MAVOS-DD: Multilingual Audio-Video Open-Set Deepfake Detection Benchmark

Florinel-Alin Croitoru[◇]
University of Bucharest

Vlad Hondru[◇]
University of Bucharest

Marius Popescu
University of Bucharest

Radu Tudor Ionescu[*]
University of Bucharest

Fahad Shahbaz Khan
MBZ University of Artificial Intelligence
Linköping University

Mubarak Shah
University of Central Florida

Abstract

We present the first large-scale open-set benchmark for multilingual audio-video deepfake detection. Our dataset comprises over 250 hours of real and fake videos across eight languages, with 60% of data being generated. For each language, the fake videos are generated with seven distinct deepfake generation models, selected based on the quality of the generated content. We organize the training, validation and test splits such that only a subset of the chosen generative models and languages are available during training, thus creating several challenging open-set evaluation setups. We perform experiments with various pre-trained and fine-tuned deepfake detectors proposed in recent literature. Our results show that state-of-the-art detectors are not currently able to maintain their performance levels when tested in our open-set scenarios. We publicly release our data and code at: `https://huggingface.co/datasets/unibuc-cs/MAVOS-DD`.

1 Introduction

The rapid progress in image, audio and video synthesis technologies has enabled the creation of realistic visual content from textual descriptions [15, 49, 53, 55, 57] and the convincing manipulation of people's identities [8, 35, 44, 51] and expressions [9, 30, 62, 64, 69, 70, 77]. This has led to a surge of innovative applications across various industries, including marketing and film making. However, these breakthroughs have also fueled the rise of malicious uses, particularly in generating deceptive synthetic audio-visual content, commonly known as deepfakes [16]. Alarmingly, a recent report shows that the incidence of deepfake-related fraud increased by a factor of 10 between 2022 and 2023[2]. In this landscape, the ability to reliably identify forged video material is more crucial than ever.

A significant body of research has emerged in response to the rising number of deepfake-related manipulation and fraud cases, aiming to detect manipulated content using advanced deep learning techniques, such as convolutional neural networks [3, 12, 14, 38, 42, 54], transformers [31, 50, 52, 58, 74, 78], and hybrid approaches [6, 11, 13, 24, 65, 76]. These methods have achieved remarkable results, often surpassing 99% accuracy on existing benchmarks [16], such as Celeb-DF [45] and FaceForensics++ [56]. Nevertheless, most evaluations are carried out in controlled environments

[*]Corresponding author: `raducu.ionescu@gmail.com`. [◇]Equal contribution.

[2]Sumsub Expert Roundtable: The Top KYC Trends Coming in 2024

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

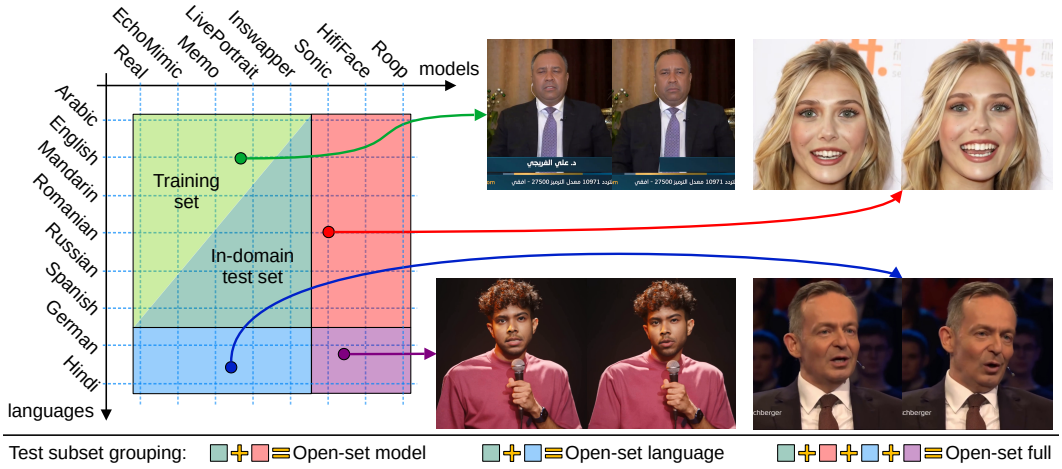


Figure 1: In MAVOS-DD, the training set and *in-domain* test set contain real and fake videos sampled from the same distribution, comprising six languages and four generative models. The *open-set model* test set extends the in-domain test set with fake samples generated by unseen models (Sonic, HifiFace, Roop). The *open-set language* test set extends the in-domain test set with samples in unseen languages (German and Hindi). The *open-set full* test set adds samples generated by unseen models in unseen languages. One fake sample from each data distribution is shown on the right-hand side. Best viewed in color.

where the synthetic and authentic samples in training and testing originate from the same video manipulation tools. This in-domain evaluation setup significantly inflates detection performance and fails to represent real-world conditions, where neither the manipulated technique nor the subject is known in advance.

To address this gap, we propose a new benchmark for evaluating audio-video deepfake detection models in a multilingual open-world setup. Our benchmark, MAVOS-DD, comprises over 35K fake and 25K real videos, totaling over 250 hours of video across eight languages: Arabic, English, German, Hindi, Mandarin, Romanian, Russian and Spanish. The fake samples are generated by seven state-of-the-art deepfake generation methods based on different approaches: talking head (EchoMimic [9], Memo [75], Sonic [32]), portrait animation (LivePortrait [25]), face swap (Inswapper[3], HifiFace [68], Roop[4]). As shown in Figure 1, we create a multi-perspective open-set benchmark. The training set comprises samples in six languages (excluding German and Hindi), where the fake samples are generated by four methods (excluding Sonic, HifiFace and Roop). We prepare an in-domain (closed) test set that is sampled from the same distribution as the training data. In addition, we create three open-set test sets: (i) *open-set model* extends the in-domain test set with fake samples generated by unseen models; (ii) *open-set language* adds German and Hindi samples to the in-domain test data; (iii) *open-set full* adds samples generated by unseen models in German and Hindi.

We perform extensive experiments using both pre-trained and fine-tuned deep fake detectors [52, 71, 80], analyzing their performance on both in-domain and open-set scenarios. While these models work well under in-domain conditions, two of them surpassing an accuracy threshold of 90%, their effectiveness drops significantly in the open-set setups. The reported performance gaps highlight a critical limitation of current deepfake detection models, namely the poor generalization across deepfake generation models and languages.

In summary, our contribution is twofold:

- We present MAVOS-DD, a comprehensive multilingual open-set benchmark for audio-video deepfake detection, encompassing over 250 hours of authentic and synthetic videos across eight languages.
- We conduct a thorough evaluation of state-of-the-art deepfake detectors, uncovering substantial performance degradation when models are tested in open-world setups, thereby emphasizing the need for more robust and generalizable detection techniques.

