Seeing Is No Longer Believing: A Security and Forensic Evaluation of Generative Video Platforms

Becky Qiao Ling Lin, Yuying Li, Fengjun Li, Bo Luo
EECS/I2S, The University of Kansas, Lawrence, KS, USA
{beckylin, yuyingli, fli, bluo}@ku.edu

***Abstract*—With the increasing capability and popularity of AI video generators, security and privacy concerns arise. In this paper, we present an external, black-box security and forensics evaluation of six state-of-the-art generative video platforms. We generate and analyze 1,100 videos across text-to-video (T2V) and image-to-video (I2V) settings, covering identity, speech, and forensic detection evaluations. We examine user-facing safeguards, including restrictions on human-image uploads, watermarking, and platform-dependent capture protections. To assess misuse-enabling fidelity and forensic robustness, we evaluate facial identity consistency, voice similarity, lip-sync accuracy, and detectability using deepfake detectors, general AI-image classifiers, and video-level forensic tools. Our results show that current generative video systems can produce highly consistent facial and vocal identities, increasing the plausibility of synthetic personas, while no evaluated forensic detector consistently and reliably identifies generated content across platforms and generation settings. A user study further shows that participants struggle to reliably distinguish generated videos from authentic content. With this empirical/measurement study, we demonstrate that modern AI video platforms are capable of generating high-fidelity videos that widen the gap between synthetic media realism and the forensic and governance mechanisms intended to detect/control misuse.**

1. Introduction

Recent generative video systems can now preserve facial identity, synthesize synchronized speech, and produce temporally coherent human behavior at a level approaching authentic media. Unlike earlier deepfake pipelines that required source videos, face-swapping tools, and technical expertise, modern generative video platforms allow ordinary users to create realistic synthetic personas from a single uploaded image or text prompt. Recent systems such as OpenAI's Sora 2 [1] and Google's Veo 3.1 [2] now support high-fidelity video synthesis with synchronized audio-visual speech generation. Sora 2 further introduces personalized cameo-based identity insertion, expanding both creative control and the potential attack surface. As identity realism improves, these systems create new attack surfaces involving impersonation, non-consensual identity reuse, reputational manipulation, misinformation, and evidentiary deception. However, the governance and forensic mechanisms surrounding these systems have not matured at the same pace. Current platforms expose inconsistent protections for identity consent, copyrighted content, provenance signaling, and media export. At the same time, existing forensic tools were largely developed for autoencoder-based and GAN-based image and face manipulation[3], [4], raising open questions about whether they generalize to modern multimodal diffusion-generated videos. Importantly, these risks extend beyond direct impersonation. Even videos that do not depict a specific real individual can fabricate events, simulate witnesses, generate misleading evidence, and undermine trust in authentic visual media. An adversary can generate videos of a target individual using only a publicly available photograph, fabricate realistic speaking personas for disinformation campaigns, or selectively choose generators whose outputs evade specific detection pipelines. Each of these actions is possible through standard platform interfaces without specialized expertise. Understanding the extent to which deployed platforms enable or mitigate these threats requires a systematic external evaluation. However, prior work on generative video evaluation and media forensics has largely proceeded in isolation, with the former targeting visual quality benchmarks and the latter focusing on GAN-based imagery and frame-level artifacts. No external evaluation has jointly examined the identity synthesis capabilities, governance mechanisms, and forensic attack surfaces of deployed generative video platforms.

Beyond generation quality, modern video generation platforms expose user-facing mechanisms with important implications for safety, privacy, and attribution. These include identity permissions, restrictions on human-image uploads, watermarking, and platform-specific capture protections. These mechanisms govern how identity, consent, provenance, and content export are handled in practice. However, their effectiveness, consistency, and limitations remain largely unexplored through independent external auditing. On the other hand, major online platforms increasingly require disclosure of AI-generated or heavily modified media, reflecting growing recognition that synthetic video can mislead viewers when provenance is unclear [5], [6]. However, such disclosure mechanisms remain platform-dependent, inconsistently enforced, and often ineffective once content is redistributed across platforms or stripped of provenance metadata.

In this paper, we aim to highlight the growing impor-

tance of robust forensic and governance mechanisms as generative video realism continues to improve. We conduct an external, black-box evaluation of six state-of-the-art generative video platforms: Sora 2 (OpenAI) [1], Veo 3.1 (Google) [2], Ray 2 (Luma AI) [7], Kling 2.6 (Kuaishou) [8], Gen 4.5 (Runway) [9], and Doubao Seedance 1.5 Pro (ByteDance) [10]. We assess both their text-to-video (T2V) and image-to-video (I2V) capabilities under realistic end-user conditions. Instead of treating these systems as pure performance benchmarks, we study them as deployed artifacts to measure observable safeguards, identity realism as a potential risk surface, and the effectiveness of current forensic detection pipelines. This approach provides empirical evidence of how state-of-the-art commercial systems behave under realistic usage conditions accessible to *ordinary users*.

