Fit for Purpose? Deepfake Detection in the Real World

Guangyu Lin[1†], Li Lin[1†], Christina P. Walker[2], Daniel S. Schiff[2], Shu Hu[1*]

[1*]School of Applied and Creative Computing, Purdue University, West Lafayette, 47907, IN, USA.
[2]Department of Political Science, Purdue University, West Lafayette, 47907, IN, USA.

*Corresponding author(s). E-mail(s): hu968@purdue.edu;
Contributing authors: lin2285@purdue.edu; lin1785@purdue.edu;
walke667@purdue.edu; dschiff@purdue.edu;
†These authors contributed equally to this work.

Abstract

The rapid proliferation of AI-generated content, driven by advances in generative adversarial networks, diffusion models, and multimodal large language models, has made the creation and dissemination of synthetic media effortless, heightening the risks of misinformation, particularly political deepfakes that distort truth and undermine trust in political institutions. In turn, governments, research institutions, and industry have strongly promoted deepfake detection initiatives as solutions. Yet, most existing models are trained and validated on synthetic, laboratory-controlled datasets, limiting their generalizability to the kinds of real-world political deepfakes circulating on social platforms that affect the public. In this work, we introduce the first systematic benchmark based on the Political Deepfakes Incident Database—a curated collection of real-world political deepfakes shared on social media since 2018. Our study includes a systematic evaluation of state-of-the-art deepfake detectors across academia, government, and industry. We find that the detectors from academia and government perform relatively poorly. While paid detection tools achieve relatively higher performance than free-access models, all evaluated detectors struggle to generalize effectively to authentic political deepfakes, and are vulnerable to simple manipulations, especially in the video domain. Results urge the need for politically contextualized deepfake detection frameworks to better safeguard the public in real-world settings.

Keywords: Deepfake, Political Science, Benchmark, Media Forensics

1 Introduction

AI-generated content (AIGC) has become increasingly prevalent in daily life, reshaping online communication [1–3]. Advances in generative adversarial networks (GANs), diffusion models (DMs), and multimodal large language models (LLMs) have lowered the barriers to producing synthetic media, including fake images and video, which are now widespread across platforms like Twitter/X, Facebook, Instagram, and Reddit [4]. However, the rapid proliferation of AIGC has introduced considerable risks [5–8]. Synthetic content, particularly *deepfakes*,[1] can blur the boundaries between authentic and fabricated information, erode trust in media and democratic institutions, fuel polarization, and enable reputational attacks or false claims of manipulation [9–12]. Among the most concerning applications of deepfakes are their use in **politics** [13–15]. High-profile political deepfakes have appeared globally: from a spoof video of former U.S. President Barack Obama, distorted footage of Indian Prime Minister Narendra Modi, to fake election ads in countries from Australia to Slovakia [16–19]. Since the popularization of diffusion and LLMs in 2022, the volume and accessibility of this type of content have surged, prompting calls for regulation, watermarking, and deepfake detection [3, 20–29].

In response, there has been a concerted effort to develop deepfake detection models [30–32]. The U.S. NIST Center for AI Standards and Innovation (formerly the AI Safety Institute), for instance, has emphasized the urgent need to detect and mitigate synthetic media, a priority echoed in several U.S. federal executive orders, the America's AI Action Plan, and initiatives from multiple international AI safety institutes, regulatory bodies, legislative frameworks, and global competitions. [22, 33–37]. However, existing models are typically trained and evaluated on synthetic datasets, such as FaceForensics++ [38], Celeb-DF [39], and DFDC [36], which are **created under controlled laboratory conditions**. While effective in their domains, these detection tools and benchmarks may fail to reflect the complexity and diversity of synthetic media encountered on social media platforms, which tend to defy neat categorization: varying in AI techniques used to create them, resolution, content structure, composition, and salience.

For instance, many detection tools are focused squarely on detecting deepfakes rather than simpler manipulations, and emphasize detection of inauthentic renderings of humans and human faces in particular [40, 41]. However, in real-world settings where individuals are most often exposed to potential misinformation, like social media, political deepfakes often contain multiple people, altered context such as background settings, mixed synthetic and real elements, and a range of image styles—cartoonish memes of a politician, burning of a flag, or an overhead view of a protest. Further,

[1] Deepfake is a portmanteau of "deep learning" and "fake." It refers to AI-generated or digitally manipulated media (such as images or videos), produced using deep neural networks that appear highly realistic. Deepfakes often depict individuals engaging in actions or speaking words they never actually performed, thereby creating a deceptive yet convincing imitation of reality.

Fig. 1: *Political deepfake samples from different sources in the PDID.*

many instances of politically misleading images and videos do not meet the strict definition of a deepfake. For example, cheapfakes and shallowfakes–simpler forms of misinformation based on Photoshop-style tools, splicing, cropping, or other simple editing–can be equally persuasive and damaging, yet are often **overlooked** in detection efforts [42–44]. This oversight can lead to unintended but serious consequences. Social media platforms that adopt detection tools or watermarking and labeling schema like C2PA's Condent Credentials [45], for example, could risk falsely reassuring users that unmarked content is trustworthy, even when it is misleading or manipulated. This presents a challenge for the evaluation and deployment of detection tools: ***how generalizable and robust are current state-of-the-art deepfake detectors to the kinds of content that actually circulate in high-stakes political contexts?***

Recent efforts have begun to address these generalization concerns. For instance, Deepfake-Eval-2024 [46] collects real-world examples from the web, while MAVOS-DD [47] introduces synthetically generated multilingual, multimodal benchmarks. Yet, despite important contributions, these benchmarks have significant limitations: they lack systematic coverage of politically salient incidents, still rely primarily on synthetic data, and fail to evaluate robustness under common post-processing operations such as blurring and compression [48]. ***No existing benchmarking effort explicitly centers on real-world political deepfakes that are publicly and organically circulated.***

In this work, we perform the **first** systematic benchmarking exercise against our current developed **Political Deepfakes Incident Database** (PDID) [16], which curates politically salient deepfake incidents primarily from a United States context (including false allegations of deepfakes) sourced directly from real-world social media dissemination spanning from 2018 to the present (see samples in Fig. 1). Specifically, (1) we first select images and videos from the PDID database spanning 2018 to September 2025 and apply a human-in-the-loop, two-stage annotation process to assign the most accurate labels (fake or real) to each sample. (2) We then analyze the dataset's distribution in terms of source platform, resolution, publication year, and

duration (for videos), and compare it with existing lab-based deepfake videos, GAN-generated images, and DM-generated images through a frequency-domain analysis to highlight distinctive characteristics of real-world fakes. (3) Finally, we categorize existing detectors into two groups based on whether they incorporate or neglect large vision-language models (LVLMs): LVLM-aware detectors and LVLM-agnostic detectors. The LVLM-agnostic group is further divided into white-box detectors (from academia and government with accessible source code) and black-box detectors (from industry). We then systematically evaluate their performance on our curated PDID dataset.

We find that (1) the fake samples in the PDID display distinct spectral characteristics compared to lab-generated fakes, while the real samples in the PDID also differ substantially from those used in laboratory-generated datasets, revealing their **real-world complexity and variability** compared to synthetic datasets; (2) detectors from academia and government perform relatively poorly, achieving a maximum AUC of only 74.78% for images and 73.67% for videos, suggesting that such detectors are **not yet suitable for direct public use** and require additional fine-tuning, development, and contexutalization to handle political deepfake scenarios effectively; (3) **paid detectors generally outperform free-access counterparts**, potentially due to their continual updates, exposure to more diverse training data, and the integration of Retrieval-Augmented Generation modules (in LVLMs); (4) most LVLMs demonstrate stable performance and exhibit **minimal sensitivity** to common post-processing operations applied to the original media; and (5) **political video deepfake detection remains a persistent challenge**, with most detectors showing a marked drop in performance compared to image-level results. While careful threshold tuning can improve accuracy and reduce false acceptance rates, this approach demands expertise **beyond** that of average users. In general then, existing detectors **struggle to generalize** to authentic political deepfakes.

Implications of these findings are significant. They bear on the suitability of both open and commercial detection tools as a solution for political misinformation, revealing gaps in the development and evaluation of these tools, as well as problematic disconnects if tool capabilities and limitations are poorly understood, including by platforms with core responsibilities to protect the public interest. Overall, detection of particular classes of inauthentic content constitutes important work, but if tools are not developed, translated, and implemented in a way that is responsive to the real-world context of political misinformation, they risk both failing to sufficiently address the core problem as well as providing false assurances when the other mitigation strategies be needed.

2 Results

2.1 PDID Dataset Statistics and Analysis

To demonstrate the significance and uniqueness of the PDID, we compare it against several widely used deepfake datasets. (1) As shown in Table 1, the PDID is unique in its focus on politically salient deepfake 'incidents' across both image and video modalities. In addition to manipulated media, it also includes authentic, real samples

(e.g., images alleged to be deepfakes but confirmed to be authentic). (2) Unlike most existing datasets that either lack label explanations or rely on heuristically generated explanations (*e.g.*, using LVLMs), our dataset includes human-verified labels with evidence of external verification by media or fact-checkers and textual explanations for both fake and real samples. This is a unique feature that enables future development of media forensics-specific LVLMs with reasoning using supervised fine-tuning. Although this application is beyond the scope of this work, it presents a promising direction for future research. (3) Moreover, our dataset is constructed entirely from media disseminated in the wild—unlike most existing datasets, which rely on lab-generated content using known deepfake, GAN, or diffusion models. This makes our fake samples significantly more representative of the types of manipulations encountered in real-world scenarios, and more likely to have real-world impact. As such, the PDID dataset provides a more realistic and useful foundation for evaluating and designing deepfake detection methods in the political context (e.g., as opposed to the corporate fraud or intimate content context which may disseminate through different channels).

2.1.1 Dataset Distribution

Since the PDID database is continually updated, for this benchmark study, we extract only the images and videos posted between 2018 and September 2025 that have undergone external verification and were subsequently labeled by both professional annotators and domain experts. This results in a curated dataset of 232 images and 173 videos. Additional details regarding the database and the extraction and labeling strategy are provided in Section 4.1. To better understand the characteristics of the extracted samples, we compute distribution statistics for both images and videos, with the results summarized in Figure 2. Specifically,

1. **Image Data**. As shown in Figure 2 top, most images originate from X (formerly Twitter) (Fig. 2(a)), while smaller portions are collected from Facebook, TruthSocial, Medium, a researcher-hosted website (Farid) [56], which archived deepfakes related to the 2024 U.S. Presidential Election, and other miscellaneous

Dataset	Year	Image		Video		Political Deepfakes	Human Explanation	In-the-Wild
		Real	Fake	Real	Fake			
FaceForensics++ [38]	2019	✗	✗	✓	✓	✗	✗	✗
DeeperForensics-1.0 [49]	2020	✗	✗	✓	✓	✗	✗	✗
Celeb-DF [39]	2020	✗	✗	✓	✓	✗	✗	✗
DFDC [36]	2020	✗	✗	✓	✓	✗	✗	✗
GenData [50]	2023	✗	✓	✗	✗	✗	✗	✗
DF-Platter [51]	2023	✗	✗	✓	✓	✗	✗	✗
DF40 [52]	2024	✗	✓	✗	✗	✗	✗	✗
MMFakeBench [53]	2024	✗	✓	✗	✗	✓✗	✗	✗
Deepfake-Eval [54]	2025	✓	✓	✓	✓	✓✗	✗	✓
AI-Face [55]	2025	✓	✓	✗	✗	✗	✗	✗
PDID (ours)	2025	✓	✓	✓	✓	✓	✓	✓

Table 1: *Comparison of existing datasets with ours. Human explanation refers to the justification or rationale provided by a person to determine whether corresponding images or videos are real or fake.*

sources. Across all platforms, fake images constitute the majority of samples. In terms of resolution (Fig. 2(b)), most images are relatively low quality, with 720p (46.1%) and 480p (40.5%) being the most common. This may be due to post-processing applied by social media platforms, which often downscale uploaded images to improve loading speed and accessibility at the expense of pixel-level detail. Finally, the yearly distribution of fake images (Fig. 2(c)) shows that the majority were generated in 2023 and 2024, aligning with the 2024 U.S. Presidential Election, during which a surge of politically motivated deepfakes circulated on social media.

2. **Video Data**. As shown in Figure 2 bottom, the majority of videos (46.2%) are sourced from X, with additional samples collected from YouTube, TikTok, and Instagram (Fig. 2(d)). Similar to the fake image distribution, fake videos represent the majority within most platform groups. In terms of resolution (Fig. 2(e)), low-resolution videos (below 720p) dominate, reflecting the post-processing and compression typically applied by social media platforms regardless of media modality. The duration distribution (Fig. 2(f)) further reveals that most videos (74.6%) are shorter than one minute, with the majority of these being fake. This trend suggests that creators of deepfakes may prefer producing short videos, which are both more difficult for the public to scrutinize and technically easier to generate, given the current limitations of generative AI in producing long and highly realistic videos. Finally, Figure 2(g) shows the temporal distribution, which mirrors the trend observed in images: most deepfake videos were produced in

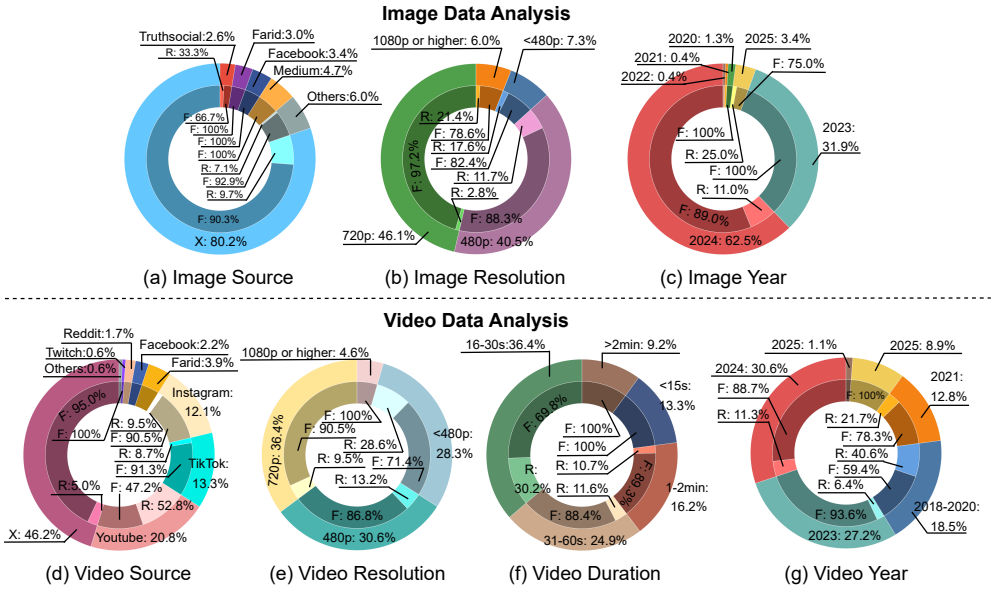

Fig. 2: *Dataset distribution across image and video sources, resolutions, durations, and years. The outer ring illustrates the overall proportion of each category, while the inner ring shows the breakdown of fake (F) and real (R) samples within each group.*