[3] https://github.com/deepinsight/insightface
[4] https://github.com/s0md3v/roop

Table 1: Comparison between MAVOS-DD and other video and audio-video (multimodal) datasets. MAVOS-DD is the largest dataset from multilingual audio-video open-set deepfake detection.

Dataset	File count		Length (h)			#Generative methods	#Languages	Open-set	Multimodal
	#Real	#Fake	Real	Fake	Total				
FaceForensics++ [56]	1,000	4,000	4.7	17.0	21.7	4	0	✗	✗
DFDC [18]	23,654	104,500	64.4	288.9	353.3	5	0	✗	✗
DeeperForensics [33]	50,000	10,000	46.3	116.7	163.0	1	0	✗	✗
ForgeryNet [28]	99,630	121,617	13.3	13.5	26.8	15	0	✗	✗
Celeb-DF [45]	590	5,639	2.1	20.4	22.5	1	0	✗	✗
WildDeepfake [79]	3,805	3,509	-	-	10.9	-	0	✗	✗
FakeAVCeleb [37]	500	19,500	1.1	41.2	42.3	3	1	✗	✓
DeepSpeak [4]	6,226	6,799	17.0	26.0	44.0	10	1	✗	✓
Deepfake-Eval-2024 [7]	1,072	964	28.9	16.2	45.1	-	49	✗	✓
MAVOS-DD (ours)	25,195	35,169	91.1	161.4	252.5	7	8	✓	✓

2 Related Work

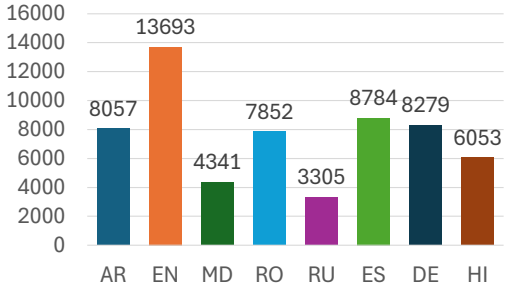
The field of deepfake generation has seen significant advancements in recent years [16], particularly with the rise of diffusion models [15, 29, 55, 57, 59]. In parallel, considerable research has been devoted to developing effective detection techniques [16, 52, 71, 80] to counter the negative effects of deepfake media. In addition, substantial efforts have been made to construct datasets for deepfake detection [18, 33, 37, 45, 56], thereby facilitating research in this domain.

Audio-visual deepfake detection. Traditional deepfake detection methods are unimodal, focusing solely on either visual artifacts, e.g. abnormal facial textures [21, 40, 42] and inconsistent lighting [23], or audio inconsistencies, e.g. speech prosody [2, 5, 63], frequency patterns [20, 60, 72, 73], and voice cloning artifacts [22, 48]. With generation methods becoming more capable, it is essential to leverage both visual and auditory modalities to improve the robustness and reliability of the forgery detection models [52, 71, 80]. Aside from unimodal cues, utilizing multimodal (audio-visual) information can naturally capitalize on the misalignment between the two modalities by examining if the audio and video signals are coherent and temporally aligned, e.g. in terms of lip movements [1, 78] or facial expressions [26].

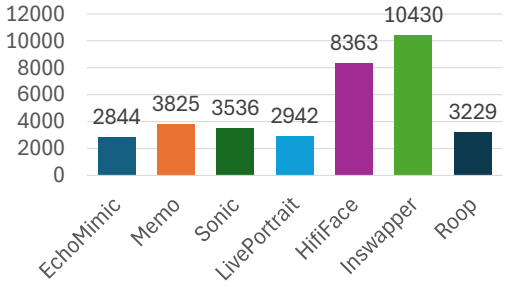
Early works on audio-visual deepfake detection used convolutional architectures [14, 38, 54]. For example, Multimodaltrace [54] extracts separate features from audio and video with residual blocks, fuses the resulting representations and further processes them to make the final prediction. Kihal *et al.* [38] also employ individual CNN-based feature extractors, but use a Random Forest model to predict the final label.

Recent works opted for architectures that leverage transformers, not only because of their higher performance, but also because of the inherent mechanism that enables fusing the information from two modalities using cross-attention modules [31, 50, 52, 58, 74, 78]. Zhou *et al.* [78] detect inconsistencies between the two modalities (focusing on lip movements and speech) by aligning their low-level latent representations and fusing them through a cross-modal attention mechanism. Nie *et al.* [50] employ two pre-trained frozen ViTs [19] to extract features, with only the `[CLS]` tokens being used for classification. To bridge the gap between modalities, the audio information is integrated into the visual tokens using an audio-distilled cross-modal interaction module. Furthermore, the authors try to detect high-frequency forgery artifacts by biasing the queries, keys, and values with learnable parameters.

Audio-visual deepfake datasets. While the advancement of deepfake generation methods has led to the development of detection methods to defend against deepfakes, it has also driven the need for extensive datasets. In the beginning, datasets comprising data from a single modality were created for both visual (image and video) data [10, 17, 18, 28, 44, 45, 56, 79] and audio data [46, 66]. Nevertheless, with the rise of multimodal models, the availability of audio-visual datasets [4, 7, 37, 41] has become essential.



(a) Number of real and deepfake videos per language.



(b) Number of deepfake videos generated with each method.

Figure 2: Distribution of videos per language and per generative method. MAVOS-DD comprises videos in eight languages, generated with seven methods. The languages are coded as follows: Arabic (AR), English (EN), German (DE), Hindi (HI), Mandarin (MD), Romanian (RO), Russian (RU) and Spanish (ES).

We present a comprehensive comparison of MAVOS-DD with other video and multimodal datasets in Table 1. DFDC [18] is among the largest video dataset for deepfake detection. However, multimodal datasets, such as FakeAVCeleb [37] and Deepfake-Eval-2024 [7] are not as large. FakeAVCeleb [37] is based on two face swapping methods and a facial reenactment method for their synthetic English-speaking videos. While DeepSpeak [4] tries to excel by employing 10 generative methods, Deepfake-Eval-2024 [7] stands out by having videos in 49 languages, although 80% is English.

One of the main limitations of the deepfake detection methods is their ability to generalize to synthetic samples generated with different methods. To this end, MAVOD-DD contains samples obtained with a variety of generative methods to facilitate training robust detection models, but also to thoroughly evaluate their ability to generalize to unseen methods. Moreover, with only one exception [7] from concurrent literature, existing datasets do not focus on the multilingual aspect of audio-visual content. Chandra *et al.* [7] collect the dataset from the web, so there is no control over the generative methods and languages. In contrast, our dataset enables an open-set evaluation in terms of both generative models and languages. Furthermore, our dataset comprises $10\times$ more deepfake content (161 hours vs. 16 hours), which enables the training of very deep models with higher generalization capacity. Although their videos span 49 languages, 80% of all videos are in English (each other language representing less than 0.5% of the dataset). In this regard, MAVOS-DD provides a more even distribution across languages (see Fig. 2a). Overall, the comparison in Table 1 shows that MAVOS-DD is the largest dataset from multilingual audio-video open-set deepfake detection.