In this paper, we provide an early empirical security baseline and offer broader insights for auditing state-of-the-art generative video systems. In particular, through the measurement study, we aim to answer the following research questions. **(RQ1)** How well do state-of-the-art generative video platforms, such as Sora, Veo, Ray, Kling, Doubao, and Runway Gen, enforce user-level safety and privacy safeguards such as human-image upload restrictions, identity permissions, and watermarking? **(RQ2)** How consistently do these platforms preserve facial identity in generated human videos under generic prompting conditions, and, where supported, how consistent are generated vocal identities? **(RQ3)** For platforms that support synchronized speech generation (Sora 2, Veo 3.1, and Doubao), how accurately do they produce temporally aligned audio-visual outputs with realistic lip synchronization? **(RQ4)** How do deepfake, AI-image, and AI-video detectors identify content generated across these platforms? **(RQ5)** How effectively can human users distinguish AI-generated videos from real videos, and how does synchronized audio influence perceived realism and detection confidence?

This paper makes the following key contributions:

1. We present an external, black-box audit showing that deployed generative video platforms now expose identity-realistic synthesis capabilities faster than their consent, provenance, and forensic defenses have matured.

2. We identify security gaps across platforms, including modality-dependent copyright bypass (where image-conditioned inputs circumvent prompt-level filters), unrestricted human-image uploads without consent controls on four of six platforms, and weak provenance protections with no visible API watermarks and fragile embedded metadata.

3. We show that generated videos preserve strong facial identity consistency and, where applicable, maintain voice similarity and lip-sync quality comparable to natural speech videos, thereby enhancing the realism and reliability of synthetic personas.

4. We demonstrate distinct performance patterns across forensic detection paradigms: deepfake image detectors fail broadly across all six generators, general AI-image detectors are substantially more effective but exhibit generator-specific blind spots, and video-level detectors provide strong coverage but exhibit substantial failures on image-conditioned videos.

5. We conduct a user study to measure participants' ability to distinguish AI-generated videos from authentic content, evaluate perceived realism across platforms, and analyze the effect of synthesized speech audio in models that support it.

2. Related Work

Generative Video and Evaluation. Text-to-video and image-to-video generation evolved rapidly from autoregressive transformer approaches to large-scale diffusion models. Early approaches, such as CogVideo [11] and Phenaki [12], demonstrated that text-to-image pipelines could be extended along the temporal axis, but were limited to short, low-resolution clips and often exhibited temporal inconsistencies. Diffusion-based methods, e.g., ModelScope [13], improved visual quality and temporal coherence through spatio-temporal denoising, while more recent modular systems, e.g., Mora [14], extended generation to longer and more diverse video sequences. Recent commercial systems further support multiple conditioning modalities, including prompt-based video synthesis and image-conditioned identity or motion generation [2], [8].

Despite these advances, prior work primarily evaluates T2V models using benchmarks focused on short, single-shot clips and visual fidelity metrics such as FVD or CLIPSIM [15], [16]. While these metrics capture perceptual quality, they do not measure properties that become critical in more advanced systems, including identity consistency across frames, audio-visual synchronization, and the generation of realistic human subjects. Recent work also explored safety assessment for T2V generation through harmful-content and jailbreak-style evaluation [17]. However, existing evaluation protocols still provide limited insight into the security, privacy, and forensic implications of deployed generative video systems, particularly those involving identity realism, consent governance, and detectability.

Deepfake and AI-Generated Media Detection. A substantial body of work studies the detection of manipulated faces and GAN-synthesized imagery. Early approaches rely on low-level visual artifacts, facial landmark inconsistencies, or physiological signals such as blinking and pulse [18], [19], while more recent methods adopt convolutional and transformer-based architectures trained on large-scale deepfake datasets. Recent surveys [3], [4] highlight a key challenge in this area: detection models often exhibit limited generalization across datasets and generation techniques. In particular, detectors trained on specific manipulation pipelines or GAN-based synthesis frequently experience performance degradation when applied to unseen generators.

More recent work on AI-generated image detection has begun to examine diffusion-based models, focusing on identifying global statistical or spectral artifacts rather than localized manipulation cues [20]. Wang et al. [21] proposed DIRE, which detects diffusion-generated images by measuring the reconstruction error when an input is passed through a pre-trained diffusion model, exploiting the observation

that generated images are reconstructed more faithfully than real ones. Earlier work by Wang et al. [22] showed that a classifier trained on a single GAN could generalize to unseen generators, establishing a widely used baseline for AI-generated image detection. However, these studies are primarily conducted on still images and do not account for temporal dynamics, compression effects, or audio-visual interactions present in modern text-to-video systems. As a result, the effectiveness of existing detection approaches on diffusion-based video remains insufficiently understood.

Video Forensics and Multimodal Detection. Video-level forensic analysis extends beyond spatial artifacts by incorporating temporal consistency, motion coherence, and cross-frame dependencies. Prior work has shown that modeling temporal dynamics improves robustness compared to frame-level detection, particularly for identifying subtle inconsistencies that are not visible in individual images [23], [3]. Forensic-oriented augmentation has been proposed recently to improve the generalizability of AI video detection [24]. In addition to temporal cues, recent work also explored multimodal signals, including audio-visual alignment and cross-modal consistency, as indicators of manipulated or synthesized content [25], [26]. For example, lip-sync accuracy and audio-visual embedding consistency have been used to detect mismatches between speech and facial motion. These approaches leverage the fact that generating temporally and semantically aligned audio and visual signals remains challenging for many generative systems.