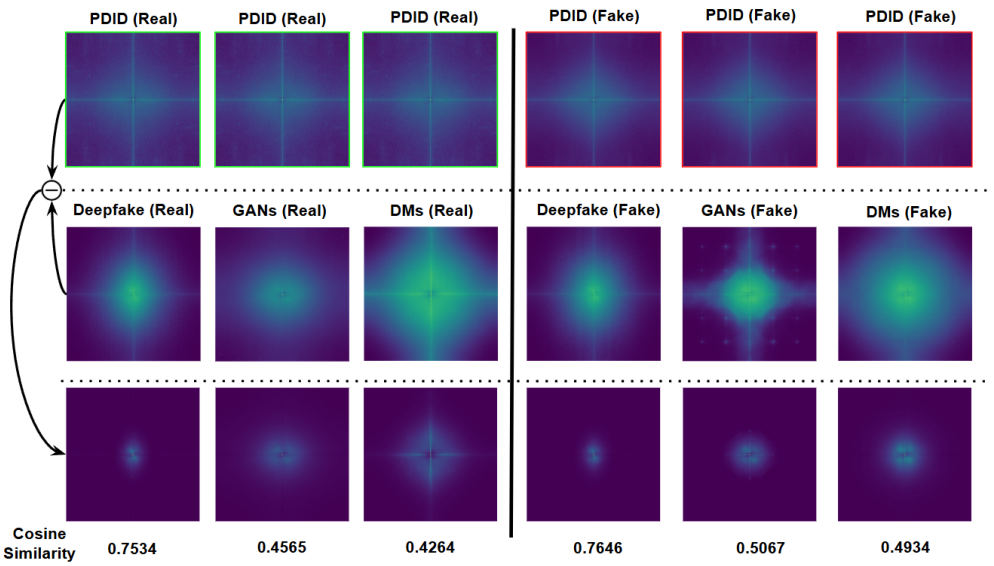

Fig. 3: *(Left three columns) Comparison of power spectra between the PDID real and others real. (Right three columns) Comparison of power spectra between the PDID fakes and other deepfakes. The bottom figures show the differences between the corresponding spectra in the top two rows, with the cosine similarity quantifying similarity.*

2023 and 2024. Notably, the proportion of deepfake videos increased from 18.5% in 2018–2020 to 30.6% in 2024. Although both 2020 and 2024 were U.S. election years, the sharp increase in 2024 can be attributed to the wider availability and accessibility of generative AI tools (*e.g.*, Sora 2 [57]), making deepfake creation significantly easier and more prevalent.

2.1.2 Power Spectrum Analysis

These diverse characteristics reflect content distributions encountered in real-world settings, which differ significantly from those found in lab-based synthetic datasets. To better understand the substantial differences between real-world fake images (from the PDID) and lab-generated fake images, we conduct a power spectrum analysis—a technique that reveals signal patterns in the frequency domain while minimizing the influence of semantic content. This analysis compares images from the PDID, conventional deepfake video datasets, GAN-generated datasets, and DM-generated datasets. The latter three categories are sourced from the AI-Face dataset [55]. Real images corresponding to each category for the fake image generation are also included for comparison. Additional details are provided in Section 4.1.3.

PDID (Real) vs Others (Real). Figure 3 (Left) compares the spectra of real images from the PDID with those from the other three categories. We conduct both qualitative and quantitative (*i.e.*, cosine similarity, where higher values indicate stronger alignment in spectral patterns) comparisons. Both visual analysis and cosine

similarity values reveal a clear distinction: the real samples in the PDID differ substantially from those in the Deepfake, GAN, and DM categories. This divergence is likely due to the constrained nature of the real data used in the latter categories:

1. In the Deepfake category, real videos are typically sourced from YouTube clips featuring trackable, mostly frontal faces with minimal occlusion, or recorded under controlled conditions with consented individuals. Some focus on celebrity interviews. These constraints limit diversity and reduce the authenticity typically observed in real-world social media content.
2. In the DM category, real images used to train diffusion models are selected from the IMDB-WIKI dataset [58], which primarily consists of celebrity portraits. Moreover, this dataset was released in 2015 and does not reflect the current visual characteristics or resolutions commonly found in modern social media images.
3. In the GAN category, real images used to train GANs come from the FFHQ dataset [59], which contains high-resolution (1024×1024), high-quality PNG images with diverse age, ethnicity, and background variations. However, this stands in contrast to the real images in the PDID, which are often heavily compressed and typically fall below 720p resolution, as shown in the resolution distributions (see Figure 2). These compression artifacts and resolution differences contribute significantly to the divergence in power spectral characteristics.

Interestingly, among the three categories, the Deepfake real images appear most similar to those in the PDID, as evidenced by its relatively higher cosine similarity score. This may be because many of the real videos in the PDID are also sourced from YouTube, aligning more closely with the data origin of the Deepfake category compared to the other two.

PDID (Fake) vs Others (Fake). Figure 3 (Right) compares the spectra of fake images in the PDID with those from the other three categories. Both visual analysis and cosine similarity scores reveal substantial differences, indicating that the PDID fake content exhibits distinct spectral characteristics. Several factors may explain this divergence:

1. Fake images and videos in the PDID are sourced from real-world social media platforms and may be produced by a wide range of generative AI tools. Many of them may not be represented in the training or generation pipelines of the other three categories. As such, the generative characteristics of the PDID fakes are more heterogeneous.[2]
2. Unlike lab-generated media, which typically undergo little to no post-processing, political fake media in the PDID from social media often experiences multiple stages of transformation before dissemination. These could include compression, editing, filtering, and platform-specific re-encoding, all of which introduce additional artifacts that alter the frequency-domain properties of the media.

[2]Notably, we witness this heterogeneity in the PDID dataset despite some areas of uniformity, as noted in the limitations section. For instance, PDID images and videos are 1) largely sourced from the US rather than a global context, 2) focus on a modest subset of contexts (*e.g.*, political events or figures rather than entertainment or sports), and 3) feature prominent individuals like political party leaders in a plurality of cases. It is striking that the PDID is potentially more structurally diverse than traditional laboratory datasets used for deepfake detection despite these built-in semantic constraints.

Benchmark	Year	Real-world Political Deepfakes	LVLM-Agnostic Detectors			LVLM-Aware Detectors
			White Box		Black Box	
			Academia	Government	Commercial	
Loc et al. [60]	2021	✗	✓	✗	✗	✗
DeepfakeBench [61]	2023	✗	✓	✗	✗	✗
CDDB [62]	2024	✗	✓	✗	✗	✗
Deng et al. [63]	2024	✗	✓	✗	✗	✗
DF40 [52]	2024	✗	✓	✗	✗	✗
MMFakeBench [53]	2024	✓✗	✗	✗	✗	✓
AI-Face [55]	2025	✗	✓	✗	✗	✗
Shield [64]	2025	✗	✗	✗	✗	✓
Ren et al. [65]	2025	✗	✗	✗	✗	✓
Ours	2025	✓	✓	✓	✓	✓

Table 2: *Comparison of existing deepfake detection benchmarks and ours.*

These observations demonstrate the unique challenges posed by real-world settings, as captured in the PDID dataset. Traditional detectors primarily developed and evaluated on lab-generated datasets may struggle to detect real-world political deepfakes. This motivates our systematic evaluation of existing detection models on the PDID dataset to expose current limitations and identify areas for improvement. These challenges and findings are further explored in the following sections.

2.2 Benchmark Evaluations

In this section, we evaluate existing, state-of-the-art deepfake detectors provided by academic researchers, government initiatives, and commercial providers. While several prior studies have conducted benchmarking, our work differs significantly in scope. Specifically, we focus on the detection of real-world political deepfakes. Moreover, existing benchmarks typically focus on a narrow range of (academic) models, whereas we expand coverage by including detectors from government agencies and industry providers, which have been largely overlooked in past efforts (see Table 2). By incorporating this broader set of detectors, our evaluation exposes both practically significant limitations that emerge when detectors are deployed in real-world settings and category-specific weaknesses that can inform directions for future improvement.

To better differentiate detector types, we organize them into two high-level categories based on whether large vision language models (LVLMs) are incorporated in their development: LVLM-agnostic detectors and LVLM-aware detectors. The LVLM-agnostic group is further divided into white-box detectors and black-box detectors (*i.e.*, commercial tools), depending on the availability of source code and implementation details. Within the white-box group, we additionally distinguish between detectors originating from academia and those specifically facilitated by government programs.

2.2.1 LVLM-Agnostic White-Box Detectors

We first evaluate twelve widely used academic deepfake detectors spanning four major categories (*i.e.*, naive backbones, frequency-based, spatial-based, and fairness-enhanced), all pretrained on the AI-Face dataset [55]. In addition, we assess three detectors developed under the DARPA Semantic Forensics (SemaFor) program [77].

Detector Type	Category	Detector Name	Image ACC↑	Image AUC↑	Image FAR↓	Video ACC↑	Video AUC↑	Video FAR↓
Academia	Naive	Xception [30]	84.57	64.30	8.67	80.63	60.09	3.03
		EfficientNet-B4 [66]	81.38	67.09	16.18	81.25	72.17	4.55
		ViT-B/16 [67]	87.77	66.36	5.78	81.25	67.38	**1.52**
	Frequency	F3Net [32]	**89.36**	71.54	**2.89**	80.00	57.41	5.30
		SPSL [68]	88.83	63.47	6.36	78.75	58.89	6.06
		SRM [69]	78.72	**74.78**	18.50	**82.50**	69.82	3.79
	Spatial	UCF [70]	86.70	72.00	6.36	78.12	60.13	8.33
		UnivFD [71]	52.66	48.82	46.24	**82.50**	69.81	4.55
		CORE [72]	52.66	57.96	46.24	80.00	66.52	8.33
	Fairness-enhanced	DAW-FDD [73]	64.36	66.15	35.26	**82.50**	68.51	3.79
		DAG-FDD [73]	85.64	52.60	6.94	81.87	54.73	**1.52**
		PG-FDD [31]	85.11	65.47	8.67	80.63	**73.67**	4.55
Government	-	CNNDetection [74]	22.87	42.52	81.50	16.27	65.95	97.20
		KitwareDetector [75]	8.51	57.23	99.42	13.86	53.50	100.00
		GANattribution [76]	22.87	29.75	82.08	67.47	50.55	25.87

Table 3: *Performance (%) comparison for LVLM-agnostic white-box detectors. The best results are shown in* ***Bold****.*

The performance of these models is summarized in Table 3. (1) Overall, all detectors exhibit low Accuracy (ACC) and Area Under the Receiver Operating Characteristic Curve (AUC) in both image and video settings. For instance, the highest AUC achieved is only 74.78% for images and 73.67% for videos. These results indicate that **detectors developed in academia and government are not sufficiently effective for detecting real-world political deepfakes disseminated on social media**. (2) When comparing detectors developed by academia and government, it is evident that academic detectors achieve substantially higher ACC and AUC scores than government-developed detectors. One possible explanation is that government detectors were trained on datasets composed of images and videos generated by a limited set of generative models under controlled conditions, typically restricted to a single generative architecture curated within the program. In contrast, the academic detectors evaluated here were trained on a more recent dataset that includes samples produced by a wide range of generative models. Additionally, since the DARPA SemaFor program concluded some time ago, the corresponding detectors may no longer receive updates or maintenance, further contributing to their relatively lower performance on contemporary, real-world political deepfakes. This finding urges that **continuous updates and adaptation of deepfake detection models are indispensable**, as detectors developed only a few years ago appear to be largely obsolete when confronted with evolving generative techniques. (3) Among the academic detectors, F3Net [32] and SPSL [69] achieve the strongest relative performance at the image level. A plausible explanation is that both methods operate in the frequency domain, making them **less sensitive to semantic content** (*e.g.*, particular individual politicians depicted in an image) and more focused on identifying *signal-level artifacts*. Consequently, the actual deepfake content contributes less to detection performance than underlying artifacts or traces produced by the generative model; these dominate the detector's decision process. Interestingly, all academic detectors demonstrate comparable ACC and favorable False Acceptance Rate (FAR) performance on the video

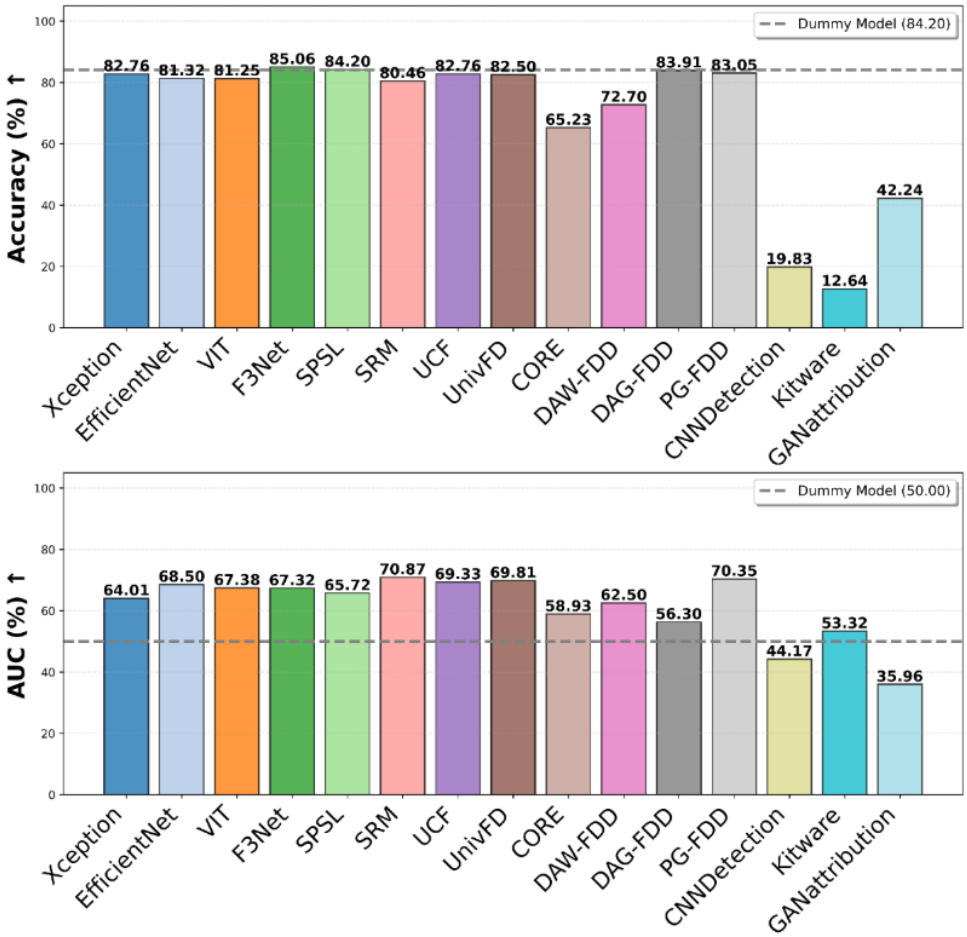

Fig. 4: *Performance comparison between LVLM-agnostic white-box detectors and "dummy model" on the image and video mixed set. The "dummy model" refers to a baseline that predicts all samples as fake.*

dataset. As detailed in Section 4.2, we randomly extract 10 frames from each video for testing and develop an aggregation approach to determine the final video prediction: temporal consistency or averaging effects across frames may help stabilize predictions and explain the relatively strong predictive performance here.

