

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

Supplementary Material

A. Appendix: Supplementary Materials

A.1. Related Work Supplementary Figures

Here we provide a detailed overview of popular deepfake detection datasets and compare them to Deepfake-Eval-2024. We focus on popular datasets released prior to the release of Deepfake-Eval-2024 (March 4, 2025).

A.1.1. Overlap between Modality Datasets

Most video datasets only include manipulated or AI-generated frames from videos without accompanying real or fake audio [23, 31, 47, 67], while a few datasets provide audio-visual (AV) data [3, 6, 27]. For datasets with AV data, if it is possible to separate audio and video components and labels, we denote the datasets in Tables S1 and S2 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 [47, 67]. To avoid reporting duplicate datasets across modalities, we omit these from Table S3.

Table S1. Survey of existing popular video deepfake detection datasets.

Dataset	Year	# Real Files	# Fake Files	Real Media Duration (hrs)	Fake Media Duration (hrs)	Total Duration (hrs)	In-the-Wild
FaceForensics++ [47]	2019	1,000	4,000	4.71 [*]	16.95 [*]	21.66 [*]	✗
Celeb-DF [31]	2019	590	5,639	2.13 [†]	20.36 [†]	22.49 [†]	✗
DFDC [12]	2020	23,654	104,500	64.43	288.88	353.31	✗
WildDeepfake [67]	2020	3,805	3,509	-	-	10.93 [*]	✓
DeeperForensics-1.0 [23]	2020	50,000	10,000	46.30 [*]	116.67 [*]	162.96 [*]	✗
DF-W [41]	2021	0	1,869	0	48.83	48.83	✓
ForgeryNet [18]	2021	99,630	121,617	13.32 [*]	13.50 [*]	26.82 [*]	✗
FakeAVCeleb (V) [27]	2021	500	19,000	1.08 [†]	41.17 [†]	42.25 [†]	✗
GOTCHA [34]	2022	409	55,838	3.13 [‡]	-	-	✗
RWDF-23 [7]	2023	0	2,000	-	48.15	48.15	✓
DF-Platter [36]	2023	764	132,496	-	-	≈736.08	✗
AV-Deepfake1M [6]	2023	286,721	860,039	-	-	1,886	✗
DeepSpeak [3]	2024	6,226	6,799	17	26	44	✗
DF40 [64]	2024	0	100k+	-	-	-	✗
Ours	2024	1,072	964	28.9	16.2	45.1	✓

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.

Table S2. Survey of existing popular audio deepfake detection datasets.

Dataset	Year	# Real Files	# Fake Files	Real Media (hrs)	Fake Media (hrs)	Total Duration (hrs)	In-the-Wild	# Languages
FoR [44]	2019	108,256	87,285	151.86 [†]	56.98 [†]	208.84 [†]	✗	1
ASVspoof (LA subset) [59, 63]	2019	12,483	108,978	5.20 [†]	45.41 [†]	50.61 [†]	✗	1
FakeAVCeleb (audio) [27]	2021	500	10,500	1.08 [†]	22.75 [†]	23.83 [†]	✗	1
WaveFake [16]	2021	0	117,985	0	≈196	≈196	✗	2
ASVspoof (DF subset) [33]	2021	20,637	572,616	-	-	325.8 [§]	✗	1
In-the-Wild [35]	2022	-	-	20.7	17.2	37.9	✓	1
SpoofCeleb [25]	2024	≈ 248,000	≈ 2,439,292	310 [†]	3,049 [†]	3,359 [†]	✗	1
Ours	2024	1,167	814	36.6	19.9	56.5	✓	42

Datasets that are not publicly available yet (such as ASVspoof5) are not included. 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 [65].

Table S3. Survey of existing popular image deepfake detection datasets.

Dataset	Year	# Real Files	# Fake Files	# Total Files	In-the-Wild	# Generation Techniques	Resolution
iFakeFaceDB [37]	2019	0	≈87,000	≈87,000	✗	2	224×224
DFFD [9]	2020	58,703	240,336	299,039	✗	4	1,024×1,024
ForenSynths [57]	2020	36,200	36,200	72,400	✗	11	256×256
ForgeryNet (image) [18]	2021	1,438,201	1,457,861	2,896,062	✗	15	Varies
DiffusionForensics [61]	2023	232,000	232,000	464,000	✗	11	256×256
CIFAKE [4]	2024	60,000	60,000	120,000	✗	1	32×32
Ours	2024	767	1,208	1,975	✓	Many	Varies

A.2. Dataset Supplementary Figures

Table S4. 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 S5. 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 S6. Deepfake-Eval-2024 Image Summary Statistics