Summary and Research Gap. Existing work primarily addresses visual quality benchmarks, detection in controlled settings, or individual safety mechanisms. Recent black-box audits have examined safety filters of generative image platforms [27], but do not jointly investigate identity governance, multimodal realism, provenance, and forensic detectability across deployed video systems. Our study addresses these dimensions in a cross-platform, external audit under ordinary-user access conditions.

3. Measurement Methodology

3.1. Design and Dataset

We conduct an external, black-box measurement of six deployed generative video platforms: Sora 2, Veo 3.1, Ray 2, Kling 2.6, Runway Gen-4.5, and Doubao Seedance 1.5 Pro. We evaluate each platform based on the capabilities exposed to ordinary users through public interfaces. Because the evaluated platforms differ in supported modalities, identity mechanisms, audio generation, watermarking behavior, and export controls, we organize the study into a shared cross-platform core and capability-specific modules.

- The shared core applies to all platforms and measures: (1) *user-facing safety and privacy safeguards*, including upload restrictions, watermarking, and provenance signals; (2) *visual realism and facial identity consistency* in generated human videos; and (3) *detectability* using frame-level and video-level forensic tools.
- Capability-specific modules, including image-conditioned identity preservation, lip-sync accuracy, voice similarity, and cameo-based consent governance, are applied only when the relevant functionality is publicly exposed. In particular, Sora 2, Veo 3.1, and Doubao are capable of natively generating synchronized speech with visible speaking behavior, hence, we further evaluate their lip-sync accuracy and audio-visual forensic risks. Finally, voice similarity is evaluated only for Sora 2, as it is the sole platform that supports voice cloning through its cameo mechanism. When a platform does not expose a given capability, we mark the measurement as not applicable, not as a failure.

Dataset. The generated dataset consists of 1,100 videos produced across six models. All videos use default model settings. For each platform, 100 videos were generated using text prompts only. Half of them use the prompt "A person talking" and half "A person not talking". 100 additional videos were generated using text and seed images by each platform, except Doubao. For Sora 2, identity-conditioned generations use its cameo-based personalization mechanism (public cameos). For Veo, Ray, Kling, and Gen, we used images from the FFHQ dataset [28] as seeds (one seed image for each video). Since Doubao does not support human image-to-video generation, we did not use image-based generation in it. We additionally generated 150 videos using complex text prompts to evaluate robustness beyond the primary dataset. We also collected 100 authentic videos to measure detector false-positive rates.

We intentionally used simple prompts to avoid prompt-dependent bias and to provide a conservative forensic benchmark. In particular, if we use complex prompts in image-based generation (e.g., ask for specific facial expressions and gestures), the identity in the generated video would significantly drift away from the uploaded face image. On the other hand, the "not talking" prompt does not produce static scenes. Generators interpret it with varied body motion, gestures, facial expressions, and background activity, introducing natural diversity in pose and movement without explicit prompt engineering.

Our measurements were conducted while all platforms were publicly accessible. Sora 2's web and mobile interfaces were later discontinued on April 26, 2026 [29]; our results should therefore be interpreted as a retrospective audit of the platform as deployed during the measurement period.

3.2. Threat Model

We formalize a threat model to capture how an ordinary-user adversary can exploit AI video generation capabilities across deployed platforms and evade safeguards and forensic tools. Unlike traditional deepfake threat models that assume access to model internals, we consider an adversary operating entirely through public interfaces and APIs.

Adversary profile and goal. The adversary holds standard, non-privileged accounts on commercial AI video platforms, with no access to model weights, training data, or internal

moderation pipelines. The adversary does not compromise, bypass, or manipulate the platform interfaces or their regular operations. For example, they do not exploit API vulnerabilities, tamper with moderation workflows, or attack/compromise platform-side provenance mechanisms beyond using them as an ordinary user. This threat model represents typical novice users with malicious intent, as much of the misuse of commercial AI video platforms came from users who rely on readily available platform capabilities instead of sophisticated attacks/exploits. They do not have the resources or knowledge to own or train a local model for video generation. They have obtained seed image(s) for the target identity (the victim). Their goal is to produce synthetic videos that (1) depict a target identity or fabricated persona with sufficient realism to be perceived as authentic, (2) evade automated forensic detection, and (3) carry no attributable provenance signal.

Attack workflow. We decompose the adversary's workflow into four stages, each mapping to an empirical component of our evaluation. *Stage 1: Safeguard*: the adversary may select the platform with weaker input-level safeguards, exploiting modality-dependent enforcement gaps (e.g., a platform that blocks a name-based prompt but permits image-conditioned generation of the same individual). *Stage 2: Identity realism*: the adversary selects the generation modality that maximizes identity consistency, choosing between image-conditioned and prompt-based generation, and selecting platforms supporting voice cloning or synchronized speech where needed. *Stage 3: Forensic evasion and provenance stripping*: the adversary selects the generator whose outputs are least detectable by the target platform's forensic pipeline, matching generators to detector blind spots. The adversary also eliminates attribution signals through API access, cropping, or exploiting inaccessible metadata. *Stage 4: Human deception*: the adversary distributes the content to human viewers, relying on multimodal realism to avoid suspicion.