We next combine the image and video sets into a single dataset and evaluate all detectors on this unified collection. To better understand their performance, we introduce a dummy baseline model (always fake prediction) that predicts every sample as fake with 100% probability, and compare all detectors against it. The dummy baseline represents a simple yet revealing reference point: it predicts every sample as "fake." While seemingly trivial, it captures an important real-world phenomenon—if a large portion of the media circulating online or within certain domains (such as

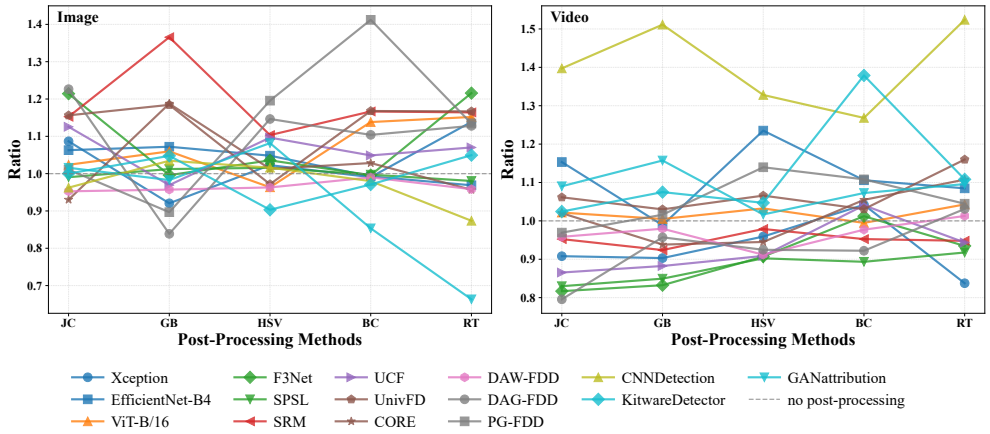

Fig. 5: *Performance (AUC) ratio (original vs. post-processed) for LVLM-agnostic white-box detectors.*

political misinformation) is indeed fake, then a naive system that always flags content as fake may appear to perform competitively under conventional metrics. This highlights a key limitation of current evaluation practices: detectors optimized for high sensitivity (flagging many fakes) can achieve deceptively strong results without true discernment. In practice, such imbalance could create a perverse incentive for developers to tune models toward over-detection, masking poor specificity, falsely assuring users, and undermining public trust in automated detection systems. The results are summarized in Figure 4. Only one detector (*i.e.*, F3Net [32]) surpasses the dummy baseline in terms of ACC. While most detectors achieve higher AUC scores than the dummy model, notable exceptions emerge among the government-developed detectors. Specifically, CNNDetection [74] and GANattribution [76], which tend to assign higher confidence scores to real samples than to fake ones, effectively predict in the wrong direction (worse than a random guesser).

However, **the low performance of these models does not necessarily imply that they are poorly designed; rather, it reflects their limited suitability for detecting real-world political deepfakes**. Many of these detectors were originally developed to target specific manipulation types, such as CNN- or GAN-generated media, under controlled conditions. Their underperformance is thus expected when faced with the unpredictable, mixed, and evolving manipulations found in social media. In contrast to the laboratory setting, users in the real world are not informed whether an image or video is GAN-generated, diffusion-based, or authentic, which exposes a critical mismatch between benchmark-specific model capabilities and practical utility. This discrepancy reiterates the importance of evaluating detectors from the perspective of ordinary users, who rely on them for trustworthy assessments without technical context. Compounding this issue, some commercial tools (as discussed

further in the next section) may overstate their effectiveness while downplaying their limitations[3].

We also conduct a robustness assessment of all detectors by applying five post-processing operations (*i.e.*, JPEG Compression (JC), Gaussian Blur (GB), HSV Shift (HSV), Brightness/Contrast Adjustment (BC), Rotation (RT)) as detailed in Section 4.2. The results are presented in Figure 5. For both image and video sets, we find that most performance change ratios (with post-processing versus without) fall within the range of [0.9, 1.1], indicating that **post-processing has minimal impact on the detectors' effectiveness**. This stability may be attributed to the fact that political deepfake samples in the PDID are sourced from social media, where they have already undergone various post-processing operations such as compression and blurring. Consequently, applying additional post-processing introduces little further change in detector performance. However, we indeed find a few detectors that are sensitive to specific post-processing approaches.

2.2.2 LVLM-Agnostic Black-Box Detectors

Recognizing that LVLM-agnostic white-box detectors from academia and government may have inherent limitations (such as training data constraints and lack of ongoing maintenance), our findings could fail to capture state-of-the-art practice in deepfake detection. To address this, we evaluate several of the most widely used commercial deepfake detection tools. Although these tools are not open-source, they are typically actively maintained and regularly updated, making them more representative of current real-world detection capabilities, particularly for ostensibly publicly-important (or commercializable) domains like deepfakes used for commercial fraud or political misinformation. To ensure a fair comparison, we conducted all tests within a short time frame (September 2025) to minimize the impact of version updates and promote consistency across evaluations.

Table 4 presents a comparative evaluation of LVLM-agnostic black-box detectors (*i.e.*, commercial tools) across both image and video modalities. The comparison includes widely used platforms such as *Incode*, *Reality Defender*, *Hive Moderation*, and others. These commercial detectors operate as closed-source platforms, each with distinct input requirements, rejection criteria, and output formats, which collectively affect their coverage, usability, and interpretability. For instance, *Incode* is a face-oriented deepfake detector primarily designed for identity verification and biometric authentication. It automatically rejects inputs when no face is detected or when the detected region is too small, resulting in partial sample exclusion. Similarly, *BrandWell* accepts only images in .jpg, .png, or .webp formats and discards all others, reducing its effective sample size for our evaluation. However, these constraints should not affect the overall conclusions of our evaluation, as we assess the tools from the perspective of average users—those without technical expertise who typically rely on such platforms and predictions at face value, and are not positioned to perform extensive testing, tweaking, or interpretation across multiple API endpoints, and so on.

[3]For example, one company advertises localization capabilities that remain non-functional even in upgraded plans, while several companies claim more than 98 or 99% accuracy, whereas four of the tools we test are below 65% accuracy.

Data Modality	Commercial Tools	ACC ↑	AUC ↑	FAR ↓	Output		Date
					Binary	Probability	
Image	Incode* [78]	91.07	52.70	**2.56**	✓	✓	9/13/2025
	Is It AI [79]	89.61	**96.15**	12.00	✓	✓	9/14/2025
	Winston [80]	61.64	85.73	43.50	✗	✓	9/14/2025
	BrandWell* [81]	59.02	39.34	31.58	✗	✓	9/14/2025
	AI or Not [82]	89.22	95.01	12.00	✓	✓	9/15/2025
	Illuminarty [83]	49.37	70.48	57.07	✗	✓	9/15/2025
	Reality Defender [84]	**93.51**	95.82	5.53	✓	✓	9/15/2025
	Hive Moderation [85]	83.62	94.47	18.50	✓	✓	9/16/2025
	Deepfake Detector [86]	49.57	68.59	58.50	✓	✓	9/16/2025
Video	Incode* [78]	**77.27**	44.74	**10.53**	✓	✓	9/24/2025
	AI or Not [82]	75.14	59.42	17.02	✓	✓	9/18/2025
	Hive Moderation [85]	44.51	**77.98**	65.96	✗	✓	9/18/2025

Table 4: *Performance (%) comparison for LVLM-agnostic black-box detectors (*i.e.*, commercial tools). * indicates that the corresponding tool rejects some test samples.*

Understanding their performance on political deepfakes is therefore particularly valuable, representing an instance where regular users (perhaps mediated through social media platforms) are relying on a detection tool fairly straightforwardly, as opposed to when commercial tools are used by corporate fraud experts, journalists, or forensic experts, for example, who have the expertise and resources to probe deeply into the suitability and interpretation of various detection tools. **Finally, it is important to note that our goal is not to determine which tool performs 'best' in an abstract sense, but rather to highlight the challenges and limitations these commercial systems face when applied to the detection of real-world political deepfakes**. This caveat is critical, as the tested dataset, determination of labels for authentic and inauthentic content, and particular evaluation regimes are at least partially subjective. For instance, even the determination of whether a given video is authentic or inauthentic is complicated (*e.g.*, a real newscast presenting coverage of a video deepfake, or an authentic image with a deceptive text caption). As a result, **we encourage readers to interpret results holistically, not limiting their attention to ranking the 'best' tools. Such comparisons require a dynamic and contextual understanding not captured by the quantitative metrics here alone**.

As shown in Table 4, at the image level, although only two tools achieve an accuracy above 90%, four tools obtain AUC values exceeding 90%. However, most tools perform less effectively in terms of FAR, with values above 10%. Specifically, *Reality Defender* achieves the highest overall accuracy (93.51%) and a strong AUC of 95.82%, indicating excellent discriminative ability in detecting political deepfakes. *Is It AI* and *AI or Not* also perform competitively, with AUCs of 96.15% and 95.01%, respectively, demonstrating reliable generalization. *Incode* attains a high accuracy of 91.07% while maintaining the lowest FAR (2.56%), reflecting a well-calibrated decision boundary that minimizes false positives, a particular concern and priority given the importance of discernment between true and false information in political contexts. In contrast, *BrandWell* and *Deepfake Detector* exhibit notably weaker performance, with AUCs

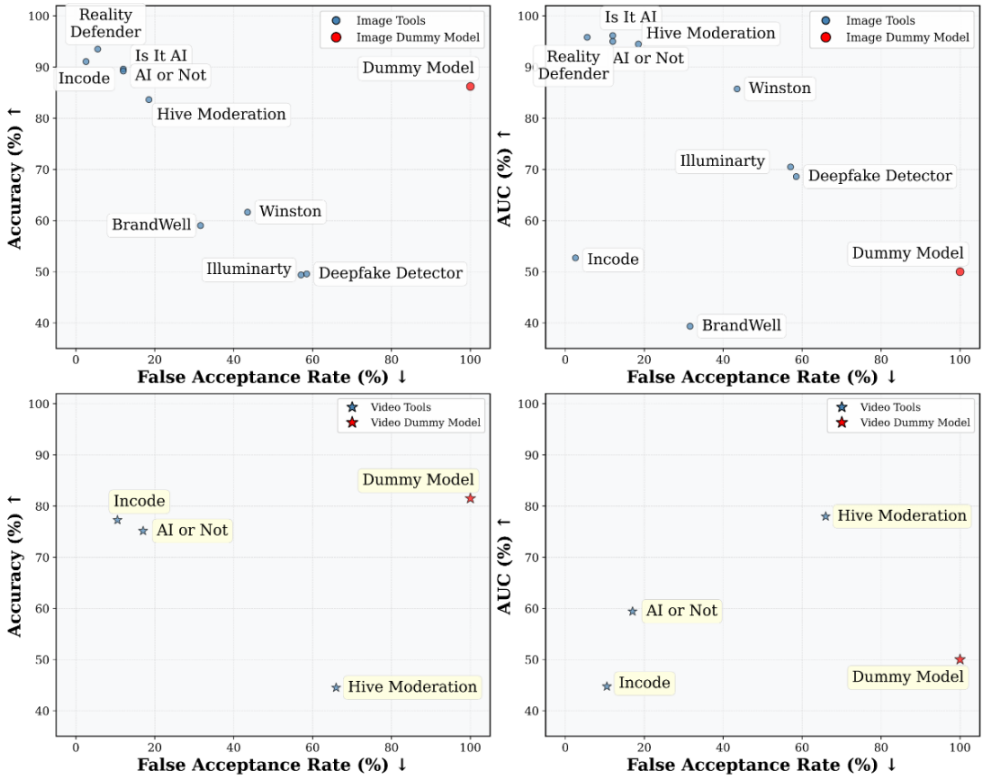

Fig. 6: *Performance comparison for commercial tools on the image set (Top) and video set (Bottom). The "dummy model" refers to a baseline that predicts all samples as fake.*

of 39.34% and 68.59%, respectively, suggesting limited robustness and poor generalization when applied to political deepfake detection, at least in the context of this study's evaluation approach and dataset.

At the video level, all three tools demonstrate lower performance compared to their results on the image set. **This indicates that detecting deepfakes in videos remains more challenging than in static images for existing commercial tools**. Among the three tools, *Incode* achieves the highest accuracy (77.27%), followed closely by *AI or Not* (75.14%), both maintaining a reasonable balance between detection reliability and false alarm rates. *Hive Moderation* attains the highest video-level AUC (77.98%); however, its elevated FAR (65.96%) suggests a strong tendency to misclassify real videos as fake, limiting its practical reliability in political video deepfake detection.