3 Dataset

Overview. Our main contribution is MAVOS-DD, a large-scale deepfake dataset consisting of 60,364 real and synthetic videos, totaling 252 hours of content across eight different languages. The synthetic content is generated using seven state-of-the-art methods: EchoMimic [9], Memo [75], Sonic [32], LivePortrait [25], Inswapper, HifiFace [68], and Roop. The deepfake methods cover three key generative tasks: talking-head generation [9, 32, 75], facial expression transfer [25], and face swapping [68]. This coverage ensures a diverse and realistic set of generated videos. The main reason for using recent generative methods is to create a challenging dataset. Yet, another level of complexity is added through the fact that the audio-video samples cover eight languages: Arabic (AR), English (EN), German (DE), Hindi (HI), Mandarin (MD), Romanian (RO), Russian (RU) and Spanish (ES). We present the video distribution per language and per generative method in Figure 2a and Figure 2b, respectively. Note that real videos are naturally included in the distribution of videos per language, but not in the distribution of videos per generative method. The distribution per language is influenced by the number of real videos that we were able to collect for each language, while the distribution per method is influenced by the speed of each generative method. The total time required to generate all videos included in MAVOS-DD amounts to roughly 88 days (time measured on a computer with an Intel i9-14900K CPU with 192 GB of RAM and an Nvidia RTX 4090 GPU with 24 GB of VRAM).

We define official training, validation, and test splits for various evaluation scenarios, as illustrated in Figure 1. The first scenario, referred to as *in-domain* evaluation, uses a test set comprising the same

Table 2: Number of real and fake videos included in the training, validation and test splits of MAVOS-DD. The test data is divided into four subsets, which generate an in-domain evaluation scenario and three open-set evaluation scenarios. The core set includes six languages (Arabic, English, Mandarin, Romanian, Russian, Spanish) and four methods (EchoMimic, Memo, LivePortrait, Inswapper). The extra languages are German and Hindi. The extra models are Sonic, HifiFace and Roop. The length (in hours) of the real and fake content in each split is reported in the last column.

Split		Video type	File count				Total count	Total length (h)
			Core set	Extra languages	Extra models	Extra models & languages		
Train		Real	10,297	0	0	0	10,297	38.5
		Fake	9,473	0	0	0	9,473	45.4
Validation		Real	1,715	0	0	0	1,715	6.5
		Fake	1,580	0	0	0	1,580	8.1
Test	In-domain	Real	5,185	0	0	0	5,185	19.3
		Fake	4,701	0	0	0	4,701	23.4
	Open-set language	Real	5,185	7,998	0	0	13,183	46.3
		Fake	4,701	4,287	0	0	8,988	46.7
	Open-set model	Real	5,185	0	0	0	5,185	19.3
		Fake	4,701	0	13,081	0	17,782	70.7
	Open-set full	Real	5,185	7,998	0	0	13,183	46.4
		Fake	4,701	4,287	13,081	2,047	24,116	107.5

languages and generative methods as the training set. The second and third scenarios, namely *open-set model* and *open-set language*, expand the in-domain test set to include samples generated by unseen models or unseen languages, respectively. The final scenario, called *open-set full*, includes samples generated by unseen models in unseen languages, presenting the most challenging evaluation setting. We present detailed statistics about MAVOS-DD and its splits in Table 2. The training and validation splits do not include videos in German or Hindi, as these languages are reserved exclusively for the test set to support open-set evaluation. Overall, the number of real and fake samples is relatively balanced. However, the *open-set model* and *open-set full* splits contain a larger number of fake samples, as they comprise synthesized videos from three additional generative methods that are not present in the training set, as illustrated in Figure 1.

Real videos. We collect real videos from YouTube, primarily sourcing content from popular news channels or street interviews in each target language (such as EasyLanguages[5]) Additionally, we include videos from well-known channels specific to each country and language, although these are not our primary focus, as they tend to lack the diversity of speaker identities found in news broadcasts. After downloading, we apply the TalkNet active speaker detection model [61] to segment the videos into shorter clips, each featuring a single speaking individual. As the process to acquire the videos and split them into smaller videos is automatic, there are some instances where the videos do not contain any humans, i.e. faces. In order to filter these out, for each video, we apply a face detector [34] on individual frames (using a step of 15 frames) and eliminate those videos that do not have a face for more than half of the evaluated frames. The final dataset comprises 25,195 high-quality videos, with resolutions ranging from 256×256 to 1920×1080, amounting to a total of 91 hours of real content.

Deepfake videos. Deepfake generation typically involves a source identity image, representing the face that is manipulated by the generative model. We take these identities from multiple sources in our experiments. The first source is a set of 500 portraits generated by us using FLUX[6]. We use the simple text prompt "A portrait of a man/woman", as it consistently produces high-quality images without compromising output diversity. For the diffusion process, we set the number of denoising steps to 50 and use a guidance scale of 3.5. Additionally, we supplement the generated portraits with real identities from well-established face datasets, specifically FFHQ [36] and CelebAMask-HQ [43],