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

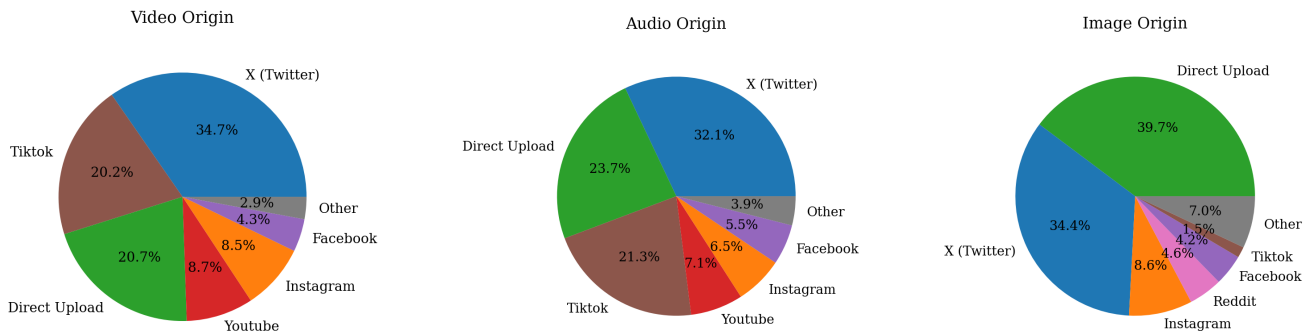


Figure S1. Origins of data in Deepfake-Eval-2024 separated by modality. In total, media was shared from 88 different web-domain names. Direct upload indicates that the media was uploaded directly to TrueMedia.org by a user, instead of the user providing a link to a social media website.

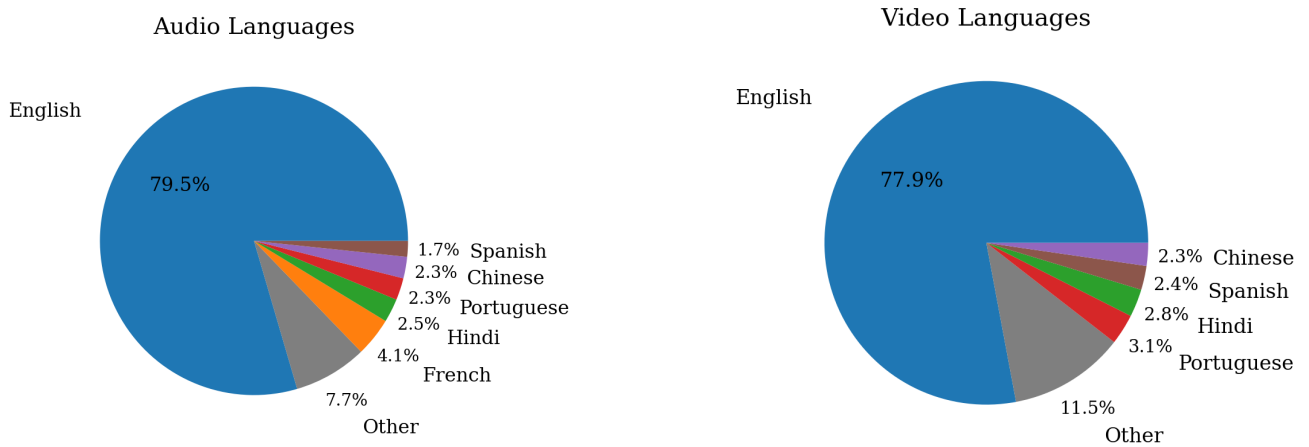


Figure S2. Language distributions for audio and video content.

Table S7. Demographic estimates of image and video subjects in Deepfake-Eval-2024.

Race/Ethnicity	Percentage
White	54.6
Multiple	16.8
Black	9.2
East Asian	6.6
South Asian	5.3
Unknown - non-white	5.0
Latino - non-white	2.4

A.2.1. Demographic Estimates of Image and Video Subjects

To characterize the demographic composition of Deepfake-Eval-2024, we manually annotated a randomly sampled 10% subset of the visual data for perceived race/ethnicity. We report the demographic estimates from this analysis in Supp. Table S7. We do not make definitive claims about subject identity.

A.3. Experimental Compute

We evaluated and finetuned all models on a single AWS A10 GPU on an g5.2xlarge instance or a single GCP L4 GPU on a g2-standard-8 instance. We found no difference between evaluating and finetuning models on the AWS or GCP instance.

A.4. Results Supplementary Figures

Table S8. Complete Off-the-Shelf Open-Source Model Results Across Modalities

Modality	Model	AUC	Accuracy	Precision	Recall	F1	FPR	FNR	EER (%)
Video	GenConViT	0.63	0.60	0.60	0.50	0.54	0.31	0.50	-
	FTCN	0.50	0.51	0.51	0.67	0.41	0.33	0.66	-
	Styleflow	0.51	0.52	0.54	0.43	0.48	0.39	0.56	-
Audio	AASIST	0.43	0.42	0.31	0.51	0.39	0.63	0.49	55.22
	RawNet2	0.53	0.48	0.66	0.39	0.49	0.36	0.61	48.20
	P3	0.58	0.36	0.36	1.00	0.53	1.00	0.00	43.00
Image	UFD	0.56	0.63	0.63	0.999	0.77	0.99	0.001	-
	DistilDIRE	0.52	0.61	0.64	0.87	0.74	0.83	0.13	-
	NPR	0.53	0.47	0.69	0.29	0.41	0.22	0.71	-