We also include dummy baseline models for both images and videos (100% of content identified as fake), as done previously for the LVLM-agnostic white-box detectors, to better understand the relative performance of the commercial tools. As shown in Figure 6, only four tools outperform the dummy model in terms of ACC on the image

Cost	LVLM	Images			Videos		
		ACC↑	AUC↑	FAR↓	ACC↑	AUC↑	FAR↓
Paid	ChatGPT(GPT-5) [87]	84.55	**96.63**	18.00	68.79	**85.28**	36.88
	Qwen-VL-Max* [88]	81.45	89.77	20.11	68.79	71.79	32.62
	Claude-Sonnet-4.5* [89]	87.50	93.01	14.00	73.99	80.04	27.66
	Gemini-2.5-pro* [90]	**88.79**	65.48	6.50	73.99	79.70	24.82
Free	LLaVA-v1.5-7B [91]	86.21	54.31	**0.00**	68.79	31.33	16.31
	LLaVA-v1.5-13B [92]	86.21	56.34	**0.00**	**81.50**	27.07	**0.00**
	LLaVA-v1.5-13B-XTuner [93]	39.11	55.91	67.88	79.77	38.61	2.13
	DeepSeek-VL-7B [94]	78.45	67.86	16.00	69.36	51.77	23.40
	CogVLM-Chat [95]	84.89	74.36	11.28	78.03	49.08	4.96
	Monkey-Chat [96]	40.00	57.36	66.15	65.32	43.66	22.70

Table 5: *Performance (%) comparison for LVLM-aware detectors.* indicates that the corresponding detector not generate correct format response for some test samples.*

set. However, all tools perform worse than the dummy model in terms of ACC on the video set, highlighting the additional difficulty of video-based deepfake detection. In contrast, most tools achieve substantially higher AUC values than the dummy model, indicating that while their decision thresholds may not be optimal, they still retain some discriminative capability.

A major epistemic and practical problem remains, however: while the latent capability of these models may exceed their measured accuracy at a standard threshold (*e.g.*, probability is greater than 0.5 for a binary determination), users and platforms are rarely in a position to meaningfully or dynamically calibrate. For instance, users of detection tools often lack access to reliable ground truth by definition, making proactive calibration difficult—if not impossible—in real-world settings. Moreover, the distribution of deepfake types continually shifts as new generation techniques, compression artifacts, and media modalities emerge, rendering optimized thresholds potentially brittle over time. Finally, contextual stakes complicate decision-making: it is unclear what the acceptable balance between false positives and false negatives should be in contexts as varied as social media moderation across diverse platforms, forensic analysis, or journalistic verification. In short, **calibration presupposes epistemic stability—yet neither the data environment nor common institutional settings provides a clear basis for knowing where, or how, to draw the line**. As these uncertainties play out in practice, platforms risk (unknowingly or untestably) deploying oversensitive or miscalibrated systems that either overflag authentic content or miss harmful manipulations, producing uneven performance across contexts and unavoidably exposing the public to harms.

2.2.3 LVLM-Aware Detectors

Recently, Large Vision-Language Models (LVLMs) have gained popularity by enabling users to upload visual media and query them directly through natural language. Many paid LVLMs regularly updated and fine-tuned with new and diverse data, improving their adaptability and alignment with recent real-world content—similar to the continuous maintenance in commercial detection tools. Beyond this, they also incorporate a Retrieval-Augmented Generation (RAG) module [97], which allows them to

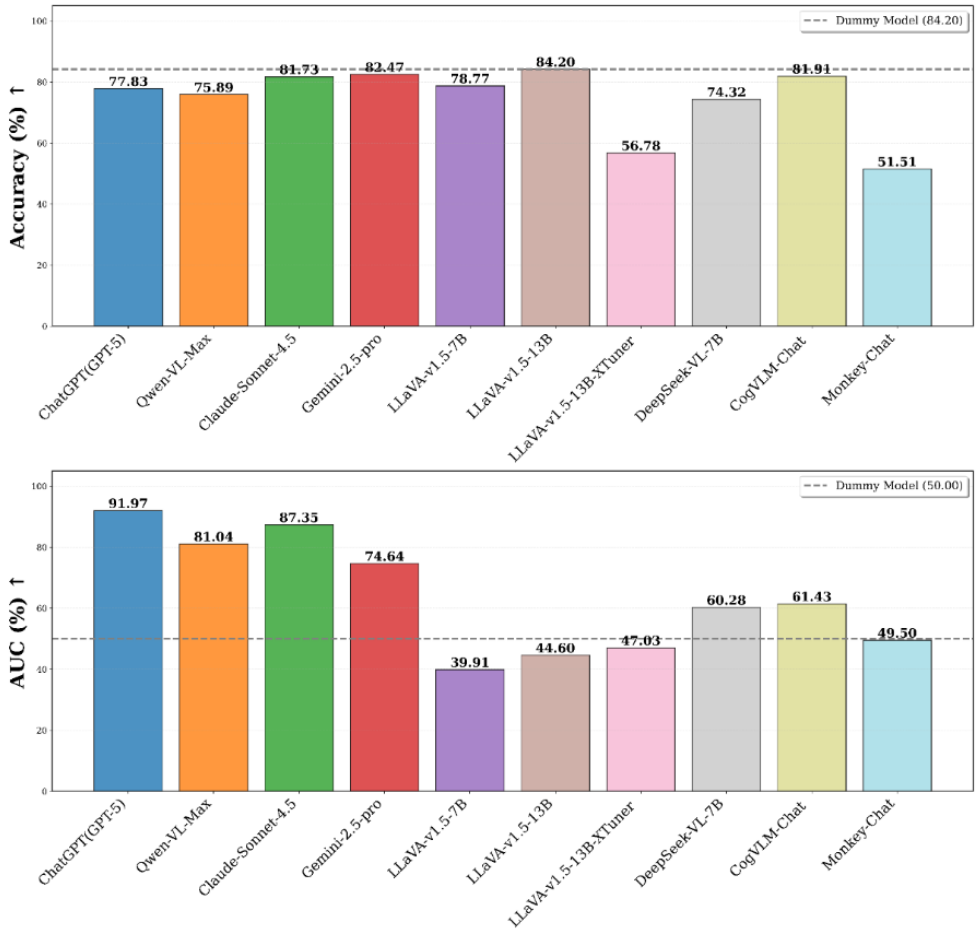

Fig. 7: *Performance comparison between LVLM-aware detectors and "dummy model" on the image and video mixed set. The "dummy model" refers to a baseline that predicts all samples as fake.*

access and analyze up-to-date online information and integrate the retrieved evidence into their reasoning process, potentially producing more contextually grounded and accurate responses. Motivated by this capability, we evaluate the effectiveness of current LVLMs in detecting political deepfakes. In this experiment, we assess ten widely used LVLMs—four requiring paid API access and six freely available online.

The performance results for both image and video evaluations are summarized in Table 5. (1) In general, none of the evaluated LVLMs achieve over 90% accuracy across both data modalities. However, in terms of AUC, two LVLMs exceed 90% performance on the image set. Interestingly, LLaVA-v1.5-7B [91] and LLaVA-v1.5-13B [92] achieve a 0% False Acceptance Rate (FAR) on images, with LLaVA-v1.5-13B maintaining this 0% FAR on videos as well. This outcome suggests that LLaVA-v1.5-13B classifies all

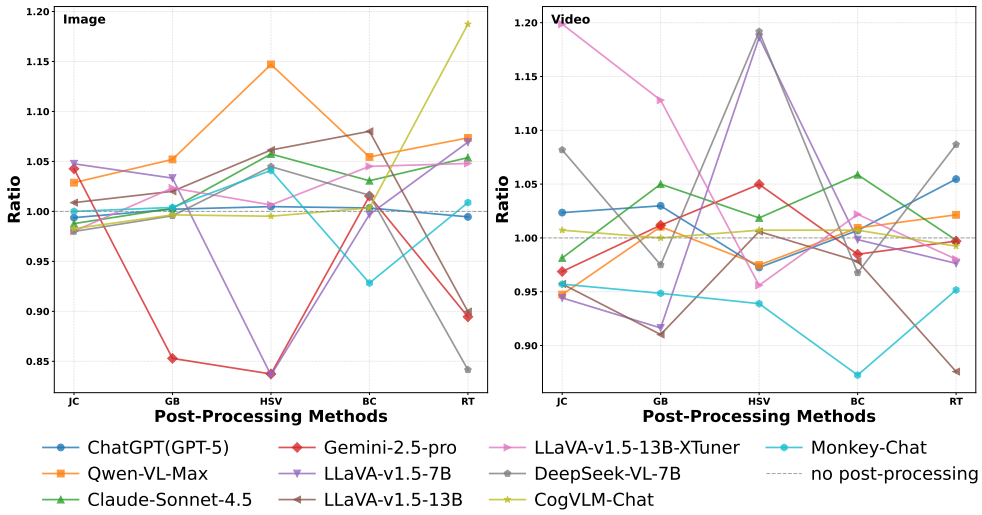

Fig. 8: *Performance (AUC) ratio (original vs. post-processed) for LVLM-aware detectors.*

samples as fake, effectively yielding no false acceptances but at the cost of high false rejections. (2) We also observe that **the paid LVLMs generally achieve higher AUC performance compared to the free-access models**. In contrast, most free LVLMs perform close to random guessing, with AUC values hovering around 50%, indicating limited capability in distinguishing real from fake political media. (3) Among all evaluated LVLMs, ChatGPT (GPT-5) demonstrates the best overall balance, achieving a strong image-level AUC of 96.63% and video-level AUC of 85.28%, indicating reliable discriminative performance across both modalities. Claude-Sonnet-4.5 follows closely, with competitive AUCs of 93.01% for images and 80.04% for videos, suggesting good generalization capability. However, both LVLMs exhibit relatively high False Acceptance Rates (FARs), which could mislead general users and increase the risk of misjudging political deepfakes as authentic.

We also combine the image and video sets into a single dataset and introduce a dummy baseline model (an "always-fake" predictor) to better understand the overall performance of LVLMs in detecting real-world political deepfakes. The results, presented in Figure 7, show that none of the LVLMs outperform the dummy model in terms of Accuracy. As discussed earlier, LLaVA-v1.5-13B [92] predicts all samples as fake, thereby matching the dummy model's behavior. In terms of AUC, all paid LVLMs exhibit relatively higher performance than the dummy model, reflecting stronger discriminative capabilities. In contrast, most free LVLMs perform near the dummy baseline, with four of them achieving even lower AUC scores, indicating a tendency to effectively predict in the wrong direction.

As with the robustness evaluation conducted for the LVLM-agnostic white-box detectors, we perform the same set of robustness experiments for all LVLMs. The

results, presented in Figure 8, reveal that **LVLMs exhibit greater robustness compared to detectors from academia and government**, as most models maintain a performance (AUC) ratio within the range of [0.95, 1.05]. A potential explanation is that these LVLMs, such as ChatGPT (GPT-5), are trained on million- to billion-scale multimodal datasets comprising images, videos, and text. Such large-scale and diverse training may enable them to handle variations in input quality more effectively and to internally normalize post-processed inputs to align with the data characteristics encountered during training. However, Qwen-VL-Max demonstrates higher sensitivity to HSV (*i.e.*, random adjustments to Hue, Saturation, and Value channels) post-processing effects on image data, showing a sharp performance decline (around 1.15 ratio). Similarly, several free-access LVLMs (*e.g.*, LLaVA-v1.5-13B-XTuner, LLaVA-v1.5-7B, DeepSeek-VL-7B) experience substantial performance degradation on the video set. Interestingly, we also observe that the performance of certain LVLMs (*e.g.*, Gemini-2.5-pro, Monkey-Chat) actually improves after post-processing (with a lowest ratio close to 0.84), a finding that urges caution rather than confidence, as these transformations should not increase model performance, should remain stable (ratio close to 1.00) if detection techniques are robust. A plausible explanation is that these transformations introduce additional visual artifacts into fake images, making the manipulations more detectable. In cases where these models initially misclassified subtle fakes, the added distortions may have amplified cues that align better with their learned representations of synthetic content, thus enhancing detection accuracy. These results indicate that, despite their overall resilience, robustness remains a persistent challenge for a few LVLMs.

3 Discussion

Our benchmark presents the first systematic evaluation of existing deepfake detectors on real-world, politically-oriented deepfakes–as well as cheapfakes and alleged instances of false content–collected from social media. Importantly, we adopt the perspective of an average end user, testing each model in its original released form, without additional fine-tuning or domain adaptation, to reflect realistic deployment conditions. According to the experimental results, we have the following findings.

Finding 1: Our analysis of the PDID dataset highlights its **unique characteristics** and provides insight into the distribution of political deepfakes observed in real-world social media settings in recent years. Compared with existing datasets (Table 1), the PDID more accurately represents in-the-wild political deepfakes (both images and videos) and provides a valuable foundation for benchmarking detection systems. This fundamentally distinguishes our work from most prior benchmarks (Table 2), which train and evaluate models exclusively on laboratory-generated deepfakes. Such datasets are typically synthesized/generated using a single generative model or tool, limiting their ability to capture the diversity and complexity of real-world manipulations. Moreover, many of these synthetic deepfakes were never publicly disseminated on social media, reducing their practical relevance (*e.g.*, real impacts on the public, elections, or institutional trust) and, consequently, the significance of benchmark results built upon them.

Finding 2: Through frequency-domain analysis of real and fake samples within the PDID, we observe that the authentic media in the PDID differs substantially from the authentic media commonly used to generate laboratory-based deepfakes. This divergence becomes even more pronounced when comparing fake media in the PDID with lab-generated deepfakes. These findings demonstrate the distinctiveness of the PDID dataset, whose **data characteristics are rarely seen in existing benchmarks**, despite the significance of studying content that has actually been disseminated and interpreted in contexts with potential political implications. The results also suggest that the generative methods underlying the political deepfakes in the PDID may involve previously unseen or heterogeneous synthesis pipelines (*i.e.*, unrestricted to any single model or framework), reflecting the uncontrolled and evolving nature of real-world generative media.

Finding 3: Our evaluation of LVLM-agnostic white-box deepfake detectors indicates that detectors developed in academia and government are **not yet suitable** for direct deployment in real-world environments. In particular, government-developed detectors exhibit notably low performance, likely due to restrictions in training data and the lack of continued maintenance or updates. This finding underscores the importance of sustained governmental support for deepfake detection research to ensure that publicly funded or endorsed systems remain trustworthy and effective for general users. Fortunately, the mission of the original SemaFor program is being continued under the Digital Safety Research Institute (DSRI) of UL Research Institutes [98], highlighting an ongoing institutional commitment to advancing media forensics. Our results also suggest that such detectors require further fine-tuning and domain adaptation before being applied to politically-oriented or real-world datasets. Moreover, the relatively stronger performance of frequency-based detectors points to a promising direction for improving the generalization ability of LVLM-agnostic white-box models—specifically, by emphasizing the learning of signal-level frequency-domain artifacts during training.