[5] `https://www.easy-languages.org/`

[6] `https://github.com/black-forest-labs/flux`

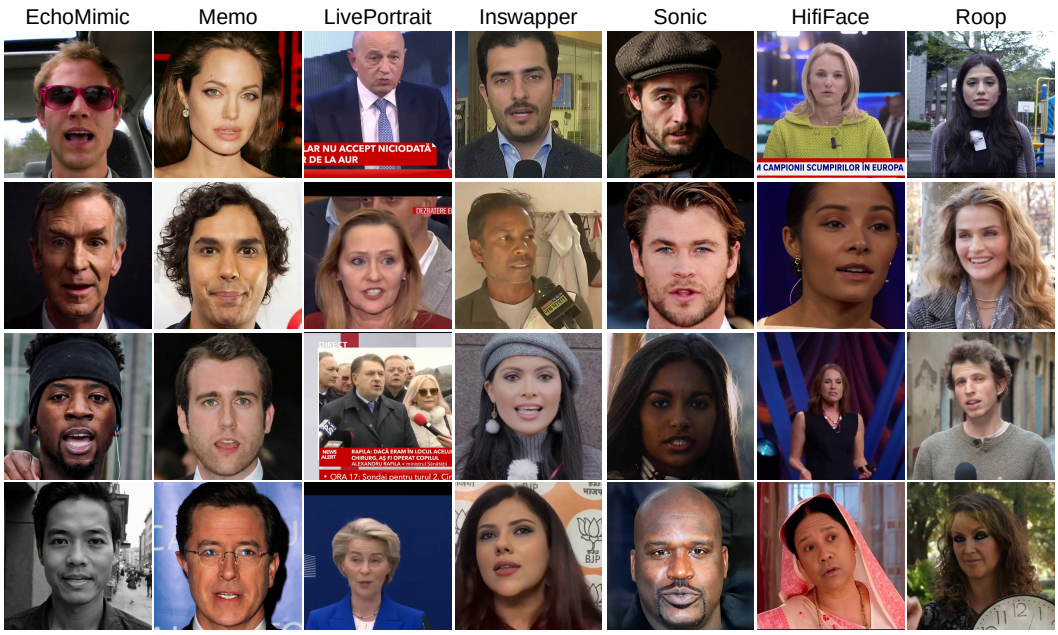


Figure 3: Fake video frames generated by each of the seven deepfake methods. Best viewed in color.

along with identities found in our real videos. These datasets have disproportional dimensions, but we sample subsets from each to ensure an almost uniform distribution across datasets.

The talking-head generation is performed with EchoMimic, Memo and Sonic. We provide these models with a portrait image, sampled from the previously described set, and an audio signal containing a person speaking. The audio also originates from the real video set described earlier. The result is a video in which the person from the portrait image utters the speech from the audio file. We emphasize that the models not only manage lip synchronization, but also effectively generate head movements and facial expressions required for this task. Furthermore, we observe that Memo and Sonic perform consistently well across multiple languages, while EchoMimic struggles with languages other than English and Mandarin. For this reason, we individually fine-tune EchoMimic on additional languages, such as Romanian and Arabic, before using it for generation. We use 1,000 real videos for each language and trained the model for 10 epochs. Finally, we synthesize over 10,000 videos using talking-head generation methods, resulting in more than 65 hours of fake content. All videos are generated at a consistent resolution of 512×512 pixels.

For facial expression manipulation, we employ LivePortrait [25]. This model can transfer facial movements (eyes, lips, and expressions) from a driving video to a source image or video. However, we observe a noticeable drop in quality when the person in the driving video is not directly facing the camera. Additionally, while lip synchronization is handled effectively, the transfer of eye movements and facial expressions is less effective. To address these limitations, we restrict our use to front-facing driving videos and focus only on lip synchronization. As a result, only the movements of the lips are synthesized in the generated samples, while all other facial attributes in the source video remain unchanged. The audio of the resulting video is taken from the driving video, to ensure alignment between the lips and the information spoken in the audio. We select front-facing driving videos from the set generated using talking-head synthesis, as these are primarily created from portrait images, and verified for the front-facing property. The source videos are represented by the real videos collected from YouTube. We generate over 2,900 videos using this method, resulting in more than 14 hours of fake content. The generated videos inherit the resolution of the source (real) videos, as the only changed aspect is the movement of the lips.

The face swapping is performed with Inswapper, HifiFace and Roop. Face swapping works by pasting the identity from a source image to a target video, while keeping the attributes that are not specific to the identity (facial expression, lip movement) unchanged. For the source images, we use portraits from the previously described dataset, which includes both synthetic and real identities. The target videos are selected from the collected set of real YouTube videos. Following face swapping, we apply GFPGAN [67] for face restoration to enhance visual quality. We generate over 22,000 videos

Table 3: Results obtained by pre-trained and fine-tuned versions of AVFF, MRDF and TALL on the MAVOS-DD official test sets: in-domain, open-set model, open-set language and open-set full. The best and second-best results on each column are highlighted in **bold blue** and orange, respectively. According to McNemar's statistical testing, all fine-tuned models are significantly better than their pre-trained counterparts (p-value < 0.001).

Method	Fine-tuned	In-domain			Open-set model			Open-set language			Open-set full		
		mAP	AUC	acc	mAP	AUC	acc	mAP	AUC	acc	mAP	AUC	acc
AVFF [52]	✗	0.51	0.51	52.45	0.50	0.50	22.58	0.51	0.51	59.46	0.50	0.50	35.34
MRDF [80]	✗	0.50	0.46	44.04	0.52	0.52	58.04	0.46	0.41	39.35	0.51	0.49	50.78
TALL [71]	✗	0.49	0.48	50.74	0.50	0.51	39.22	0.48	0.47	50.78	0.50	0.49	44.63
AVFF [52]	✓	**0.95**	**0.95**	**86.93**	**0.85**	**0.89**	75.34	**0.90**	**0.90**	**84.26**	**0.87**	**0.89**	77.68
MRDF [80]	✓	0.90	0.90	84.27	0.78	0.88	**78.32**	0.88	0.88	82.15	0.82	0.86	**78.87**
TALL [71]	✓	0.87	0.86	78.07	0.79	0.84	66.20	0.80	0.80	73.25	0.77	0.79	67.42

using this deepfake method, totaling 81 hours of fake content. The resolution of the resulting videos matches that of the target (real) videos.