3.3. Evaluation Framework

We describe the metrics, tools, and procedures for the four evaluation dimensions: governance auditing, identity realism, forensic detectability, and human perception.

1. Safety, privacy, and access-control audit. Following established frameworks in algorithmic auditing [30], [31], we systematically tested platform behavior under *six* aspects: prompt-only generation, human-image uploads, synchronized speech generation, cameo insertion, privacy settings, and access-control functions.

Sora 2 was evaluated through its public web and mobile interfaces, while the remaining models were evaluated through their UIs and public APIs. For each platform, we documented observable safeguards related to content restrictions, identity handling, privacy controls, provenance indicators, and media-generation permissions. We did not use third-party generation services or wrapper tools.

2. Multimodal quality (identity realism). Multimodal quality of the AI videos is evaluated in three aspects: (A) *Face recognition.* For each identity, we build a reference gallery and extract embeddings with three standard face-recognition models: DeepFace, FaceNet, and InsightFace [32], [33], [34]. A frame is counted as PASS when the model verifies that the face matches the gallery identity; we report the Detection Pass Rate (DPR) as the percentage of PASS frames per video. DPR measures machine-verifiable identity consistency rather than human-perceived similarity. All models use default thresholds with no tuning or calibration. (B) *Voice similarity.* We evaluate vocal identity preservation using four complementary representations: Log-Mel and MFCC features [35] for low-level acoustics, wav2vec2-base embeddings [36] for phonetic structure, and Resemblyzer embeddings [37] for speaker identity. For each identity, we compute a real–real baseline from two authentic recordings and an AI–real score from generated clips, then perform paired t-tests with Holm correction for multiple comparisons. We evaluate voice similarity only for Sora 2, as it is the sole platform supporting voice cloning. (C) *Lip-sync quality.* We evaluate lip–audio synchronization using the LSE-D and LSE-C metrics from SyncNet [38], following the standard Wav2Lip evaluation pipeline. We compare generated-video scores against published benchmarks on news and YouTube speech datasets rather than our own human reference videos, which are drawn from online creators and do not form a controlled speech corpus.

3. Forensic detectors. We evaluate detectability using ten forensic systems spanning three paradigms. *Deepfake-image detectors* include the commercial Sightengine Deepfake Detector [39] and three ViT-based models: Ashish [40], Dima806 [41], and Wvolf [42]. *AI-image detectors* include Sightengine AI Image Detection [43], Ateeqq [44], Hemg [45], and UnivFD [20]. *AI-video detectors* include Sightengine Video Detection [46] and AI or Not [47]. Video-level detectors analyze each clip directly, whereas image-level detectors independently analyze every extracted frame.

4. Human perception study. We recruited 146 participants (all $\geq$18 years old; IRB-approved), primarily from a U.S. university, a European university, and a Japanese research institute, to distinguish AI-generated videos from authentic ones. The study consisted of two parts. In Part 1, each participant viewed seven muted videos (four AI-generated and three authentic), randomly sampled from pools spanning the six generators and authentic sources, respectively. In Part 2, each participant viewed three audio-enabled videos (two randomly selected from Sora 2, Veo 3.1, or Doubao, and one authentic). Videos were randomly shuffled without repetition for each participant. Because videos were randomly sampled, the number of evaluations per generator varies across platforms in both parts.

Authentic videos were extracted from YouTube and TikTok, which were timestamped significantly earlier than AI video platforms were released. Participants could replay video clips as needed but receive no correctness feedback before completing the survey. For each decision, participants reported confidence on a 5-point Likert scale. We analyze detection accuracy, confidence calibration, and the effect of audio on perceived realism.

4. Empirical Findings

4.1. Governance and Access-Control Gaps

Human-Identity and Copyright Policy. We first evaluate platform policies/practices related to human-identity and copyrighted content in both prompt-based and image-conditioned generation. The results are summarized in Table 1. We observe substantial differences in how platforms enforce these restrictions depending on the input modality: (1) Ray 2, Gen 4.5, and Kling 2.6 all allow video generation from uploaded images of intellectual property (IP) characters (e.g., Mickey Mouse), public figures (e.g., Michael Jordan), and generic people, while Veo 3.1 blocks video generation from images of IP characters or public figures. Doubao takes images of non-human IP characters to generate videos, but rejects any human images. Sora 2 used a complex Cameo mechanism (to be examined later). (2) Doubao and Kling 2.6 accept text prompts to generate videos from IP characters and public figures, when their names are explicitly referenced in the prompt (e.g., "a video of Taylor Swift talking"). Veo 3.1, Sora 2, and Gen 4.5 all reject such prompts (with error messages). Ray 2 denies text prompts to generate videos from IP characters, but accepts prompts to generate videos from public figures. (3) In general, Veo and Sora appear to be the most restrictive, while Kling is the most permissive, which may raise significant privacy and ethical concerns. (4) The publicly announced policies, when available, appear to be inconsistent with the actual practices of the generation platforms.

