denote the datasets in Tables 8 and 9 with (A) or (V) to describe which part of the datasets we are reporting on. Similarly, there is often overlap between video and image datasets; some popular datasets used for image deepfake detection training and evaluation are composed of individual frames from video datasets [35, 15]. To avoid reporting duplicate datasets across modalities, we omit these from Table 10.

Table 8: **Survey of existing popular video deepfake detection datasets.** When duration values are not directly provided, values are estimated using several methods: * indicates calculation from frame count assuming 30fps (the most commonly encountered frame rate among published video datasets), † indicates derivation from average clip lengths, ‡ indicates values estimated from reported estimates, and ≈ indicates direct reported estimates.

Dataset	Year	# Real Files	# Fake Files	Real Media Duration (hrs)	Fake Media Duration (hrs)	Total Duration (hrs)	In-the-Wild
FaceForensics++ [35]	2019	1,000	4,000	4.71*	16.95*	21.66*	✗
Celeb-DF [36]	2019	590	5,639	2.13†	20.36†	22.49†	✗
DFDC [12]	2020	23,654	104,500	64.43	288.88	353.31	✗
WildDeepfake [15]	2020	3,805	3,509	-	-	10.93*	✓
DeeperForensics-1.0 [19]	2020	50,000	10,000	46.30*	116.67*	162.96*	✗
DF-W [16]	2021	0	1,869	0	48.83	48.83	✓
ForgeryNet [14]	2021	99,630	121,617	13.32*	13.50*	26.82*	✗
FakeAVCeleb (V) [21]	2021	500	19,000	1.08†	41.17†	42.25†	✗
GOTCHA [50]	2022	409	55,838	3.13‡	-	-	✗
DF-Platter [51]	2023	764	132,496	-	-	≈736.08	✗
AV-Deepfake1M [13]	2023	286,721	860,039	-	-	1,886	✗
DeepSpeak [18]	2024	6,226	6,799	17	26	44	✗
DF40 [52]	2024	0	100k+	-	-	-	✗
Ours	2024	1,072	964	28.9	16.2	45.1	✓

Table 9: **Survey of existing popular audio deepfake detection datasets.** Similar to video datasets, when duration values are not directly provided, values are estimated using several methods: † indicates derivation from average clip lengths, ≈ indicates direct reported estimates, and § indicates values provided by a survey paper [53].

Dataset	Year	# Real Files	# Fake Files	Real Media (hrs)	Fake Media (hrs)	Total Duration (hrs)	In-the-Wild	# Languages
FoR [20]	2019	108,256	87,285	151.86†	56.98†	208.84†	✗	1
FakeAVCeleb (audio) [21]	2021	500	10,500	1.08†	22.75†	23.83†	✗	1
WaveFake [54]	2021	0	117,985	0	≈196	≈196	✗	2
ASVspoof 2021 (DF subset) [11]	2021	20,637	572,616	-	-	325.8§	✗	1
In-the-Wild [17]	2022	-	-	20.7	17.2	37.9	✓	1
Ours	2024	1,167	814	36.6	19.9	56.5	✓	42

Table 10: **Survey of existing popular image deepfake detection datasets.**

Dataset	Year	# Real Files	# Fake Files	# Total Files	In-the-Wild	# Generation Techniques	Resolution
iFakeFaceDB [55]	2019	0	≈87,000	≈87,000	✗	2	224×224
DFFD [56]	2020	58,703	240,336	299,039	✗	4	1,024×1,024
ForenSynths [42]	2020	36,200	36,200	72,400	✗	11	256×256
ForgeryNet (image) [14]	2021	1,438,201	1,457,861	2,896,062	✗	15	Varies
DiffusionForensics [49]	2023	232,000	232,000	464,000	✗	11	256×256
CIFAKE [57]	2024	60,000	60,000	120,000	✗	1	32×32
Ours	2024	767	1,208	1,975	✓	Many	Varies

A.2 Dataset Supplementary Figures

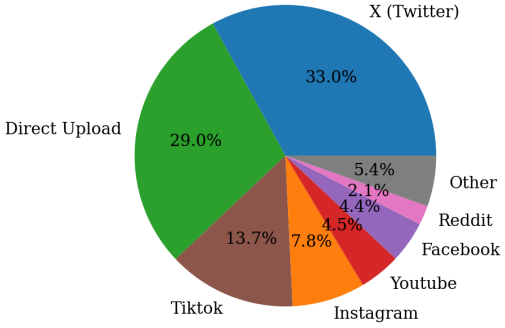

Figure 4: Origins of data in Deepfake-Eval-2024 combined for all modalities. Direct upload indicates that the media was uploaded directly to TrueMedia.org by a user, instead of the user providing a link to a social media website.