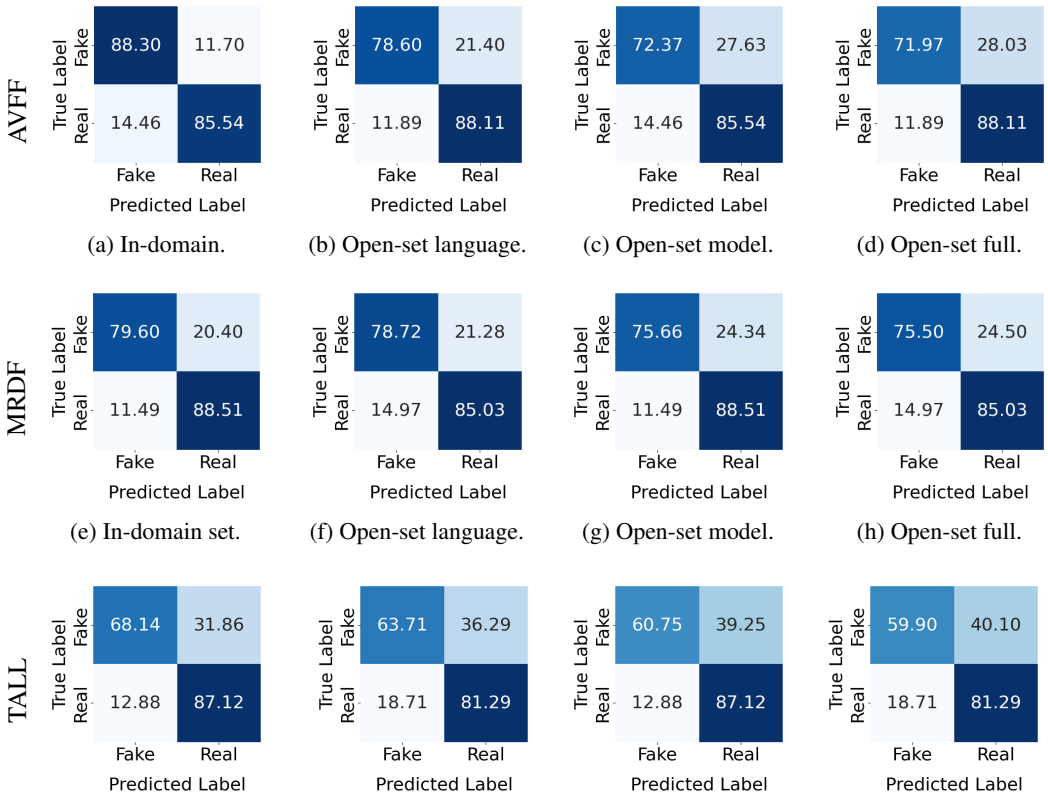
In Figure 3, we present synthetic video frames produced by each of the seven deepfake methods. The samples are diverse and have a high degree of realism, confirming that MAVOS-DD represents a challenging dataset for existing deepfake detectors. For both real and generated videos, we highlight that the number of frames per second (FPS) ranges from 23 to 60. The audio bitrate varies between 88 and 140 `kbps`, with the audio sample rate spanning from 16 to 44.1 `kHz`. The video bitrate ranges from 40 to over $10,000$ `kbps`.

4 Experiments

Baselines and hyperparameters. We conduct experiments using thee state-of-the-art deepfake detectors. Two of them, namely AVFF [52] and MRDF [80], are multimodal, while the third one, TALL [71], analyzes only the video input. AVFF employs two unimodal encoders based on transformer blocks, each of them being trained to predict features of the opposite modality. The outputs from both encoders are concatenated and passed to a binary classifier for deepfake detection. Similarly, MRDF uses two encoders to extract features from each modality. The two encoders are based on ResNet-18 [27]. Their output is concatenated and further processed by an audio-visual transformer module for deepfake detection. TALL is a spatio-temporal modeling method that captures both spatial and temporal inconsistencies. The method is applicable to multiple architectures. In our work, we use TALL-Swin, which is based on Swin Transformer [47]. We conduct the experiments using both pre-trained and fine-tuned versions of each model. We fine-tune MRDF for 5 epochs, TALL for 15 epochs and AVFF for 10 epochs on MAVOS-DD. The number of epochs are established based on early stopping. To optimize the models, we employ Adam [39] with a learning rate of 10^{-3} for MRDF, $2 \cdot 10^{-5}$ for TALL and 10^{-5} for AVFF, respectively. We keep the default values for the other hyperparameters of Adam. We set the batch size to 4 for AVFF and MRDF, and 32 for TALL. All the experiments are carried out on a computer with an Intel i9-14900K CPU with 192 GB of RAM and an Nvidia RTX 4090 GPU with 24 GB of VRAM.

Results. In Table 3, we report the results for the three baseline models across three evaluation metrics: mean average precision (mAP), area under the ROC curve (AUC), and accuracy (acc). We report these values on all four test sets: in-domain, open-set model, open-set language and open-set full.

The results demonstrate that MAVOS-DD is a difficult data set for existing deepfake detection methods, since all the employed and publicly available pre-trained models perform close to random chance, regardless of the test set. We can attribute the performance gap of pre-trained models to the fact that MAVOS-DD typically contains examples that are more challenging to detect, since they are generated with models that exhibit a high degree of realism. The fine-tuned versions perform much better, especially in the in-domain scenario. With respect to the in-domain scenario, their performance levels decline in open-set setups, indicating that further developments are needed to improve the generalization of state-of-the-art detectors. As expected, the most significant performance drop is



(a) In-domain. (b) Open-set language. (c) Open-set model. (d) Open-set full.

(e) In-domain set. (f) Open-set language. (g) Open-set model. (h) Open-set full.

(i) In-domain. (j) Open-set language. (k) Open-set model. (l) Open-set full.

Figure 4: Confusion matrices obtained by AVFF, MRDF and TALL after fine-tuning them on MAVOS-DD.

observed in the open-set model setup. This drop indicates that detectors still fail to generalize from a set of deepfake methods to another. The performance drop is lower in the open-set language case. However, when we examine the number of real samples incorrectly predicted by the fine-tuned MRDF model as fake across in-domain and open-set language scenarios, we observe a difference of 1,378 samples, increasing from 596 to 1,974. This suggests that a significant portion of misclassified samples are likely labeled as fake simply because the audio is in a language not included in the training set. Another important observation is the noticeable performance gap between the unimodal TALL method and the two multimodal approaches (AVFF and MRDF), suggesting that jointly analyzing visual and audio modalities provides a significant advantage on MAVOS-DD.

We report the confusion matrices obtained by AVFF, MRDF and TALL, for each of the four test scenarios in Figure 4. In the open-set scenarios, AVFF shows a significant drop in its ability to detect fake videos. The same observation applies to MRDF, although the number of false positives with respect to the in-domain test case drops by less than 4.1%. TALL exhibits a poor ability to detect deepfakes, regardless of the target test set. These observations strengthen the claim that MAVOS-DD represents a challenging deepfake benchmark for modern deepfake detectors. Finally, to attest the usefulness of the provided training data, we compute McNemar's statistical test between pre-trained and fine-tuned versions of each model, obtaining a p-value lower than 0.001 in all cases.

Error analysis. We investigate which of the deepfake generative methods poses the greatest challenge for MRDF in terms of detection accuracy. We find that samples generated by LivePortrait and Roop are the most difficult, with 80% of the samples being labeled as real. Roop is one of the methods included in the test set only, and we believe that this explains the poor performance of MRDF in identifying samples generated by Roop. In contrast, LivePortrait is part of the in-domain set, but the poor performance of the detector on this method can be attributed to the fact that we only synchronize the lips, leaving everything else as in the original video. In Figure 5, we illustrate such a scenario where we show, side-by-side, frames from a real video and its corresponding fake video modified with LivePortrait. In the illustrated video, MRDF fails to detect the fake, misclassifying it as real.

Figure 5: A real video and its corresponding fake sample generated using LivePortrait. The MRDF detector incorrectly classifies the fake sample as real. Best viewed in color.