Finding 4: Through our evaluation of LVLM-agnostic black-box deepfake detectors (*i.e.*, commercial tools) from industry, we observe that a few tools do achieve relatively strong AUC performance on the image set. However, most tools exhibit low accuracy and a high FAR, indicating that **accurate decision-making depends heavily on careful threshold selection**. This suggests that commercial platforms should provide clearer, more intuitive guidance to help non-expert users interpret results effectively to maintain user trust and engagement, though helping users calibrate is far from straightforward and represents an important research challenge. Next, although several tools perform reasonably well on images, they demonstrate a significant drop in performance on videos, revealing that political video deepfake detection remains a major challenge for current commercial tools. Consequently, further development is needed to enhance robustness in dynamic visual contexts. This is critically significant because of the diversity of multimodal content: video with altered frames, authentic videos with authentic audio, authentic images with misleading text captions, and so on. Until such improvements are achieved and tools are carefully, contextually implemented, average users should exercise significant caution when interpreting results from commercial tools, particularly when testing political video deepfakes.

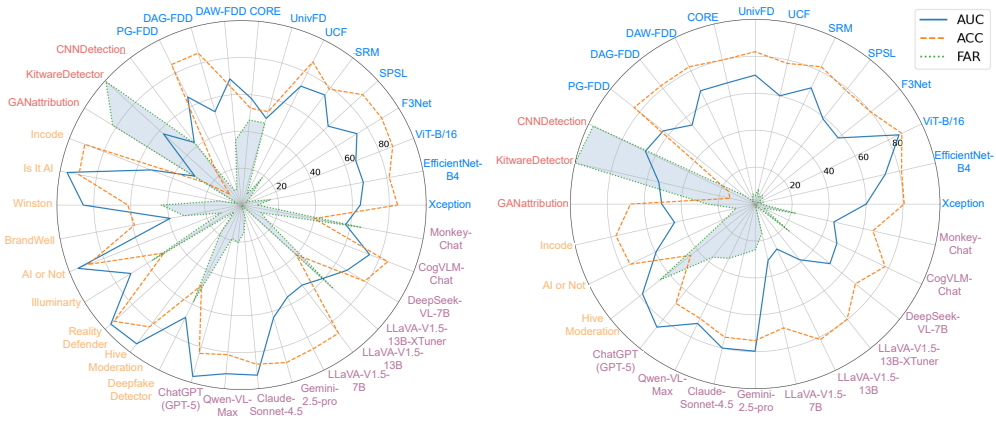

Fig. 9: *Performance overview of all detectors on images (Left) and videos (Right).*

Finding 5: Recently, the use of LVLMs by the general public has increased substantially across a wide range of tasks. Leveraging LVLMs for political deepfake detection has therefore become a promising direction. Our experimental results indicate that **paid LVLMs relatively outperform free-access counterparts**. This advantage can likely be attributed to two key factors: (1) paid LVLMs are regularly updated and fine-tuned with new and diverse data, improving their adaptability and alignment with recent real-world content—similar to the continuous maintenance in commercial detection tools; and (2) many paid LVLMs incorporate a RAG module [97], which enhances performance through online information retrieval and implicit fact-checking. This built-in retrieval capability allows the models to cross-validate visual or textual inputs against up-to-date external knowledge, thereby improving their accuracy in identifying political deepfakes. While the paid LVLMs demonstrate strong potential, achieving high AUC scores in both image and video detection, their overall reliability remains limited due to consistently high FARs. This indicates that users must carefully adjust or interpret probability thresholds before making final authenticity judgments. Nevertheless, one notable advantage of LVLM-based detection is their robustness: **most LVLMs exhibit stable performance and are largely insensitive to common post-processing operations** applied to the original media, suggesting better resilience to real-world variations in content quality.

Finding 6: To better visualize and interpret the performance of current detectors in identifying political deepfakes, we generate radar plots for both image and video sets, as shown in Figure 9. Overall, it is evident that **paid detectors consistently outperform free-access detectors in terms of AUC**. This discrepancy can likely be attributed to the continued maintenance and regular updates of paid models, which often incorporate newer and more diverse training data, thereby enhancing their generalization capability. However, **political video deepfake detection remains a major challenge for most detectors**, as reflected by the noticeable performance decline compared to image-level AUC results. Furthermore, our analysis suggests that with careful threshold tuning, it is possible to achieve both high ACC and low FAR, as observed in some academic video detectors. Nonetheless, this process requires

substantial technical knowledge and contextual decision-making, which **limits its practicality for non-expert users** in real-world scenarios.

In general, our systematic benchmarking reveals that existing deepfake detectors struggle to generalize effectively to real-world political cases–the very cases most likely to promote propaganda, distort democratic processes, and undermine trust in news and political institutions. Most detectors exhibit substantial drops in accuracy and AUC when tested on authentic political deepfakes than their original reported performance, potentially due to domain shifts, limited diversity in training data, and overreliance on laboratory-generated synthetic media. Even advanced LVLMs, while more robust to post-processed media, show inconsistent decision thresholds and high false acceptance rates that could mislead average users. These findings reveal an **urgent need to rethink current detection paradigms**. Future research could prioritize politically contextualized and diversified datasets, cross-modal reasoning, and explainable detection frameworks capable of operating under dynamic real-world conditions. Building detectors that not only identify manipulation but also communicate uncertainty and reasoning transparently will be critical for restoring public trust and ensuring accountability in the age of AI-based generative political media.

Complicating this is the difficulty in calibrating detection tools, communicating appropriate and inappropriate uses clearly, and encouraging contextual adaptation (*e.g.*, by platforms) beyond straightforward integration or binary prediction determinations. Deepfake detection needs to remain accessible from a user design perspective, not just technically robust, or else it risks undermining its own goals and providing false assurance. For instance, a detection tool that only operates on human faces will fail to capture deepfake content of a stolen ballot box, while a detection tool that promises accuracy while failing to address simple manipulations like cheapfakes risks falsely assuring the public, while leaving open major threat vectors for hostile actors to exploit. In sum, despite promising advances made by deepfake detection tools, these tools must be developed, evaluated, implemented, and communicated in the context of real-world problems if they are to deliver on their promise.

Moreover, our results also reiterate the need to approach detection as one layer in a multi-pronged effort to improve public discernment and safeguard platforms. Even effective use of technical detection tools requires, as suggested above, careful communication, user education, and contextual implementation. Beyond that, our study's identification of ground truth labels required a sociotechnical effort–not just examining technical artifacts, but necessarily relying on experts who analyzed source cues and esoteric political context. This strongly suggests that appropriate safeguards require going far beyond technical detection. Media and digital literacy, human-centered design, platform policies, formal regulation, and other layers of governance are required to realize and complement technical detection efforts, if they are to be ultimately successful.

Although our benchmarking study is systematic, several limitations remain. (1) *Limited sample size.* Due to the nature of real-world political deepfakes, collecting a large number of verified samples is extremely difficult, despite our continued efforts through the PDID. Consequently, the relatively small test set may introduce minor statistical bias in the results. Moreover, the majority of deepfakes come from the US

political context, limiting both representativeness of the content and generalizability of the evaluation (e.g., a large portion of the content focuses on just a few political figures from a few ethnic backgrounds) (2) *Focus on visual deepfakes.* Our benchmark exclusively targets visual deepfakes. However, many political manipulations are multimodal, such as those involving synthetic or manipulated audio. While such cases were excluded during data preparation, our findings may not directly extend to more complex multimodal deepfakes that combine visual and audio manipulations. (3) *Face-centered training bias in academic detectors.* The academic detectors included in this study were primarily trained on facial images. As a result, manipulations that occur outside the face region may go undetected if the facial region remains unaltered, implicating the appropriateness of our testing and scoring approach and urging against simplistic interpretations of the results. (4) *Temporal limitation of commercial tool evaluation.* Our evaluation of commercial tools was conducted within a short time window. Given their frequent updates, the performance of individual tools may fluctuate over time. Moreover, due to budget constraints and high subscription costs[4], we were unable to continuously monitor performance changes or conduct a broader suite of robustness assessments using multiple post-processing variations. (5) *Subjectivity in ground truth determinations.* Despite efforts to use the most robust existing dataset and incorporate multiple sources of evidence (professional annotators, political and deepfake experts, source cues from social media, external verification from fact checkers), ground truth determinations are not unimpeachable. For instance, it is not given whether an authentic newscast featuring coverage of a deepfake should be considered true or false from a holistic perspective, and either determination could be 'unfair' from the perspective of a certain detection procedure or tool. (6) *Scope of evaluated LVLMs.* The LVLMs included in our benchmark were not specifically designed for media forensics. Although deepfake-specialized LVLMs [99] may yield stronger results, our study intentionally focuses on widely accessible, general-purpose LVLMs to reflect the realistic experience of average users. (7) *Unified prompting strategy.* To maintain consistency across LVLMs, we used a single, standardized prompt. Testing with multiple or adaptive prompts could potentially improve prediction quality but would also introduce greater variability, increase experimental complexity, and compromise comparability across models.

Despite these limitations, we believe that **our findings are already valuable in revealing the existing challenges and limitations** faced by current detectors in identifying real-world and potentially consequential political deepfakes. The insights gained from this study also highlight several promising directions for future research. In the future, we plan to expand the benchmark by increasing the test sample size and incorporating multimodal deepfakes. Additionally, we intend to evaluate newly developed detectors and deepfake-specific LVLMs to obtain deeper and more comprehensive insights into their effectiveness and generalization across diverse real-world scenarios.

[4]Evaluation procedures for commercial tools relied on their official APIs and documentation. Some providers offered complimentary credits or trial access for preliminary testing; however, all reported results were generated using independently secured access.

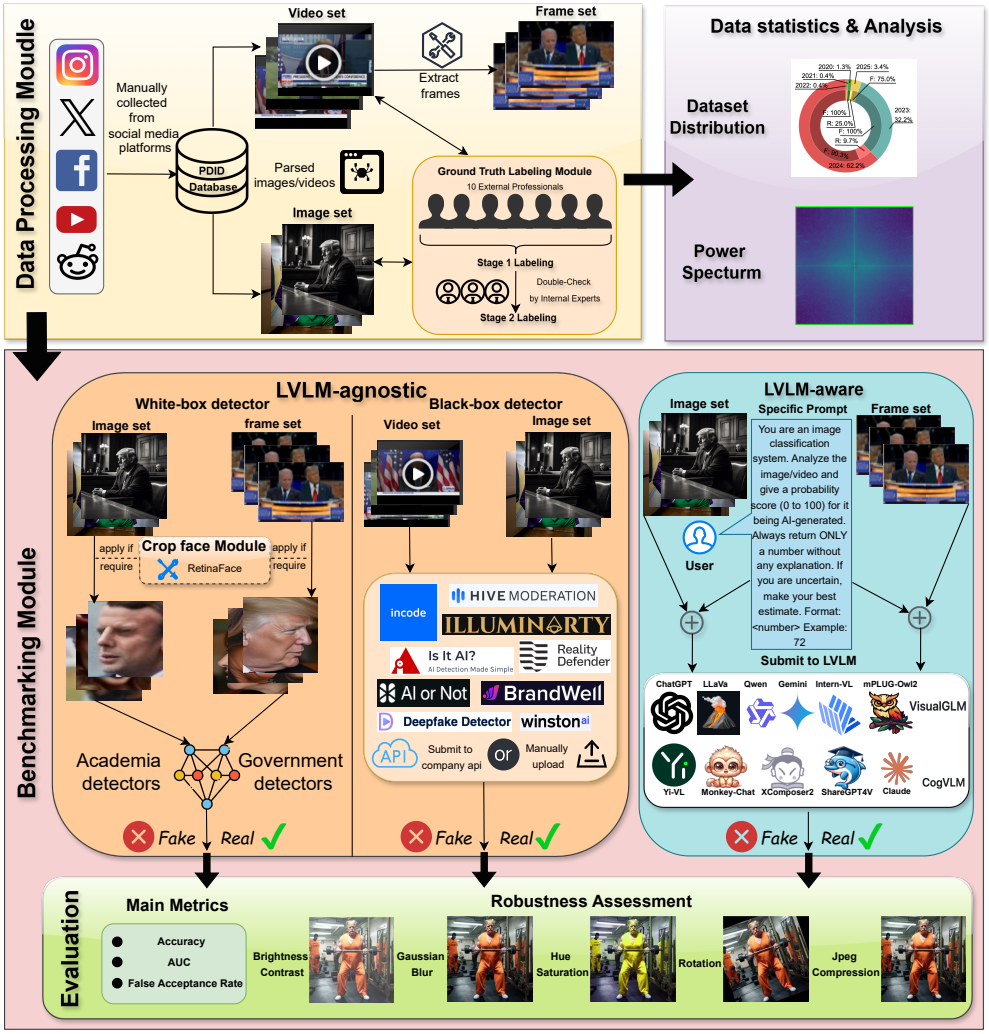

Fig. 10: *Overview of the entire pipeline for benchmarking.*

4 Methods

In this section, we describe in detail the methods used for data processing and benchmarking of existing detectors. An overview of the entire pipeline is illustrated in Figure 10.

4.1 PDID

4.1.1 Data Collection

The PDID provides a novel repository of politically oriented synthesized media. The dataset is inclusive of deepfakes, cheapfakes, and other incidents *alleged* to be deepfakes or cheapfakes. As described in [100], data collection involves a team of individuals-primarily graduate and undergraduate researchers with expertise in political deepfakes-sourcing, ingesting, and annotating identified incidents according to a structured cookbook. Analysts source incidents primarily from five major social media platforms, including monitoring each platform weekly, utilizing independent Democratic-leaning and Republican-leaning accounts, designed to facilitate more robust coverage reflective of the kinds of content that might be targeted at individuals and algorithmically mediated. The data collection process also involves following major news articles: deepfakes reaching this level of attention are de facto considered prominent enough to constitute incidents. Finally, data are ingested from other initiatives aimed at collecting deepfakes, like PolitiFact, Hany Farid's Deepfakes in the 2024 Election Dataset, the WIRED AI Elections Project, and the AI Incidents Database. To improve annotation quality and reliability, incidents are currently primarily limited to English-language and often US-based incidents.