Table S9. Complete Open-Source Model Finetuning Results Across Modalities

Modality	Model	AUC	Accuracy	Precision	Recall	F1	FPR	FNR	EER (%)
Video	Genconvit	0.82	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
	RawNet2	0.876	0.817	0.818	0.908	0.860	0.334	0.092	20.91
	P3	0.920	0.855	0.802	0.818	0.810	0.122	0.182	15.38
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	0.55	0.58	0.61	0.81	0.70	0.76	0.19	-

Table S10. Best Commercial Model Performance on Deepfake-Eval-2024

Modality	Model Rank	Accuracy	AUC	Precision	Recall	F1
Video	#1	0.78	0.79	0.77	0.77	0.77
	#2	0.66	0.70	0.78	0.47	0.59
	#3	0.59	0.64	0.59	0.69	0.64
Audio	#1	0.89	0.93	0.89	0.84	0.87
	#2	0.88	0.93	0.88	0.80	0.84
	#3	0.86	0.90	0.83	0.83	0.84
Image	#1	0.86	0.88	0.97	0.79	0.87
	#2	0.82	0.90	0.99	0.71	0.83
	#3	0.77	0.89	0.98	0.64	0.77

Table S11. Open-Source Multimodal Model Results

Model	AUC on Deepfake-Eval-2024	AUC on Original Publication Test Dataset
AVF [15]	0.58	0.945 (FakeAVCeleb), 0.87 (KoDF)
FGI [2]	0.42	0.845 (FakeAVCeleb), 0.98 (DFDC)

A.5. Dataset Access and Ethics

Deepfakes pose an established threat to society, which is why the release of updated deepfake detection benchmarks is important for improving our response to this threat. However, there is a potential for released datasets to be used with malicious intentions to create deepfake generation technologies that are more realistic and that evade existing detectors. As such, we gate access to this dataset through Huggingface to individuals verifiably at research institutions or doing work related to deepfake detection. Prospective users must accept a terms of use agreement. Then they must provide an institutional / company email address and link that provides additional evidence of work related to deepfake detection, which is used for verification. We intend to maintain this system of access gating until May 2027, at which point generative AI is likely to have advanced far beyond the deepfakes represented in Deepfake-Eval-2024, and thus there will be a much lower risk of making Deepfake-Eval-2024 available to all individuals. Starting in May 2027, we will modify the access process so that any individual who agrees to the Terms of Use can get access to the dataset. Further, in the Terms of Use, we state that we release the dataset with a CC-BY-SA-4.0 license, designating it primarily as a research artifact. While the dataset is freely available, users are responsible for ensuring its legal use in commercial settings. Users must independently verify compliance with applicable laws before employing the dataset for commercial purposes.

B. Appendix: Labeling Criteria

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

B.1. Image labeling codebook

AI-generated video/image traits adapted from Kamali et al. [26].

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 photo-shopped 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: <ul style="list-style-type: none">• 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. [26]

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
Audio without speech such as 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 $> 80\%$ likelihood of fake PLUS there’s some additional reason to believe it’s fake (ie. the audio quality sounds synthetic, or sociocultural implausability) Example 11	Voice is off camera and unable to locate original source
Human impersonator Example 12		

C. Appendix: Verification Process

Reverse Image Search

- If a media item did not contain common AI traits to help us determine ground truth, we used Google’s reverse image search to locate the original source of the item, or to find a professional fact-checking source that confirmed the item’s ground truth.

Source Trustworthiness

- When we located the original source or fact-checking source for an item, we used tools such as All Sides and Ad Fontes Media Bias to judge the trustworthiness of the source before determining the ground truth.

ChatGPT

- While we did not trust GPT implicitly, we did use it to point us in the right direction. For example, if a video showed Kamala Harris saying “xyz,” we used the following prompt as a first step to determine its veracity: “Did Kamala Harris say ‘xyz?’ Give me 3 reputable sources confirming or denying this claim.”

Google

- We used Google Search to find primary sources confirming or denying media claims. For example, if a video showed Donald Trump saying “they’re eating the pets of the people who live there,” we ran the search “Did Trump say ...” Or, if an image or video depicted Joe Biden falling asleep at a press conference, we ran the search “Did Biden fall asleep at ...” The results often pointed us to primary sources that we used to determine ground truth.

C.1. Reverse Image Search Verification Process

	Click on “See Exact Matches”	
No Match Found	If no match and no clues, mark as “Unknown”	
Match on Unknown Source	If match is found on a lesser-known site, check if the image is credited to a reputable source (AP, Reuters, etc.)	If credited, confirm by checking the site. If the site is legitimate, mark as “Real.”
Match on Social Media	If found on social media, read comments for clues.	If comments suggest it is fake due to artifacts in the media, mark as “Fake.” If credible, mark as “Real.”
Verified Source	If found on a reputable source’s social media (NBC, White House, etc.), mark as “Real.”	
Edited Media	If you find edited media (<i>e.g.</i> , face swapped or text altered in a sign), pay close attention to details.	Mark as “Fake.”