5 Broader Impact and Limitations

The advancements of deepfake generation models have significant implications for society, as it facilitates the widespread of misinformation. As synthetic media becomes increasingly realistic and accessible, the risk of misuse continues to grow. To fight against this, not only more competent models are required, but also varied datasets, as robust detection systems heavily depend on the utilized training data. Our research fosters the development of such models, as it addresses some of the limitations of previous datasets: a wide range of generation methods, multiple languages, and a meticulously designed split that translates into challenging open-set evaluation scenarios. Robust deepfake detection models may be beneficial for journalists, social media platforms and even governmental agencies. It could also help to protect individuals from having their reputation damaged.

Nevertheless, we also acknowledge that the development of detection methods can also lead to more sophisticated generative models, the research in the generative AI domain being restless. Still, we are convinced that MAVOS-DD will continue to be very useful, as we aim to continuously update it with state-of-the-art generative models.

A potential limitation of our benchmark consists of the hardware requirements to carry out experiments on it. Some minimum resources, e.g. CPU for loading the videos and GPU for deep learning models, must be utilized for training and evaluating on such a dataset. Another possible limitation is represented by the fact that the dataset inadvertently has a demographic bias, corresponding to the set of eight languages, which could result in reduced performance between different populations. This requires a continued evaluation of fairness and increased responsibility when deploying deepfake models trained on our dataset.

6 Conclusion and Future Work

In this work, we introduced MAVOS-DD, a large-scale open-set benchmark for multilingual audio-video deepfake detection, comprising over 250 hours of real and generated videos. We further proposed a test split that creates four different evaluation scenarios: in-domain, open-set model, open-set language and open-set full. The resulting scenarios are aimed to assess the performance and robustness of deepfake detectors in challenging situations. We evaluated three different state-of-the-art deepfake detectors on the newly proposed benchmark, and observed significant performance drops across all four evaluation setups. The empirical results highlight the need to develop more robust deepfake detectors for practical scenarios.

In future work, we aim to continuously update the dataset by adding deepfake samples generated with models that are going to be released after our first release date. Thus, MAVOS-DD will keep up with the development pace of generative models, so that it will stay relevant for a long period of time. Additionally, we target the development of novel deepfake detectors that specifically address the challenges of the proposed open-set setups, which closely resemble real-world scenarios.

References

[1] Shruti Agarwal, Hany Farid, Ohad Fried, and Maneesh Agrawala. Detecting deep-fake videos from phoneme-viseme mismatches. In *Proceedings of CVPR*, pages 660–661, 2020.

[2] Luigi Attorresi, Davide Salvi, Clara Borrelli, Paolo Bestagini, and Stefano Tubaro. Combining automatic speaker verification and prosody analysis for synthetic speech detection. In *Proceedings of ICPR*, pages 247–263, 2023.

[3] Zhongjie Ba, Qingyu Liu, Zhenguang Liu, Shuang Wu, Feng Lin, Li Lu, and Kui Ren. Exposing the deception: Uncovering more forgery clues for deepfake detection. In *Proceedings of AAAI*, pages 719–728, 2024.

[4] Sarah Barrington, Matyas Bohacek, and Hany Farid. DeepSpeak Dataset v1.0. *arXiv preprint arXiv:2408.05366*, 2024.

[5] Logan Blue, Kevin Warren, Hadi Abdullah, Cassidy Gibson, Luis Vargas, Jessica O'Dell, Kevin Butler, and Patrick Traynor. Who Are You (I Really Wanna Know)? Detecting Audio DeepFakes Through Vocal Tract Reconstruction. In *Proceedings of USENIX*, pages 2691–2708, 2022.

[6] Nicolo Bonettini, Edoardo Daniele Cannas, Sara Mandelli, Luca Bondi, Paolo Bestagini, and Stefano Tubaro. Video face manipulation detection through ensemble of CNNs. In *Proceedings of ICPR*, pages 5012–5019, 2021.

[7] Nuria Alina Chandra, Ryan Murtfeldt, Lin Qiu, Arnab Karmakar, Hannah Lee, Emmanuel Tanumihardja, Kevin Farhat, Ben Caffee, Sejin Paik, Changyeon Lee, Jongwook Choi, Aerin Kim, and Oren Etzioni. Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024. *arXiv preprint arXiv:2503.02857*, 2025.

[8] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. SimSwap: An Efficient Framework For High Fidelity Face Swapping. In *Proceedings of ACMMM*, pages 2003–2011, 2020.

[9] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. EchoMimic: Lifelike Audio-Driven Portrait Animations through Editable Landmark Conditions. In *Proceedings of AAAI*, pages 2403–2410, 2024.

[10] Zhongxi Chen, Ke Sun, Ziyin Zhou, Xianming Lin, Xiaoshuai Sun, Liujuan Cao, and Rongrong Ji. DiffusionFace: Towards a Comprehensive Dataset for Diffusion-Based Face Forgery Analysis. *arXiv preprint arXiv:2403.18471*, 2024.

[11] Jongwook Choi, Taehoon Kim, Yonghyun Jeong, Seungryul Baek, and Jongwon Choi. Exploiting Style Latent Flows for Generalizing Deepfake Video Detection. In *Proceedings of CVPR*, pages 1133–1143, 2024.

[12] Andrea Ciamarra, Roberto Caldelli, Federico Becattini, Lorenzo Seidenari, and Alberto Del Bimbo. Deepfake Detection by Exploiting Surface Anomalies: The Surfake Approach. In *Proceedings of WACV*, pages 1024–1033, 2024.

[13] Davide Alessandro Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. Combining EfficientNet and Vision Transformers for Video Deepfake Detection. In *Proceedings of ICIAP*, pages 219–229, 2022.

[14] Davide Cozzolino, Alessandro Pianese, Matthias Nießner, and Luisa Verdoliva. Audio-visual person-of-interest deepfake detection. In *Proceedings of CVPR*, pages 943–952, 2023.

[15] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion Models in Vision: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023.

[16] Florinel-Alin Croitoru, Andrei-Iulian Hiji, Vlad Hondru, Nicolae Catalin Ristea, Paul Irofti, Marius Popescu, Cristian Rusu, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Deepfake Media Generation and Detection in the Generative AI Era: A Survey and Outlook. *arXiv preprint arXiv:2411.19537*, 2024.

[17] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *Proceedings of CVPR*, pages 5781–5790, 2020.

[18] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The DeepFake Detection Challenge (DFDC) Dataset. *arXiv preprint arXiv:2006.07397*, 2020.

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of ICLR*, 2021.

[20] Cunhang Fan, Jun Xue, Shunbo Dong, Mingming Ding, Jiangyan Yi, Jinpeng Li, and Zhao Lv. Subband fusion of complex spectrogram for fake speech detection. *Speech Communication*, 155:102988, 2023.