Watermarks. For Sora accessed through the app and on the website, all generated videos include a visible digital watermark, both in draft previews and in the final downloaded versions. For all other models, videos generated through the web interfaces do contain visible watermarks. However, they could be easily evaded with API access, as videos generated through API do not contain visible watermarks.

To assess the consistency of Sora 2 watermarks, we conducted multiple export trials for different videos. The watermark consistently appeared in the same regions of the frame, making it potentially removable or concealable through simple cropping or by overlaying stickers using third-party applications. If taking screenshots, the watermarks could be partially obscured using freely available online editing tools, as shown in Appendix Figure 2. OpenAI also indicates that each video is embedded with C2PA metadata [1]; however, this information is generally not readily visible to typical end users in standard viewing workflows.

We further examined embedded C2PA provenance using online verification tools and local Python-based inspection. C2PA credentials were detectable in all original videos tested from Sora 2, Veo 3.1, and Runway Gen 4.5, although Runway's credentials required local inspection. However, after simple FFmpeg trimming or lossless remuxing, the credentials were no longer detectable in any tested videos. These findings suggest that embedded provenance metadata, while effective for identifying unmodified outputs, may not survive common video-processing operations.

TABLE 1: Policies and practices for IP, public-figure, and generic-person requests. IP: IP character; PF: public figure; GP: generic person; ×: blocked; ✓: allowed.

Model	Image+Text Prompt			Text Prompt		
	IP	PF	GP	IP	PF	GP
Veo 3.1	×	×	✓	×	×	✓
Sora 2	×	×	×	×	×	✓
Ray 2	✓	✓	✓	×	✓	✓
Doubao	✓	×	×	✓	✓	✓
Gen 4.5	✓	✓	✓	×	×	✓
Kling 2.6	✓	✓	✓	✓	✓	✓

Sora 2-Specific Governance Mechanisms. Unlike the other evaluated platforms that serve primarily as generation tools, Sora 2 operated as an integrated creative and social platform. As such, it implemented a complex set of user-facing identity governance and content protection mechanisms through its public web and mobile interfaces, including cameo-based permission management, selective identity sharing, content deletion controls, visible watermarking, and mobile-level screen recording restrictions.

We observed that cameo creators can restrict who is allowed to generate videos using their likeness and may later revoke access or delete generated content. However, these controls are not fully retroactive. Videos published prior to a permission change remain accessible until manually removed, limiting dynamic consent and creating a temporal window during which previously authorized content may continue to circulate. We also observed that published videos cannot be hidden after release, further reducing post-publication control over generated media.

Sora 2 also implements platform-dependent capture protections. On mobile devices, attempting screen recording causes the display to black out during playback, preventing direct visual capture of generated content. However, screenshots and audio capture remain possible, leaving partial extraction pathways available to users. In contrast, the web interface does not block recording and instead relies primarily on persistent visible watermarking.

These mechanisms demonstrate a stronger emphasis on interface-level governance and provenance signaling than is observable through the API-based workflows of the other evaluated platforms, while also illustrating the practical limitations of retroactive consent enforcement and client-side capture protections. Additional platform-specific observations are summarized in Appendix Table 10.

Security implications. These gaps directly enable Stages 1 and 3 of the adversarial workflow in Section 3.2. Modality-dependent copyright enforcement allows bypass through image-conditioned inputs, unrestricted human-image uploads enable non-consensual identity reuse, and absent or fragile provenance signals leave generated content without user-visible attribution once exported.

4.2. Identity Realism and Multimodal Consistency

To assess how consistently the models preserve identity, voice characteristics, and speech alignment across diverse conditions, we evaluated their videos using face-recognition, voice-similarity, and lip-sync metrics.

Face recognition. Table 2 reports the DPR for each image-based generator, condition, and face-recognition model. We excluded three videos generated by Ray 2, as they contained fewer than three frames with a detectable face.

Results are highly consistent across the three independent recognition models. Most generators preserve identity well under both talking and no-talking conditions: (1) Veo achieves the highest and most stable performance with near-perfect DPR: the faces in the AI-generated videos are almost always recognized as the seed face. (2) Kling follows closely, with DPR above 94% in both conditions and little sensitivity to whether the subject is speaking. Gen 4.5 maintains solid scores in the no-talking condition ($\geq$93%) but exhibits a small, consistent drop under the talking condition, suggesting that lip and jaw motion slightly perturbs the generated appearance. (3) Ray 2 presents a different pattern. Across all three detectors, its no-talking DPR falls to 66–72%, far below its own talking DPR of 96–100%. A plausible explanation is that without the constraint of speech dynamics, Ray 2's generation drifts further from the seed image through stronger stylization or scene adaptation, whereas the talking condition keeps the face anchored to a naturalistic, recognizable appearance.