After sourcing, two individuals annotate every incident according to the structured codebook, with reconciliation by a third coder in cases of persistent disagreement. This effort includes capturing an extensive set of metadata, such as media format (image, video, audio-visual), whether the content is *presented* as real or fake, whether the content is determined to be a deepfake or cheapfake, who created or shared the deepfake, metrics on social media views or reposts, and information about who is targeted in the content. We capture incidents that are politically salient deepfakes or cheapfakes, or images or videos that people have called fake, but might be authentic. To be politically salient, the image or video has to either be posted by a politician, include a political figure in the content, or be about a politically-connected event (e.g., a natural disaster response or election). Additional theoretically informed variables are annotated, covering domains such as communication objectives, depictions of harm, policy narratives, and evidence of real-world consequences. Additionally, evidence of external verification of truth or falsity is captured, focusing on objective information from fact checkers, journalists, or forensic experts, and annotators additionally note both technical artifacts (e.g., too many fingers) and contextual cues (e.g., a certain political figure was no longer living in a given year). Information on self-disclosure (e.g., the poster indicated the image is fake) and on watermarking or labeling is included as well, and annotators incorporate additional qualitative commentary about context and verification to explain incidents and coding decisions and improve research transparency. The complete codebook and associated typologies are available at https://doi.org/10.17605/OSF.IO/FVQG3.

The initial resulting dataset from which this study's test data set is drawn has a total of 939 images and 502 videos, as of October 2025, covering the time period of 2018 through the present. Unlike purely technical data sets or mere collections of images or videos, the PDID bridges the gap between AI research and sociopolitical analysis

by contextualizing incidents within relevant political frameworks, policy narratives, and documented societal impacts, thereby enabling rigorous investigation of synthetic media's role in information integrity challenges. For instance, the rich metadata fields or human explanations provided in the PDID can enable additional research needed to fine-tune reasoning-based LVLMs, drawing from information about source cues, text from broader social media threads, or context about political events that may be difficult to retrieve otherwise, especially retrospectively.

4.1.2 Two-Stage Label Generation

In this work, we select images and videos from the PDID generated between 2018 and September 2025, initially excluding those where the research team felt authenticity was too difficult to determine, resulting in 232 images and 179 videos. While the PDID provides evidence of external verification (e.g., by professional fact checkers), this form of third-party verification only covers a modest subset of images/videos, many of which are labeled as "unclear either way." As such, we engaged in a more intensive **two-stage** labeling process to identify the ground truth (either *Fake* or *Real*) for all images and videos.

Stage 1 (External Professional Labelers): We first engaged ten professional annotators from an external data labeling company. Since the labelers did not have prior expertise in deepfakes, we provided them with a detailed image and video labeling instruction document (see Supplementary Information for details). To facilitate unbiased human judgment, labelers were not permitted to conduct external research, search online, use fact-checking websites, or employ any automated detection tools. Their decisions were to be based solely on visual and contextual cues described in the provided instructions. Each labeler assigned one of three possible labels to every image or video:

- *Real*: Appears authentic, with no clear signs of manipulation.
- *Fake*: Appears manipulated, generated, or altered. This includes: deepfakes and AI-generated content, photoshopped images, memes and cartoons, cheapfakes (simple edits), and any content that has been significantly modified from reality.
- *Uncertain*: Insufficient evidence to confidently determine authenticity.

Specifically, labelers were instructed to consider visual/image cues, social media and contextual cues, real-world plausibility, and self-disclosure indicators when making their determinations. Random guessing was strictly prohibited, and the Uncertain label was to be used only when the labeler was genuinely unable to make a confident judgment. To minimize potential ordering bias, the presentation order of images and videos was randomized for each labeler. After all annotations were completed, we aggregated the results and computed a **confidence level** for the "Fake" label for each sample, defined as follows:

- *Low confidence*: Consensus among 0–4 labelers.
- *Medium confidence*: Consensus among 5–7 labelers.
- *High confidence*: Consensus among 8–10 labelers.

Stage 2 (Internal Political Deepfake Expert Review): Next, a set of researchers with specific expertise in political deepfakes and in the U.S. political context conducted a second round of verification to ensure labeling accuracy. Specifically,

we re-examined all images and videos that received **low** or **medium confidence** ratings from Stage 1 using a rigorous cross-checking strategy. This included using fact-checking websites such as Snopes [101] and AFP fact check [102] to verify whether images and videos were real or fake, as well as utilizing knowledge about the political context and real-world events, drawing on self-disclosure indicators and information about posters or sharers of content (e.g., usernames, country of origin). Because the dataset–focused on deepfake incidents–is inherently imbalanced (with a significantly higher number of fake samples than real ones), there is a high risk of mislabeling real images/videos. To mitigate this, the political deepfake expert team additionally reviewed all samples that were labeled as Real by at least one professional labeler in Stage 1. During this stage, we exclude previously labeled real video samples that contain fake audio, as this study focuses exclusively on **visual deepfakes** and a holistic judgment of truth sensitive to audio would not align with results based on audio-agnostic detection tools (the majority assessed in our study). Based on our expert assessment, we assign final labels.

It is important to note that many artifacts fell into hybrid categories—for instance, real footage with fake captions (*e.g.*, video footage of a real war, but described as taking place elsewhere), or authentic news segments displaying deepfake clips. In these cases, labeling was based on the dominant communicative intent of the artifact as a whole rather than the presence of localized synthetic elements. This distinction matters for the utility of deepfake detection tools as well as how to fairly evaluate them. Yet in practical detector deployment, providers may overpromise the capabilities of their detectors or underestimate associated challenges, and users likewise may apply such tools to content without such awareness, potentially leading to false confidence. Detection tools that attend narrowly to technical artifacts of diffusion models or GANs risk missing the forest through the trees, as 1) simple edits or manipulations (cheapfakes), 2) strategic hybrid use of authentic and synthetic content or 3) of real and misleading content, or of 4) multi-modal content could induce serious errors in the accuracy of detection tools and impact on the public.

In turn, and to promote a fair evaluation, if a sample was deemed too ambiguous to confidently categorize, we re-designated it as Uncertain and excluded it from the final dataset. Only a small number of samples were excluded in this way. The remaining samples, which achieved high confidence and passed all review criteria, were retained as fake or real. This two-stage validation process resulted in a final curated dataset containing 232 images and 173 videos with verified authenticity labels (*i.e.*, 214 fake images, 18 real images, 146 fake videos, 27 real videos).

4.1.3 Expose Frequency Domain Artifacts

To identify the specific manipulation artifacts present in the PDID dataset and compare them with those produced by established techniques (*e.g.*, variational autoencoders, generative adversarial networks, and diffusion models), we conducted a frequency domain analysis. This analysis aims to assess whether recent real-world fake images are indeed generated by known academic methods. To this end, we randomly extract one frame from each video using the OpenCV library (`cv2`) [103] and add

them to the original image set, resulting in the final test set with 405 (*i.e.*, 232+173) samples used for this experiment. All images and frames are resized to 224×224.

Inspired by [104], we extract all fake images. For each image x_i, where $i \in \{1, \cdots, I\}$ and I is the total number of fake images, we extract its noise residual r_i by subtracting its denoised version. This process is formulated as follows:

$$r_i(m,n) = x_i(m,n) - D\big(x_i(m,n)\big), \quad i = 1, 2, \ldots, I. \tag{1}$$

Here, (m, n) denotes the pixel coordinates, and $D(\cdot)$ represents the denoising filter. In this study, we use the learned denoising network DnCNN [105] as the denoising filter.

Next, we apply the Discrete Fourier Transform (DFT) to convert each noise residual r_i into its frequency domain representation. Let M and N denote the image height and width, respectively. The transformation is defined as follows:

$$F_i(k,l) = \sum_{m=1}^{M} \sum_{n=1}^{N} r_i(m,n)\, e^{-j\,2\pi\left(\frac{k}{M}m + \frac{l}{N}n\right)}. \tag{2}$$

Here, (k, l) denotes the frequency domain coordinates, and $F_i(k, l)$ is the complex Fourier coefficient at location (k, l) for the i-th image.

To capture the characteristic artifacts of the image source, we compute the average power spectral density $S_x(k, l)$ across all images, defined as:

$$S^{\text{fake}}_{\text{PDID}}(k,l) = \frac{1}{I} \sum_{i=1}^{I} \left|F_i(k,l)\right|^2. \tag{3}$$

Similarly, we extract all real images and apply the same procedure to obtain $S^{\text{real}}_{\text{PDID}}(k, l)$.

For comparison, we utilize subsets from the latest lab-based AI-Face dataset [55], which includes traditional deepfake images from four sources, GAN-generated images from ten sources, DM-generated images from eight sources, and their corresponding real images. Specifically, the real images are sourced from FF++ [38], DFDC [36], DFD [106], and Celeb-DF [39] for the deepfake category; FFHQ [59] for GANs; and IMDB-WIKI [58] for DMs. For each category (*i.e.*, Deepfake, GAN, and DM), we randomly select 1,000 fake images and 1,000 real images from the corresponding source. Following the same procedure described above, we compute $S^{\text{fake}}_{\text{category}}(k, l)$ and $S^{\text{real}}_{\text{category}}(k, l)$, where category $\in \{\text{Deepfake}, \text{GAN}, \text{DM}\}$. Figure **??** illustrates the spectral artifacts across different categories for both fake and real images.

4.2 Benchmark Settings

4.2.1 Detectors

To build a comprehensive benchmark, we systematically evaluated a diverse set of deepfake detectors, categorized as described below. All detectors were used with their publicly available pre-trained weights or accessed through APIs. No additional training, fine-tuning, or adaptation was performed.

1. **LVLM-agnostic white-box detectors**. In this category, we include detectors developed from academia and government. Specifically, we consider the following models:
   (a) For detectors developed by the **academic** community, we adopt 12 pre-trained detectors from the latest benchmark study: AI-Face [55]. These detectors are organized into four major types:
      (i) *Naive detectors*: Backbone architectures used directly for binary classification, including convolutional models such as Xception [30] and EfficientNet-B4 [66], as well as transformer-based models like ViT-B/16 [67].
      (ii) *Frequency-based Detectors*: Models that exploit frequency-domain cues for forgery detection, such as F3Net [32], SPSL [68], and SRM [69].
      (iii) *Spatial-based Detectors*: Models that leverage spatial features (*e.g.*, textures or inconsistencies) for detection, including UCF [70], UnivFD [71], and CORE [72].
      (iv) *Fairness-enhanced Detectors*: Approaches designed to improve fairness in AI-generated face detection by incorporating fairness-aware learning objectives. This includes DAW-FDD, DAG-FDD [73], and PG-FDD [31].
   (b) From the **government** sector, we include three detectors developed under the DARPA Semantic Forensics (SemaFor) program [77], each of which accepts images as input: CNNDetection [74] is a classifier trained on ProGAN [107]-generated images. KitwareDetector [75] is a detector developed by Kitware, Inc., designed to detect GAN-generated images. GANAttribution [76] is a detector that goes beyond binary detection to attribute GAN-generated images to their source generators by identifying distinctive fingerprints left by different generative models.
2. **LVLM-agnostic black-box detectors**. To evaluate the performance of existing commercial deepfake detection tools on our dataset, we first identify nine image-based services and three video-based services. Access to these tools is obtained either through direct purchase, subscription, or by contacting the vendors to request testing access. Specifically,
   (a) The image-based commercial tools evaluated are: *Incode, Is It AI, Winston, BrandWell, AI or Not, Illuminarty, Reality Defender, Hive Moderation, and Deepfake Detector.*
   (b) The video-based commercial tools include: *Incode, AI or Not, and Hive Moderation.*

   It is important to note that a few of these tools reject a subset of the test samples. In such cases, performance metrics are calculated only on the accepted samples for each specific tool.
3. **LVLM-aware detectors**. Recognizing the rapid progress in LVLMs, we evaluate ten models capable of accepting image inputs and generating textual responses with explanations. Specifically, we assess the performance of 10 LVLMs: ChatGPT(GPT-5) [87], Qwen-VL-Max [88], Claude-Sonnet-4.5 [89], Gemini-2.5-pro [90], LLaVA-v1.5-7B [91], LLaVA-v1.5-13B [92], LLaVA-v1.5-13B-XTuner [93], DeepSeek-VL-7B [94], CogVLM-Chat [95], and Monkey-Chat [96]. For the first four LVLMs, we purchased API access for evaluation. The

remaining LVLMs were freely accessible online. To ensure consistent and token-efficient outputs, we used the following prompt for all of them:

```
You are an image classification system. Analyze the image/video and give
a probability score (0 to 100) for it being AI-generated. Always return
ONLY a number without any explanation. If you are uncertain, make your
best estimate. Format: <number> Example: 72
```

4.2.2 Dataset

In our evaluation, we use the newly constructed image set and video set to assess all detectors. To make the comparison fair, we design the following strategies for each detector category.

For **LVLM-agnostic white-box detectors**, we leverage access to the source code to adapt our data according to each detector's input requirements.

1. If a detector expects the full image, we retain the original image format for the image evaluation. For video evaluation, since all detectors in this category are image-based and do not support direct video input, we test the 10 frames extracted for each video. To compute a final prediction score per video, we adopt the following *frame-based aggregation rule*:
   (a) If any frame yields a detection probability greater than 0.5 (*i.e.*, likely fake), we average only those frame probabilities to obtain the final video-level probability.
   (b) If all extracted frame probabilities are below 0.5 (*i.e.*, likely real), we compute the average over all frames.

   This rule ensures that detectors remain sensitive to subtle manipulations, enabling the identification of a fake video even when only a single frame is manipulated.
2. If the detector requires cropped facial regions, we apply the RetinaFace model [108, 109] to automatically locate and extract face regions. The resulting face crops are then resized to the detector-specific input dimensions (*e.g.*, 224×224, 256×256). Note that if no faces are detected in an image or in a whole video, that sample is excluded from the final evaluation. Therefore,
   (a) For image evaluation, the above procedure produces a total of 586 cropped face images from the original 232 images. The final prediction score for each original image is defined as the highest face-level prediction score among all detected faces within that image.
   (b) For video evaluation, following the previous rule, we use the 10 frames extracted for each video from the preprocessing procedure. Face cropping is applied where required by the detector. The final prediction score for each frame is defined as the highest face-level prediction score among all detected faces within that frame. The final prediction score for each video is computed according to the *frame-based aggregation rule* described previously.

For **LVLM-agnostic black-box detectors**, all models in this category are proprietary and developed by commercial vendors, typically requiring a subscription or one-time purchase for access. We evaluate these detectors using all 232 images and

173 videos. Notably, we do not extract frames from videos for this evaluation, as most detectors in this category return prediction probabilities directly at the video level.