[21] Shuaijv Fang, Zhiyong Zhang, and Bin Song. Deepfake Detection Model Combining Texture Differences and Frequency Domain Information. *ACM Transactions on Privacy and Security*, 28(2):21, 2025.

[22] Yang Gao, Tyler Vuong, Mahsa Elyasi, Gaurav Bharaj, and Rita Singh. Generalized Spoofing Detection Inspired from Audio Generation Artifacts. In *Proceedings of INTERSPEECH*, pages 4184–4188, 2021.

[23] Candice R. Gerstner and Hany Farid. Detecting real-time deep-fake videos using active illumination. In *Proceedings of CVPR*, pages 53–60, 2022.

[24] Jiazhi Guan, Hang Zhou, Zhibin Hong, Errui Ding, Jingdong Wang, Chengbin Quan, and Youjian Zhao. Delving into sequential patches for deepfake detection. In *Proceedings of NeurIPS*, pages 4517–4530, 2022.

[25] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. LivePortrait: Efficient Portrait Animation with Stitching and Retargeting Control. *arXiv preprint arxiv:2407.03168*, 2024.

[26] Alexandros Haliassos, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Leveraging real talking faces via self-supervision for robust forgery detection. In *Proceedings of CVPR*, pages 14930–14942, 2022.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of CVPR*, pages 770–778, 2016.

[28] 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. In *Proceedings of CVPR*, pages 4360–4369, 2021.

[29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of NeurIPS*, volume 33, pages 6840–6851, 2020.

[30] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. HeadNeRF: A Real-time NeRF-based Parametric Head Model. In *Proceedings of CVPR*, pages 20374–20384, 2022.

[31] Hafsa Ilyas, Ali Javed, and Khalid Mahmood Malik. AVFakeNet: A unified end-to-end Dense Swin Transformer deep learning model for audio-visual deepfakes detection. *Applied Soft Computing*, page 110124, 2023.

[32] Xiaozhong Ji, Xiaobin Hu, Zhihong Xu, Junwei Zhu, Chuming Lin, Qingdong He, Jiangning Zhang, Donghao Luo, Yi Chen, Qin Lin, Qinglin Lu, and Chengjie Wang. Sonic: Shifting focus to global audio perception in portrait animation. In *Proceedings of CVPR*, 2025.

[33] 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 CVPR*, pages 2886–2895, 2020.

[34] Glenn Jocher, Jing Qiu, and Ayush Chaurasia. Ultralytics YOLO, January 2023. URL `https://github.com/ultralytics/ultralytics`.

[35] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar Fine-Tuning for 3D Human Model Fitting Towards In-the-Wild 3D Human Pose Estimation. In *Proceedings of IC3DV*, pages 42–52, 2021.

[36] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12): 4217–4228, 2021.

[37] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. In *Proceedings of NeurIPS*, 2021.

[38] Marouane Kihal and Lamia Hamza. Robust multimedia spam filtering based on visual, textual, and audio deep features and random forest. *Multimedia Tools and Applications*, 82(26):40819–40837, 2023.

[39] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of ICLR*, 2015.

[40] Staffy Kingra, Naveen Aggarwal, and Nirmal Kaur. LBPNet: Exploiting texture descriptor for deepfake detection. *Forensic Science International: Digital Investigation*, 42–43:301452, 2022.

[41] Pavel Korshunov and Sébastien Marcel. DeepFakes: a New Threat to Face Recognition? Assessment and Detection. *arXiv preprint arXiv:1812.08685*, 2018.

[42] Romeo Lanzino, Federico Fontana, Anxhelo Diko, Marco Raoul Marini, and Luigi Cinque. Faster than lies: Real-time deepfake detection using binary neural networks. In *Proceedings of CVPR*, pages 3771–3780, 2024.

[43] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of CVPR*, pages 5548–5557, 2020.

[44] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Advancing High Fidelity Identity Swapping for Forgery Detection. In *Proceedings of CVPR*, pages 5073–5082, 2020.

[45] Yuezun Li, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. In *Proceedings of CVPR*, pages 3204–3213, 2020.

[46] 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.

[47] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proceedings of ICCV*, pages 9992–10002, 2021.

[48] Juan Manuel Martín-Doñas and Aitor Álvarez. The Vicomtech Partial Deepfake Detection and Location System for the 2023 ADD Challenge. In *Proceedings of IJCAI*, pages 37–42, 2023.

[49] Alexander Quinn 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. In *Proceedings of ICML*, pages 16784–16804, 2022.

[50] Fan Nie, Jiangqun Ni, Jian Zhang, Bin Zhang, and Weizhe Zhang. Frade: Forgery-aware audio-distilled multimodal learning for deepfake detection. In *Proceedings of ACMMM*, page 6297–6306, 2024.

[51] Yuval Nirkin, Yosi Keller, and Tal Hassner. FSGAN: Subject Agnostic Face Swapping and Reenactment. In *Proceedings of ICCV*, pages 7184–7193, 2019.

[52] Trevine Oorloff, Surya Koppisetti, Nicolò Bonettini, Divyaraj Solanki, Ben Colman, Yaser Yacoob, Ali Shahriyari, and Gaurav Bharaj. AVFF: Audio-Visual Feature Fusion for Video Deepfake Detection. In *Proceedings of CVPR*, pages 27102–27112, 2024.

[53] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125*, 2022.

[54] Muhammad Anas Raza and Khalid Mahmood Malik. Multimodaltrace: Deepfake Detection Using Audiovisual Representation Learning. In *Proceedings of CVPR*, pages 993–1000, 2023.

[55] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of CVPR*, pages 10684–10695, 2022.

[56] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to Detect Manipulated Facial Images. In *Proceedings of ICCV*, pages 1–11, 2019.

[57] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Proceedings of NeurIPS*, pages 36479–36494, 2022.

[58] Davide Salvi, Honggu Liu, Sara Mandelli, Paolo Bestagini, Wenbo Zhou, Weiming Zhang, and Stefano Tubaro. A Robust Approach to Multimodal Deepfake Detection. *Journal of Imaging*, 9 (6), 2023.

[59] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Proceedings of NeurIPS*, pages 11918–11930, 2019.

[60] Kaavya Sriskandaraja, Vidhyasaharan Sethu, Phu Ngoc Le, and Eliathamby Ambikairajah. Investigation of Sub-Band Discriminative Information Between Spoofed and Genuine Speech. In *Proceedings of INTERSPEECH*, pages 1710–1714, 2016.

[61] Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is Someone Speaking? Exploring Long-term Temporal Features for Audio-visual Active Speaker Detection. In *Proceedings of ACMMM*, pages 3927–3935, 2021.

[62] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. EMO: Emote Portrait Alive Generating Expressive Portrait Videos with Audio2Video Diffusion Model Under Weak Conditions. In *Proceedings of ECCV*, pages 244–260, 2024.