Because Sora operates through a fundamentally different mechanism where Cameo directly embeds the user-provided face into the generation process rather than conditioning on an image, we report its results separately as a reference. Across all seven prompt conditions (frontal, three-quarter, profile, bright, dim, jogging, dancing), Sora achieves an overall average DPR of 94.7%, with near-perfect scores on frontal and well-lit conditions and a moderate drop on profile views (84.9%), where face-recognition models inherently struggle. These results confirm that Sora's Cameo mechanism reliably preserves identity, and the profile-view drop reflects the limits of the evaluators rather than a failure of generation quality. Although direct comparison with the image-conditioned models is not appropriate given the differing evaluation setup, the overall DPR range is broadly consistent with the better-performing generators, suggesting that both paradigms can achieve strong identity preservation under favorable conditions.

Voice similarity. Table 3 summarizes the comparison between the real–real baseline and the AI–real similarity across four voice similarity models. As Sora is the only generator that supports joint facial and voice conditioning via its Cameo feature, results are reported for Sora only. Overall, the generated voices closely match the real voices.

Low-level acoustic and content-level representations show minimal degradation: Log-Mel and wav2vec2 yield no significant difference between real and generated speech (p=0.055 and p=0.971, respectively), indicating that the generated audio preserves both spectral envelope and phonetic structure. MFCC reveals a small gap (p=0.042), which is not significant after Holm correction (p_{adj}=0.126), pointing to minor spectral differences that do not substantially affect perceived similarity. Resemblyzer, which targets speaker identity rather than acoustic detail, shows the largest and most significant gap ($p_{\text{adj}}<0.004$), with AI–real similarity dropping from 0.862 to 0.780. This indicates that while Sora's generated voices capture most acoustic characteristics of the source speaker, fine-grained identity-discriminative features are less fully preserved.

Overall, the AI–real similarity scores remain within the range observed for recordings of the same speaker. Sora's audio generation preserves most speaker characteristics and produces voices that sound close to the real speakers.

Lip-Sync Quality. Last, we evaluate lip–audio synchronization using the LSE-D (lower is better) and LSE-C (higher is better) metrics from SyncNet [38]. Table 4 reports results for Veo, Sora, and Doubao, alongside published benchmarks on real-world speech video datasets (news and YouTube).

All three generators produce speaking videos with strong audio–visual alignment, although their lip-sync performance remains below natural-video benchmarks on LSE-D. Veo achieves the highest LSE-C (8.08), while Doubao achieves the lowest LSE-D (7.32). Doubao's low LSE-D suggests tight mouth–motion timing, yet its LSE-C (7.53) is lower than Veo's, indicating weaker overall audio–visual confidence. Sora shows the highest LSE-D (7.76) and lowest LSE-C (7.32), suggesting less consistent synchronization across clips. Overall, Veo provides the highest audio–visual confidence, while Doubao achieves the best temporal alignment according to LSE-D. All three models fall short of natural YouTube video in terms of LSE-C, indicating room

TABLE 2: Face identity consistency (%) across video generators and face-recognition models. **IF** = InsightFace (ArcFace), **FN** = FaceNet (VGGFace2), **DF** = DeepFace (ArcFace + RetinaFace). †Sora uses the Cameo mechanism; results are reported as a reference and are not directly comparable to image-conditioned models.

	No-talking			Talking		
Generator	**IF**	**FN**	**DF**	**IF**	**FN**	**DF**
Kling 2.6	97.0	99.3	100.0	94.3	97.8	99.3
Veo 3.1	99.8	100.0	100.0	100.0	99.3	100.0
Ray 2	66.1	67.9	72.2	96.5	99.5	99.7
Gen 4.5	93.2	95.0	99.5	90.2	91.2	98.5
Sora†	Overall avg: IF = 98.6			FN = 92.1		DF = 93.5

TABLE 3: Paired t-test comparing real–real baseline similarity vs. AI–real similarity across six cameo identities using four independent voice similarity estimators.

Model	**Real Mean**	**AI Mean**	**t-statistic**	**p-value**
Log-Mel	0.9964	0.9934	2.49	0.055
MFCC	0.9745	0.9533	2.71	**0.042**
wav2vec2	0.8570	0.8562	0.04	0.971
Resemblyzer	0.8622	0.7804	9.11	**0.00027**

TABLE 4: Lip-sync performance of AI-generated videos compared with reference benchmarks. Lower LSE-D and higher LSE-C indicate better synchronization. †Two outliers removed via IQR filtering ($n = 48$).

	Veo 3.1	**Sora 2**	**Doubao**†	**News** [38]	**YouTube** [38]
LSE-D ↓	7.61	7.76	7.32	6.83	4.96
LSE-C ↑	8.08	7.32	7.53	7.00	8.51

for further improvement in audio–visual confidence.

4.3. Human Perception of Synthetic Video

We next evaluate whether human observers can reliably distinguish AI-generated videos from authentic videos. Across all 1460 classification annotations from 146 participants, the overall detection accuracy was 69.04%, indicating that modern generative video systems produce content that a substantial portion of users perceive as authentic.