For **LVLM-aware detectors**, several models (*e.g.*, ChatGPT, Qwen) require a subscription or purchase for access. To manage cost while maintaining consistency, we evaluate all detectors in this category using the original curated set of 232 images and 173 videos. For video testing, we use the 10 frames extracted for each video. We then follow the same aggregation rule used for LVLM-agnostic white-box detectors: if any frame yields a probability above 0.5 (indicating likely fake), we compute the final video score by averaging those frames; otherwise, we average all extracted frame scores. This approach ensures consistent and cost-effective evaluation across detectors.

Robustness Assessment. Although our media are originally collected from social media, they may be redistributed across various platforms in the future. It is therefore critical to evaluate the performance of existing detectors under such dissemination scenarios. To this end, we conducted a robustness assessment for both LVLM-agnostic black-box detectors and LVLM-aware detectors. Following the protocol in [55], we applied five post-processing operations to all images and extracted video frames to simulate real-world degradations. The transformations include:

- JPEG Compression (JC) – re-encoding images at a fixed quality level;
- Gaussian Blur (GB) – applied with a predefined kernel size;
- HSV Shift (HSV) – random adjustments to hue, saturation, and value channels;
- Brightness/Contrast Adjustment (BC) – random modifications within specified bounds;
- Rotation (RT) – applied up to a fixed angle, with padding added to preserve image dimensions.

4.2.3 Evaluation Metrics

To evaluate detection performance, we use the following utility metrics: ACC, AUC, and FAR. Higher values of ACC and AUC indicate better performance, while a lower value of FAR is preferred. Unless otherwise specified, a fixed threshold of 0.5 is applied to compute ACC and FAR. For LVLM-agnostic black-box detectors, if a tool provides binary output (*i.e.*, "Real" or "Fake"), we use it directly. If only a probability score is returned, we apply a threshold of 0.5 to determine the final prediction label for each sample.

Data availability

The raw PDID database can be found at http://bit.ly/pdid. The curated PDID dataset for the benchmarking can be accessed at https://purdue0my.sharepoint.com/:f:/g/personal/lin1785_purdue_edu/EpQPBynMPlJDkdQ05Q-Ej5EBNzoiQIhZgFHLo6cbrnYNMw?e=AWkXKE.

Code availability

All of the code used in this paper is available under the MIT License at https://github.com/Purdue-M2/Political_Deepfakes_Benchmark.

Acknowledgements

This work is supported by the U.S. National Science Foundation (NSF) under grant IIS-2434967 and the National Artificial Intelligence Research Resource (NAIRR) Pilot and TACC Lonestar6. The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of NSF and NAIRR Pilot.

Author contributions

S.H. and D.S.S. conceived the project and designed the study as well as the benchmarking framework. G.L. collected and curated datasets, conducted experiments, and contributed to data analysis. C.P.W. assisted with dataset curation and contributed to manuscript drafting. L.L. performed experiments, analyzed data, and co-drafted the manuscript. S.H. and D.S.S. provided major revisions, and supervised the overall research. All authors contributed to the discussion, interpretation of results, and final manuscript preparation.

Competing interests

The authors declare no competing interests.

Declarations

- Funding
  U.S. National Science Foundation (NSF) under grant IIS-2434967.
- Conflict of interest/Competing interests (check journal-specific guidelines for which heading to use)
  The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
- Ethics approval and consent to participate
  Not applicable
- Consent for publication
  Not applicable
- Materials availability
  Not applicable

Supplementary Information

A Image and Video Labeling Instructions

You will evaluate a dataset of images and videos (around 400 items). Each item will be reviewed by 100 labelers. Your task is to use only the visual content and the information presented in or around the item itself to decide whether it is real, fake, or uncertain.

You are not expected to conduct outside research or use specialized detection tools.

A.1 Your Task

For each image or video, select one label:

- **Verified Real** — Appears authentic, with no clear signs of manipulation.
- **Verified Fake** — Appears manipulated, generated, or altered (including deepfakes, cheapfakes, photoshops, memes, cartoons, or cases where context strongly suggests alteration).
- **Uncertain** — Not enough evidence to confidently judge either way.

A.2 What to Consider

A.2.1 Visual/Image Cues

- **Distortions:** unnatural hands, teeth, or ears; mismatched lighting; warped or duplicated backgrounds.
- **Overlays:** added text, watermarks, or meme elements.
- **Cartoonish style:** clearly non-photographic or drawn.

A.2.2 Social Media and Contextual Cues

- **Source clues:** verified accounts may increase credibility; parody accounts may signal unreliability. Neither is conclusive on its own.
- **Timing cues:** inconsistent timestamps or dates may raise suspicion.
- **Format clues:** meme layouts, joke captions, or doctored screenshots are signals.

A.2.3 Real-World Plausibility

- **Event plausibility:** if the depicted situation is highly improbable, it may suggest fakery.
- **Location cues:** unexpected or impossible settings for the subject.
- **Timing plausibility:** if the same person appears in two places simultaneously, it's likely fake.

A.2.4 Self-Disclosure and Labels

Explicit labels (e.g., "this is a parody," "AI-generated") or usernames like `@fun-deepfake-editor` are strong signals but not absolute proof. Treat them as evidence to weigh holistically.

A.3 What NOT to Do

- Do not Google, check news, or verify against external fact-checks.
- Do not use technical deepfake detection tools.
- Do not guess randomly—if uncertain, select **Uncertain**.

A.4 Best Practices

- Use all cues together: visuals, context, plausibility, and disclosures.
- Stay balanced—no single cue determines the label.
- Be consistent—apply the same reasoning across items.
- Multiple annotators will review each item; thoughtful, careful judgments are more valuable than quick guesses.

A.5 Examples

A.5.1 Verified Fake

- TikTok video watermarked "deepfake parody" with clear facial distortions.
- Photoshop of a politician shaking hands with an alien, mismatched lighting.
- Meme screenshot with overlaid joke captions.

A.5.2 Verified Real

- Press photo of a candidate at a rally, with no visible anomalies.
- Screenshot of a televised news broadcast showing plausible visuals and context.

A.5.3 Uncertain

- Medium-quality video where lip-syncing looks off.
- Screenshot from a non-verified account showing plausible but unconfirmed content.
- Official post with slight oddities—conflicting signals.

A.6 Frequently Asked Questions (FAQ)

- **Q1:** What if the content is a meme or cartoon? **A:** Generally treat as Fake since the image has been altered.
- **Q2:** What if the content comes from a parody or joke account? **A:** Treat as a signal but not definitive; weigh it against visual cues.
- **Q3:** What if the content is labeled "AI-generated"? **A:** Treat as strong evidence of fakery but verify visually.
- **Q5:** What if glitches might be due to compression or manipulation? **A:** Choose Uncertain.
- **Q6:** What if the event seems implausible? **A:** If extremely implausible, lean Fake; if merely unusual, consider Uncertain.
- **Q7:** What if the item is a screenshot of another post? **A:** Use both the image and contextual cues to evaluate authenticity.

B Details of Commercial Tools Testing

All commercial tools are tested through their official APIs; if an API is not available, we manually upload images or videos one by one for evaluation. All 9 tools successfully processed all image and video samples, except for two tools that automatically rejected a subset of inputs due to internal filtering or format restrictions.

Incode is a face-oriented deepfake detector primarily designed for identity verification and biometric authentication. It automatically rejects inputs when no face is detected or when the detected region is too small. **BrandWell** accepts only images in .jpg, .png, or .webp formats.

C Details of LVLM Testing

Paid LVLMs were evaluated through their official APIs. However, we occasionally observed that the models failed to return responses in the expected format. To maintain experimental consistency, we did not modify the original prompts (although we attempted several variations to elicit valid score outputs, without success). Additionally, due to company-imposed content restrictions, the models sometimes halted generation when the input was deemed inappropriate.

References

[1] Sarah Kreps, R. Miles McCain, and Miles Brundage. "All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation". In: *Journal of Experimental Political Science* 9.1 (2022/ed), pp. 104–117. ISSN: 2052-2630, 2052-2649. DOI: 10.1017/XPS.2020.37. (Visited on 04/05/2022).

[2] Josh A. Goldstein and Shelby Grossman. *How Disinformation Evolved in 2020*. 2021. URL: https://www.brookings.edu/techstream/how-disinformation-evolved-in-2020/ (visited on 05/26/2022).

[3] World Economic Forum. *These Are the 3 Biggest Emerging Risks the World Is Facing*. 2024. URL: https://www.weforum.org/agenda/2024/01/ai-disinformation-global-risks/ (visited on 06/27/2024).

[4] Bobby Chesney and Danielle Citron. "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security". In: *California Law Review* 107.6 (2019), pp. 1753–1820. URL: https://heinonline.org/HOL/P?h=hein.journals/calr107&i=1789 (visited on 06/01/2021).

[5] Joakim Laine, Matti Minkkinen, and Matti Mäntymäki. "Understanding the Ethics of Generative AI: Established and New Ethical Principles". In: *Communications of the Association for Information Systems* 56.1 (2025). ISSN: 1529-3181. URL: https://aisel.aisnet.org/cais/vol56/iss1/7.

[6] Mark Coeckelbergh. "LLMs, Truth, and Democracy: An Overview of Risks". In: *Science and Engineering Ethics* 31.1 (2025), pp. 1–13. ISSN: 1471-5546. DOI: 10.1007/s11948-025-00529-0. (Visited on 01/27/2025).

[7] Julia Barnett. "The Ethical Implications of Generative Audio Models: A Systematic Literature Review". In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 2023, pp. 146–161. DOI: 10.1145/3600211.3604686. arXiv: 2307.05527 [cs, eess]. (Visited on 04/04/2024).

[8] Laura Weidinger et al. *Ethical and Social Risks of Harm from Language Models*. 2021. DOI: 10.48550/arXiv.2112.04359. arXiv: 2112.04359 [cs]. (Visited on 01/03/2023).

[9] Will Hawkins, Brent Mittelstadt, and Chris Russell. "Deepfakes on Demand: The Rise of Accessible Non-Consensual Deepfake Image Generators: The Rise of Accessible Non-Consensual Deepfake Image Generators". In: *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. Athens Greece: ACM, 2025, pp. 1602–1614. ISBN: 979-8-4007-1482-5. DOI: 10.1145/3715275.3732107. (Visited on 06/27/2025).

[10] Morgan Wack et al. "Generative Propaganda: Evidence of AI's Impact from a State-Backed Disinformation Campaign". In: *PNAS Nexus* 4.4 (2025), pgaf083. ISSN: 2752-6542. DOI: 10.1093/pnasnexus/pgaf083. (Visited on 04/02/2025).

[11] Sophie Loewenstein. "Make America Fake Again?: Banning Deepfakes of Federal Candidates in Political Advertisements Under the First Amendment". In: *Fordham Law Review* 93.1 (2024), p. 273. URL: https://ir.lawnet.fordham.edu/flr/vol93/iss1/7.

[12] Sacha Altay and Fabrizio Gilardi. "People Are Skeptical of Headlines Labeled as AI-generated, Even If True or Human-Made, Because They Assume Full

AI Automation". In: *PNAS Nexus* 3.10 (2024), pgae403. ISSN: 2752-6542. DOI: 10.1093/pnasnexus/pgae403. (Visited on 10/02/2024).

[13] Paulina Trifonova and Sukrit Venkatagiri. "Misinformation, Fraud, and Stereotyping: Towards a Typology of Harm Caused by Deepfakes". In: *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*. San Jose Costa Rica: ACM, 2024, pp. 533–538. ISBN: 979-8-4007-1114-5. DOI: 10.1145/3678884.3685938. (Visited on 11/18/2024).

[14] Charlotte Bird, Eddie Ungless, and Atoosa Kasirzadeh. "Typology of Risks of Generative Text-to-Image Models". In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. Montr\'{e}al QC Canada: ACM, 2023, pp. 396–410. ISBN: 979-8-4007-0231-0. DOI: 10.1145/3600211.3604722. (Visited on 12/04/2024).

[15] Ben Bariach, Bernie Hogan, and Keegan McBride. *Towards a Harms Taxonomy of AI Likeness Generation*. 2024. arXiv: 2407.12030. URL: http://arxiv.org/abs/2407.12030 (visited on 12/04/2024).

[16] Christina P. Walker, Daniel S. Schiff, and Kaylyn Jackson Schiff. "Merging AI Incidents Research with Political Misinformation Research: Introducing the Political Deepfakes Incidents Database". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 2024, pp. 23053–23058. DOI: 10.1609/aaai.v38i21.30349. (Visited on 03/28/2024).

[17] Isabel Linzer. *Adaptation and Innovation: The Civic Space Response to AI-Infused Elections*. Tech. rep. Washington, D.C.: Center for Democracy and Technology, 2025. URL: https://cdt.org/insights/adaptation-and-innovation-the-civic-space-response-to-ai-infused-elections/ (visited on 03/22/2025).

[18] Lluis De Nadal and Peter Jančárik. "Beyond the Deepfake Hype: AI, Democracy, and "the Slovak Case"". In: *Harvard Kennedy School Misinformation Review* (2024). DOI: 10.37016/mr-2020-153. (Visited on 09/02/2024).

[19] Renee Barnes et al. "Disinformation and Deepfakes Played a Part in the US Election. Australia Should Expect the Same - University of the Sunshine Coast, Queensland". In: (2024). URL: https://research.usc.edu.au/esploro/outputs/magazineArticle/Disinformation-and-deepfakes-played-a-part/991084998702621 (visited on 12/16/2024).

[20] Lee Rainie and Jason Husser. *AI & Politics '24*. Tech. rep. Imagining the Digital Future Center, 2024. URL: https://imaginingthedigitalfuture.org/reports-and-publications/ai-politics-24/ (visited on 06/27/2024).

[21] Siddarth Srinivasan. *Detecting AI Fingerprints: A Guide to Watermarking and Beyond*. 2024. URL: https://www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/ (visited on 08/20/2024).

[22] White House. *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. Tech. rep. Washington, D.C.: White House, Office of Science and Technology Policy, 2023. URL: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/ (visited on 11/06/2023).

[23] Emmie Hine and Luciano Floridi. “New Deepfake Regulations in China Are a Tool for Social Stability, but at What Cost?” In: *Nature Machine Intelligence* 4.7 (2022), pp. 608–610. ISSN: 2522-5839. DOI: 10.1038/s42256-022-00513-4. (Visited on 08/01/2022).