[63] Chenglong Wang, Jiangyan Yi, Jianhua Tao, Chu Yuan Zhang, Shuai Zhang, and Xun Chen. Detection of Cross-Dataset Fake Audio Based on Prosodic and Pronunciation Features. In *Proceedings of INTERSPEECH*, pages 3844–3848, 2023.

[64] Haodi Wang, Xiaojun Jia, and Xiaochun Cao. EAT-Face: Emotion-Controllable Audio-Driven Talking Face Generation via Diffusion Model. In *Proceedings of FG*, pages 1–10, 2024.

[65] Tianyi Wang and Kam Pui Chow. Noise Based Deepfake Detection via Multi-Head Relative-Interaction. In *Proceedings of AAAI*, pages 14548–14556, 2023.

[66] 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. *Computer Speech & Language*, 64:101114, 2020.

[67] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards Real-World Blind Face Restoration with Generative Facial Prior. In *Proceedings of CVPR*, pages 9164–9174, 2021.

[68] Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji. HifiFace: 3D Shape and Semantic Prior Guided High Fidelity Face Swapping. In *Proceedings of IJCAI*, pages 1136–1142, 2021.

[69] Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024.

[70] Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. Vasa-1: Lifelike audio-driven talking faces generated in real time. In *Proceedings of NeurIPS*, pages 660–684, 2024.

[71] Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. TALL: Thumbnail Layout for Deepfake Video Detection. In *Proceedings of ICCV*, pages 22601–22611, 2023.

[72] Jun Xue, Cunhang Fan, Zhao Lv, Jianhua Tao, Jiangyan Yi, Chengshi Zheng, Zhengqi Wen, Minmin Yuan, and Shegang Shao. Audio Deepfake Detection Based on a Combination of F0 Information and Real Plus Imaginary Spectrogram Features. In *Proceedings of DDAM*, pages 19–26, 2022.

[73] Jichen Yang, Rohan Kumar Das, and Haizhou Li. Significance of Subband Features for Synthetic Speech Detection. *IEEE Transactions on Information Forensics and Security*, 15:2160–2170, 2020.

[74] Yibo Zhang, Weiguo Lin, and Junfeng Xu. Joint Audio-Visual Attention with Contrastive Learning for More General Deepfake Detection. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(5):137, 2024.

[75] Longtao Zheng, Yifan Zhang, Hanzhong Guo, Jiachun Pan, Zhenxiong Tan, Jiahao Lu, Chuanxin Tang, Bo An, and Shuicheng Yan. MEMO: Memory-Guided Diffusion for Expressive Talking Video Generation. *arXiv preprint arXiv:2412.04448*, 2024.

[76] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of ICCV*, pages 15024–15034, 2021.

[77] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, and Otmar Hilliges. IMavatar: Implicit Morphable Head Avatars from Videos. In *Proceedings of CVPR*, pages 13545–13555, 2022.

[78] Yipin Zhou and Ser-Nam Lim. Joint Audio-Visual Deepfake Detection. In *Proceedings of ICCV*, pages 14800–14809, 2021.

[79] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection. In *Proceedings of ACMMM*, pages 2382–2390, 2020.

[80] Heqing Zou, Meng Shen, Yuchen Hu, Chen Chen, Eng Siong Chng, and Deepu Rajan. Cross-modality and within-modality regularization for audio-visual deepfake detection. In *Proceedings of ICASSP*, pages 4900–4904, 2024.

A Ethical Statement

We share MAVOS-DD under the International Attribution Non-Commercial Share-Alike 4.0 (CC BY-NC-SA 4.0) license, aiming for open and responsible research on deepfake detection. All real data samples are collected from public YouTube videos. Since the videos are gathered from a public website, we adhere to the European regulations[7] allowing researchers to use and store data from the public web domain for non-commercial research purposes. Moreover, we respect the individual privacy rights, including the right to be forgotten. If any individual identifies themselves in the dataset and wishes to have their data removed, they can contact us and we will promptly address the request by removing the respective video(s), in compliance with data protection principles.

[7] `https://eur-lex.europa.eu/eli/dir/2019/790/oj`

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper introduces a new benchmark for deepfake detection, which was made available through the Huggingface platform. To attest the scale and benefits of our benchmark, we have included a comparison with other datasets in Table 1. Furthermore, we demonstrated the experiments carried out through the presented results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Although we have tried to address the limitations of the previous datasets, our benchmark presents a few limitations as well. These are all discussed in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA] .

Justification: Given the scope of our paper, we do not have any theoretical results to present, only experimental results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Firstly, we only used open-source code repositories available on GitHub. Secondly, the models, the training hyperparameters and the environments in which the experiments were carried out are clearly detailed in the first paragraph of Section 4, so that every technical person could easily reproduce our results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The introduced dataset has been made publicly available on one of the most well-known and used platforms, namely Huggingface, and it can be easily accessed with the `datasets` python package.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: On the one hand, one of the main strengths of our dataset is data organization: MAVOS-DD not only has the classic train/val/test splits, but also includes testing subsets for unseen generation models and languages. On other hand, the training and test details (including all the aforementioned elements) are clearly specified in the **Baselines and hyperparameters** paragraph of Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We have computed McNemar's statistical test, resulting in a p-value lower than 0.001. This is also mentioned at the end of the **Results** paragraph in Section 4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We described the compute resources of our working environment in the first paragraph of Section 3. Furthermore, we mentioned that we used the same environment for fine-tuning the detectors, as mentioned in Section 4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We confirm that we carefully read the NeurIPS Code of Ethics and we respect all the points. The most important concerns are related to Privacy, Copyright and Fair Use, nevertheless, our data originates from publicly available data which we ensured that it can be used for academic purposes. Additionally, the accompanying README file of the data set includes a note with instructions for individuals who wish to request the removal of content involving them, directing them to contact us via email.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our work aims to facilitate the deepfake detection research, and as a result, contribute to the development robust models and thus fight against the misuse. These positive societal impacts were discussed in Section 5.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Given that the objective of our proposed dataset is to detect deepfake media, and thus prevent the misuse, we consider that there is no need to implement such safeguards.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: On the one hand, we have made all the efforts to ensure we do not break any license for distributing the acquired data only for research purposes. On the other hand, we have the necessary rights for the generated samples and release them within the scope of our license.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The new asset i.e. the benchmark introduced is thoroughly documented in Section 3. Moreover, we provide a detailed documentation on how to access the dataset on its Huggingface page, such as how to filter the data for each split.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: For our dataset, we did not carry out any activities involving human subjects, but rather collected the data from online public sources.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We did not work with any human subjects, and thus we do not require any Institutional Review Board (IRB) approvals.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: Neither our work nor any components rely on LLMs.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.