Overall Classification Behavior. Participants correctly identified AI-generated videos as synthetic in 71.92% of cases (630/876), but were considerably less accurate at confirming authentic videos, correctly classifying only 64.73% of real videos (378/584). This association is statistically significant (χ^2=192.26, p<0.001), indicating that participants' annotations were strongly associated with the true label rather than random guessing. However, participants also exhibited a systematic tendency to label uncertain content as AI-generated, producing a false-positive rate of 35.27% for authentic videos. Conversely, 28.08% of AI-generated videos were misclassified as real, demonstrating that current generators produce content sufficiently realistic to deceive human observers in roughly one-quarter of cases.

Per-Platform Detection Accuracy and Effect of Audio. To enable fair comparison across all six generators, we first report per-platform detection accuracy using only Part 1 (muted) responses, in which videos from all platforms are presented without audio. We then compare these muted baselines with Part 2 (audio-enabled) results for the platforms that support synchronized speech generation, Sora 2, Veo 3.1, and Doubao, as well as for real videos. Table 5 reports the combined results.

For muted video clips, detection difficulty varies substantially across generators. Doubao and Veo 3.1 were the easiest to identify, with accuracies of 86.17% and 80.41%, respectively. Ray 2 followed at 76.84%, while Kling 2.6 and Gen 4.5 fell in an intermediate range at 69.15% and 67.31%. Sora 2 was the most difficult to detect (63.00%), approaching the detection rate for authentic content. Notably, authentic videos were classified correctly at the lowest rate of any category (62.33%), only slightly above chance, suggesting that exposure to generative media may cultivate a skeptical prior that degrades trust in authentic content.

Comparing audio-enabled results with the muted baselines, we observe an asymmetry in how synchronized speech affects perception. Overall detection accuracy did not differ significantly between muted and audio conditions (68.79% vs. 69.63%; χ^2=0.10, p=0.748), but this aggregate masks

TABLE 5: Human detection accuracy (%) for muted and audio-enabled videos. Negative Δ indicates reduced detection accuracy with audio.

Source	Doubao	Veo	Ray	Kling	Gen	Sora	Real
Muted Acc.	86.17	80.41	76.84	69.15	67.31	63.00	62.33
Audio Acc.	89.86	64.81	–	–	–	59.13	71.92
Δ	+3.69	-15.60	–	–	–	-3.87	+9.59

divergent per-source effects. Veo 3.1's detection accuracy dropped from 80.41% to 64.81% when audio was present, a decrease of 15.60 percentage points that is not significant after Holm correction (p_{adj}=0.076), suggesting that its synthesized speech may enhance perceived realism. Sora 2 showed a modest decrease from 63.00% to 59.13% ($\Delta = -3.87$; Fisher's exact p=0.578), suggesting that its visual output alone is already difficult to distinguish from real content. In contrast, Doubao's detection accuracy increased from 86.17% to 89.86% ($\Delta = +3.69$; Fisher's exact p=0.630), indicating that its synthesized audio does not enhance perceived realism. Detection accuracy for real videos increased from 62.33% to 71.92% when audio was present ($\Delta = +9.59$; χ^2=4.41, p=0.036), an improvement that is not significant after Holm correction (p_{adj}=0.108). Synchronized audio thus shows generator-dependent trends, potentially making Veo 3.1 videos harder to detect while improving recognition of real videos.

Confidence and Familiarity. Additional results on detection accuracy by confidence and AI-video familiarity are provided in Appendix Table 11.

Security implications. These results empirically validate Stage 4 of our threat model (§3.2): human detection accuracy is limited overall, with generator-dependent effects of synchronized audio. Notably, participants were more accurate at detecting AI-generated videos than confirming real ones (71.92% vs. 64.73%), misclassifying real videos as AI-generated at a rate of 35.27% compared to 28.08% for the reverse error. This asymmetry suggests that as generative quality improves, the dominant failure mode may shift from false negatives (failing to detect synthetic content) toward false positives (eroding trust in authentic media). Because our study uses simple, unoptimized prompts, these detection rates may differ under adversarially crafted content.

4.4. Forensic Detectability Across Generators

To evaluate how well current forensic tools identify AI-generated content, we conducted a series of tests using deepfake detectors, AI-image classifiers, and AI-video detectors.

Deepfake Image Detection. Table 6 summarizes the results from four deepfake detectors on frames extracted from AI videos. All detectors operate at the frame level, assessing the spatial realism of individual images without temporal context. Across models, detector behavior varies, depending on both the generator and the conditioning modality.

SightEngine is the weakest of the four detectors overall, classifying the majority of generated frames as real

TABLE 6: Deepfake image detection: the percentage of generated frames misclassified as real. Only frames with detectable faces are evaluated. SE: SightEngine. Note: Sora 2 uses Cameo while all others use image.