[24] Tenneth Fairclough and Lew Blank. *Voters Overwhelmingly Believe in Regulating Deepfakes and the Use of Artificial Intelligence*. 2024. URL: https://www.dataforprogress.org/blog/2024/2/8/voters-overwhelmingly-believe-in-regulating-deepfakes-and-the-use-of-artificial-intelligence (visited on 02/22/2024).

[25] Claire R. Leibowicz and Christian H. Cardona. *From Principles to Practices: Lessons Learned from Applying Partnership on AI's (PAI) Synthetic Media Framework to 11 Use Cases*. 2024. arXiv: 2407.13025. URL: http://arxiv.org/abs/2407.13025 (visited on 11/04/2024).

[26] Partnership on AI. *PAI's Responsible Practices for Synthetic Media*. Tech. rep. Partnership on AI, 2023. URL: https://syntheticmedia.partnershiponai.org/ (visited on 06/06/2023).

[27] Ramak Molavi Vasse'i and Gabriel Udoh. *In Transparency We Trust? Evaluating the Effectiveness of Watermarking and Labeling AI-Generated Content*. Tech. rep. Mozilla Foundation, 2024. URL: https://foundation.mozilla.org/en/research/library/in-transparency-we-trust/research-report/ (visited on 03/13/2024).

[28] Justyna Lisinska. *Why Watermarking Text Fails to Stop Misinformation and Plagiarism*. 2024. URL: https://datainnovation.org/2024/09/why-watermarking-text-fails-to-stop-misinformation-and-plagiarism/ (visited on 09/23/2024).

[29] David Evan Harris and Lawrence Norden. “Meta's AI Watermarking Plan Is Flimsy, At Best”. In: *IEEE Spectrum* (). URL: https://spectrum.ieee.org/meta-ai-watermarks (visited on 03/05/2024).

[30] François Chollet. “Xception: Deep Learning with Depthwise Separable Convolutions”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 1251–1258.

[31] Li Lin et al. “Preserving Fairness Generalization in Deepfake Detection”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 16815–16825.

[32] Yuyang Qian et al. “Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues”. In: *European Conference on Computer Vision*. Springer, 2020, pp. 86–103.

[33] *AMERICA'S AI ACTION PLAN*. https://www.whitehouse.gov/articles/2025/07/white-house-unveils-americas-ai-action-plan/. 2025.

[34] National Institute of Standards and Technology. *Frontier Research on Mitigating Risks from Synthetic Content: A Call to Action*. 2024. URL: https://www.nist.gov/aisi/frontier-research-mitigating-risks-synthetic-content-call-action.

[35] National Institute of Standards and Technology. *Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content*

Transparency. NIST AI 100-4. U.S. Department of Commerce, National Institute of Standards and Technology, 2024. URL: https://airc.nist.gov/docs/NIST.AI.100-4.SyntheticContent.ipd.pdf (visited on 06/27/2025).

[36] Brian Dolhansky et al. "The Deepfake Detection Challenge (Dfdc) Dataset". In: *arXiv preprint arXiv:2006.07397* (2020). arXiv: 2006.07397.

[37] "The Competition of Fairness in AI Face Detection". In: *https://sites.google.com/view/aifacedetection/home*.

[38] Andreas Rossler et al. "Faceforensics++: Learning to Detect Manipulated Facial Images". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 1–11.

[39] Yuezun Li et al. "Celeb-Df: A Large-Scale Challenging Dataset for Deepfake Forensics". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 3207–3216.

[40] Luisa Verdoliva. "Media Forensics and DeepFakes: An Overview". In: *IEEE Journal of Selected Topics in Signal Processing* 14.5 (2020), pp. 910–932. ISSN: 1932-4553, 1941-0484. DOI: 10.1109/JSTSP.2020.3002101. (Visited on 09/15/2025).

[41] Ruben Tolosana et al. "Deepfakes and beyond: A Survey of Face Manipulation and Fake Detection". In: *Information Fusion* 64 (2020), pp. 131–148. ISSN: 15662535. DOI: 10.1016/j.inffus.2020.06.014. (Visited on 09/15/2025).

[42] Michael Hameleers, Toni van der Meer, and Rens Vliegenthart. "How Persuasive Are Political Cheapfakes Disseminated via Social Media? The Effects of out-of-Context Visual Disinformation on Message Credibility and Issue Agreement". In: *Information, Communication & Society* 0.0 (2024), pp. 1–18. ISSN: 1369-118X. DOI: 10.1080/1369118X.2024.2388079. (Visited on 08/20/2024).

[43] Michael Hameleers. "Cheap Versus Deep Manipulation: The Effects of Cheapfakes Versus Deepfakes in a Political Setting". In: *International Journal of Public Opinion Research* 36.1 (2024), edae004. ISSN: 1471-6909. DOI: 10.1093/ijpor/edae004. (Visited on 03/10/2024).

[44] Demetrios Ioannou. "Deepfakes, Cheapfakes, and Twitter Censorship Mar Turkey's Elections". In: *Wired* (). ISSN: 1059-1028. URL: https://www.wired.com/story/deepfakes-cheapfakes-and-twitter-censorship-mar-turkeys-elections/ (visited on 05/30/2023).

[45] Coalition for Content Provenance and Authenticity (C2PA). *C2PA Technical Specification 2.2*. https://c2pa.org/specifications/specifications/2.2/index.html. Accessed: 2025-10-09. 2025.

[46] Nuria Alina Chandra et al. *Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024*. 2025. DOI: 10.48550/arXiv.2503.02857. arXiv: 2503.02857 [cs]. (Visited on 06/24/2025).

[47] Florinel-Alin Croitoru et al. *MAVOS-DD: Multilingual Audio-Video Open-Set Deepfake Detection Benchmark*. 2025. DOI: 10.48550/arXiv.2505.11109. arXiv: 2505.11109 [cs]. (Visited on 06/24/2025).

[48] Yuhang Lu and Touradj Ebrahimi. "Impact of Video Processing Operations in Deepfake Detection". In: *2023 24th International Conference on Digital Signal*

Processing (DSP). Rhodes (Rodos), Greece: IEEE, 2023, pp. 1–5. ISBN: 979-8-3503-3959-8. DOI: 10.1109/DSP58604.2023.10167906. (Visited on 09/15/2025).

[49] Liming Jiang et al. "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*. 2020, pp. 2889–2898.

[50] Christopher Teo, Milad Abdollahzadeh, and Ngai-Man Man Cheung. "On Measuring Fairness in Generative Models". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 10644–10656.

[51] Kartik Narayan et al. "Df-Platter: Multi-face Heterogeneous Deepfake Dataset". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 9739–9748.

[52] Zhiyuan Yan et al. "DF40: Toward next-Generation Deepfake Detection". In: *arXiv preprint arXiv:2406.13495* (2024). arXiv: 2406.13495.

[53] Xuannan Liu et al. "Mmfakebench: A mixed-source multimodal misinformation detection benchmark for lvlms". In: *ICLR*. 2024.

[54] Nuria Alina Chandra et al. "Deepfake-eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024". In: *arXiv preprint arXiv:2503.02857* (2025).

[55] Li Lin et al. "AI-face: A Million-Scale Demographically Annotated AI-generated Face Dataset and Fairness Benchmark". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2025.

[56] "Deepfakes in the 2024 US Presidential Election". In: *https://farid.berkeley.edu/deepfakes2024election/*.

[57] *Sora 2*. https://openai.com/index/sora-2/. 2025.

[58] Rasmus Rothe, Radu Timofte, and Luc Van Gool. "Dex: Deep Expectation of Apparent Age from a Single Image". In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2015, pp. 10–15.

[59] Tero Karras, Samuli Laine, and Timo Aila. "A Style-Based Generator Architecture for Generative Adversarial Networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 4401–4410.

[60] Loc Trinh and Yan Liu. "An Examination of Fairness of Ai Models for Deepfake Detection". In: *arXiv preprint arXiv:2105.00558* (2021). arXiv: 2105.00558.

[61] Zhiyuan Yan et al. "DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection". In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., 2023, pp. 4534–4565. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/0e735e4b4f07de483cbe250130992726-Paper-Datasets_and_Benchmarks.pdf.

[62] Chuqiao Li et al. "A Continual Deepfake Detection Benchmark: Dataset, Methods, and Essentials". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023, pp. 1339–1349.

[63] Jingyi Deng et al. "Towards Benchmarking and Evaluating Deepfake Detection". In: *IEEE Transactions on Dependable and Secure Computing* (2024). DOI: 10.1109/TDSC.2024.3369711.

[64] Yichen Shi et al. “Shield: An evaluation benchmark for face spoofing and forgery detection with multimodal large language models”. In: *Visual Intelligence* 3.1 (2025), p. 9.
[65] Simiao Ren et al. “Can Multi-modal (reasoning) LLMs work as deepfake detectors?” In: *arXiv preprint arXiv:2503.20084* (2025).
[66] Mingxing Tan and Quoc V. Le. “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks”. In: *International conference on machine learning* (2019), pp. 6105–6114.
[67] Alexey Dosovitskiy et al. “An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *arXiv preprint arXiv:2010.11929* (2020). arXiv: 2010.11929.
[68] Honggu Liu et al. “Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 772–781.
[69] Yuchen Luo et al. “Generalizing Face Forgery Detection with High-Frequency Features”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 16317–16326.
[70] Zhiyuan Yan et al. “Ucf: Uncovering Common Features for Generalizable Deepfake Detection”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 22412–22423.
[71] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. “Towards Universal Fake Image Detectors That Generalize across Generative Models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 24480–24489.
[72] Yunsheng Ni et al. “Core: Consistent Representation Learning for Face Forgery Detection”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 12–21.
[73] Yan Ju et al. “Improving Fairness in Deepfake Detection”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024, pp. 4655–4665.
[74] Sheng-Yu Wang et al. “CNN-generated Images Are Surprisingly Easy to Spot... for Now”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 8695–8704.
[75] Arslan Basharat. *Kitware Generated Image Detector*. 2021. URL: https://github.com/Kitware/generated-image-detection (visited on 03/24/2025).
[76] Yaser Yacoob. *Ganattribution*. 2024. URL: https://gitlab.umiacs.umd.edu/yaser/ganattribution (visited on 03/24/2025).
[77] DARPA. *SemaFor: Semantic Forensics*. https://www.darpa.mil/research/programs/semantic-forensics. Accessed: 2025-10-09. 2025.
[78] *Incode – Identity Verification and Biometric Authentication Platform*. https://incode.com/. Incode Technologies, 2025.
[79] *Is It AI?* https://isitai.com/. 2025.
[80] *Winston AI – The Most Trusted AI Detector*. https://gowinston.ai/. 2025.
[81] *Advanced AI Image Detector*. https://brandwell.ai/ai-image-detector/. BrandWell, 2025.

[82] *AI or Not*. https://www.aiornot.com/dashboard/home. 2025.
[83] *Illuminarty — AI Generated Content Detection*. https://illuminarty.ai/en/. Illuminarty, 2025.
[84] *Reality Defender — Deepfake Detection Platform*. https://www.realitydefender.com/. Reality Defender, 2025.
[85] *Hive Moderation*. https://hivemoderation.com/. Hive Moderation, 2025.
[86] *Deepfake Detector — AI-Powered Deepfake Detection Tool*. https://deepfakedetector.ai/. 2025.
[87] OpenAI. *ChatGPT (GPT-5)*. https://chat.openai.com/. Large language model. 2025.
[88] Jinze Bai et al. "Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond". In: *arXiv preprint arXiv:2308.12966* (2023).
[89] Anthropic. *Claude Sonnet 4.5*. https://www.anthropic.com/claude. Large language model by Anthropic. Accessed: 2025-10-07. 2025.
[90] Gheorghe Comanici et al. "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities". In: *arXiv preprint arXiv:2507.06261* (2025).
[91] Bin Lin et al. "Video-Llava: Learning United Visual Representation by Alignment before Projection". In: *arXiv preprint arXiv:2311.10122* (2023). arXiv: 2311.10122.
[92] Haotian Liu et al. "Visual Instruction Tuning". In: *NeurIPS*. 2023.
[93] XTuner Contributors. *XTuner: A Toolkit for Efficiently Fine-Tuning LLM*. https://github.com/InternLM/xtuner. Accessed: 2023. 2023.
[94] Haoyu Lu et al. "Deepseek-vl: towards real-world vision-language understanding". In: *arXiv preprint arXiv:2403.05525* (2024).
[95] Weihan Wang et al. "Cogvlm: Visual expert for pretrained language models". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 121475–121499.
[96] Zhang Li et al. "Monkey: Image Resolution and Text Label Are Important Things for Large Multi-modal Models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, pp. 26763–26773.
[97] Yunfan Gao et al. "Retrieval-augmented generation for large language models: A survey". In: *arXiv preprint arXiv:2312.10997* 2.1 (2023).
[98] *SemaFor Content Assessment*. https://semafor.dsri.org/. 2025.
[99] Li Lin et al. "Detecting multimedia generated by large ai models: A survey". In: *arXiv preprint arXiv:2402.00045* (2024).
[100] Christina P. Walker, Daniel S. Schiff, and Kaylyn Jackson Schiff. "Merging AI Incidents Research with Political Misinformation Research: Introducing the Political Deepfakes Incidents Database". In: *[Journal information missing]* (2024).
[101] *Snopes*. https://www.snopes.com/. 2025.
[102] *AFP Face Check*. https://factcheck.afp.com/. 2025.

[103] G. Bradski. "The OpenCV Library". In: *Dr. Dobb's Journal of Software Tools* (2000).

[104] Chende Zheng et al. "Breaking Semantic Artifacts for Generalized AI-generated Image Detection". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 59570–59596.

[105] Riccardo Corvi et al. "Intriguing Properties of Synthetic Images: From Generative Adversarial Networks to Diffusion Models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 973–982.

[106] Google Research. *Contributing Data to Deepfake Detection Research*. 2019. URL: https://research.google/blog/contributing-data-to-deepfake-detection-research/.

[107] Tero Karras et al. "Progressive growing of gans for improved quality, stability, and variation". In: *arXiv preprint arXiv:1710.10196* (2017).

[108] Jiankang Deng et al. "Retinaface: Single-stage Dense Face Localisation in the Wild". In: *arXiv preprint arXiv:1905.00641* (2019). arXiv: 1905.00641.

[109] Sefik Ilkin Serengil and Alper Ozpinar. "LightFace: A Hybrid Deep Face Recognition Framework". In: *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, 2020, pp. 23–27. DOI: 10.1109/ASYU50717.2020.9259802.