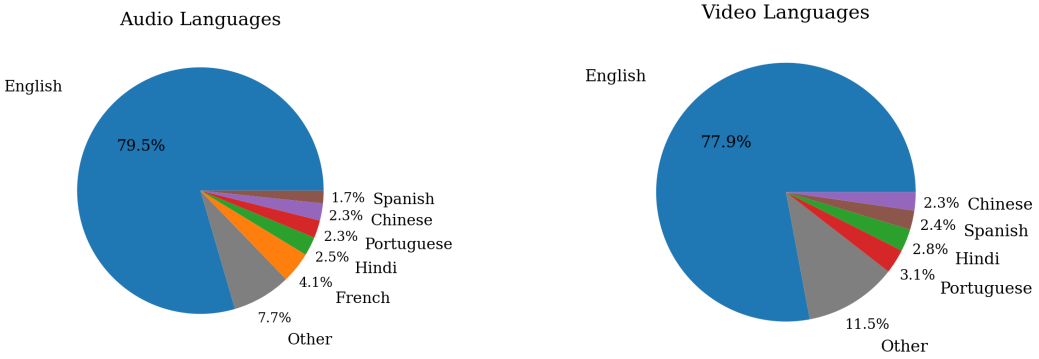

Figure 5: Language distributions for audio and video content.

A.3 Results Supplementary Figures

Table 11: Combined Open-Source Model Finetuning Results Across Modalities

Modality	**Model**	**AUC**	**Accuracy**	**Precision**	**Recall**	**F1**	**FPR**	**FNR**	**EER (%)**
Video	Genconvit	0.18	0.75	0.78	0.65	0.71	0.17	0.35	-
	FTCN	0.71	0.65	0.64	0.61	0.62	0.30	0.39	-
	Styleflow	0.56	0.53	0.52	0.66	0.58	0.61	0.34	-
Audio	AASIST	0.906	0.836	0.797	0.761	0.778	0.118	0.239	16.99
	P3	0.920	0.855	0.802	0.818	0.810	0.122	0.182	15.38
	RawNet2	0.876	0.817	0.818	0.908	0.860	0.334	0.092	20.91
Image	UFD	0.56	0.63	0.63	1.00	0.77	1.00	0.00	-
	DistilDIRE	0.56	0.61	0.64	0.87	0.74	0.85	0.13	-
	NPR	0.73	0.69	0.74	0.78	0.76	0.46	0.22	-

B Appendix: Labeling Codebooks

We present the labeling criteria for all modalities. The complementary examples mentioned in this section can be found here.

B.1 Image labeling codebook

AI generated video/image traits adapted from Kamali et al. [25]

Real (no AI manipulation)	Fake (AI manipulation)	Unknown
Original, reputable source confirms no AI manipulation	If any portion is AI, then entire item is fake	Cartoons, animations, and photoshopped images such as swapped signs, hats, or t-shirts (unless evidence of AI manipulation)
Fact-checking source confirms no AI manipulation	Fact-checking source confirms AI manipulation	Unable to confirm AI manipulation or not
Real media in which a person is lying, or real images presented out of context and misleading	Contains 3+ of the following AI traits: • Stylistic Artifacts: hyper-realistic or inconsistent detail, smooth or plastic/waxy looking skin (Example 1), cartoonish appearance (Example 2), too perfect, inconsistent lighting or reflections etc. • Anatomical Implausibilities: irregular pupils, mangled/ missing/disproportionate limbs, incorrect/merged fingers, inconsistent facial features of famous personas compared to their real images etc. • Sociocultural Implausibilities: unlikely scenarios or historical inaccuracies • Functional Implausibilities: misspelled/backwards text, impossible words, impossible structure of buildings, vehicles, food etc.	
	Face swapping and face morphing for media created in 2023 or later	Face swapping and face morphing for media created prior to 2023
Content from film or TV with no evidence of AI manipulation		
Media manipulation using text and non-AI-generated image overlays such as stickers (Example 3 and Example 4)		

B.2 Video Codebook

AI generated video/image traits adapted from Kamali et al. [25]

Real (no AI manipulation)	Fake (AI manipulation)	Unknown
Lips and mouth are crisp, clear, nuanced, and match sound perfectly.	Lips are roughly in sync Example 5 with audio, but clearly not crisp or natural	
Lips and audio are completely out of sync Example 6, (and you find original source to confirm that audio was dubbed onto a real video)		Lips and audio are completely out of sync, but you cannot find the original source to confirm if video is real or manipulated
Located original source and confirmed no AI manipulation	Located original source and confirmed AI manipulation was used	Video quality is too poor to determine if mouth movements are crisp and nuanced
Highly edited Example 7, but every individual clip is real		Filters Example 8, effects, GIFs
Real person is obviously "lip syncing" Example 9 or parody, no evidence of AI manipulation.		
Talking head Example 10 pasted on background (predominant in many tiktok videos)		

B.3 Audio Codebook

Real (no AI manipulation)	Fake (AI manipulation)	Unknown
Lips and mouth are crisp, clear, nuanced, and match sound perfectly.	If lip sync is off AND 2 or more audio models say >80%	If lip sync is off and you cannot discern if AI or human impersonator
Music, silence, and sound effects were labeled as real unless there was other evidence of AI manipulation.	Audio-Only: if 2 or more models say >80% PLUS there's some additional reason to believe it's fake (ie. the audio quality sounds synthetic) Example 11	Voice is off camera and unable to locate original source
human impersonator Example 12		