Model	Image/Cameo				Prompt			
	SE	Wvolf	Dima	Ashish	SE	Wvolf	Dima	Ashish
Veo 3.1	81.56	86.70	85.45	77.16	78.75	99.62	99.62	98.75
Sora 2	85.82	97.92	98.53	95.47	92.33	100.00	100.00	94.79
Ray 2	87.11	56.93	61.59	44.86	70.85	100.00	100.00	89.67
Gen 4.5	84.50	90.75	91.00	82.75	53.82	99.25	99.00	97.12
Kling 2.6	93.88	34.62	37.12	26.38	99.12	91.83	93.34	79.15
Doubao	–	–	–	–	74.55	100.00	99.87	97.42

regardless of generator or modality. Real-classification rates remain above 70% for nearly all conditions, with prompt-based Kling 2.6 reaching 99.12% and Sora 2 reaching 92.33%. The only notable exception is prompt-based Gen 4.5, where the rate drops to 53.82%. These results suggest that SightEngine's spatial forensic model is largely ineffective against current diffusion-based video generators.

The three open-source ViT-based detectors (Wvolf, Dima, and Ashish) are similarly ineffective on prompt-based generations: outputs from Veo 3.1, Sora 2, Ray 2, Gen 4.5, and Doubao are almost all classified as real by Wvolf and Dima (frequently 99–100%), with Ashish only slightly better (89–98%). However, these detectors show considerably stronger performance on content generated with seed images. Kling 2.6 image-conditioned frames were classified as real at 34.62%, 37.12%, and 26.38% by Wvolf, Dima, and Ashish, respectively, compared to 91.83%, 93.34%, and 79.15% for its prompt-based outputs. Ray 2 image-conditioned frames show a similar gap, falling to 56.93% (Wvolf), 61.59% (Dima), and 44.86% (Ashish). This indicates that seed images introduce forensically exploitable artifacts that these ViT-based models can partially detect, even when they fail almost entirely on prompt-based content from the same generator.

Overall, deepfake image detectors exhibit highly inconsistent behavior across modern diffusion-based video generators. No single detector provides reliable detection performance. SightEngine fails broadly across conditions, while the ViT-based detectors fail selectively on prompt-based outputs but show meaningful detection capability for image-conditioned content from certain generators.

AI Image Detection. Table 7 reports the performance of four AI-image detectors on frames generated by the six evaluated models. Compared with the deepfake detectors, these classifiers generally identify synthetic content more reliably, although performance varies substantially across generators and conditioning modalities. For prompt-based generations, SightEngine and Ateeqq are the strongest detectors: frames from Veo 3.1, Gen 4.5, and Kling 2.6 were almost all correctly classified as AI, while less than 7% of the frames from Sora 2 and Doubao were misclassified as real. Hemg is considerably less reliable on prompt-conditioned content, misclassifying 17–40% of the frames. For image-conditioned generations, the pattern partially reverses: Hemg

TABLE 7: AI-image detection on generated video frames. We report the percentage of frames classified as real (Real%); lower values indicate better detection. SE: SightEngine, UF: UnivFD.

Model	Image/Cameo				Prompt			
	SE	Ateeqq	Hemg	UF	SE	Ateeqq	Hemg	UF
Veo 3.1	22.3	0.9	26.0	99.0	0.3	0.0	20.3	99.9
Sora 2	14.0	13.9	8.8	97.7	2.8	6.6	16.9	98.0
Ray 2	48.1	45.3	29.3	94.9	3.1	21.9	24.3	98.4
Gen 4.5	13.3	38.5	30.5	97.1	0.4	0.0	29.9	100.0
Kling 2.6	0.4	29.5	22.1	92.6	0.3	0.9	40.0	96.5
Doubao	–	–	–	–	2.3	2.0	17.0	100

outperforms SightEngine and Ateeqq on Sora 2 (misclassification of 8.8% vs. 14.0% and 13.9%) and Ray 2 (29.3% vs. 48.1% and 45.3%), suggesting that image conditioning introduces artifacts that different detectors exploit in complementary ways. More broadly, image-conditioned frames are misclassified as real at substantially higher rates than prompt-based frames across most generators.

UnivFD, a CLIP-based detector designed for cross-generator generalization, misclassifies 92–100% of AI video frames as real, despite achieving strong results on diffusion-generated images [20]. This indicates that cross-generator generalization on still images does not transfer to frames extracted from diffusion-generated video, likely because the temporal generation process produces spatial artifact distributions that differ from those in single-image synthesis.

Overall, the three AI-image detectors substantially outperform both the deepfake detectors and the research prototype UnivFD, but their performance remains dependent on both the generator and conditioning modality. No single detector dominates across all settings, and image-conditioned content is generally harder to detect.

Cross-Detector Agreement. To quantify whether different forensic paradigms capture overlapping or complementary signals, we compute pairwise Cohen's κ between all eight frame-level detectors. Because deepfake detectors operate only on frames with faces while AI-image detectors operate on all extracted frames, each pairwise score is computed over the intersection of frames evaluated by both detectors. All comparisons are computed exclusively over AI-generated frames, so the reported agreement reflects how consistently detectors classify the same synthetic content, without being inflated by trivial agreement on authentic frames that both detectors would classify identically. Figure 1 presents the resulting agreement matrix, with detectors grouped by paradigm.

Among the deepfake detectors, only the three ViT-based models show meaningful agreement: Dima and Wvolf are nearly interchangeable (κ=0.90), with Ashish moderately aligned (κ=0.55, 0.56). Deepfake SightEngine, however, disagrees with all three (κ between -0.10 and -0.06), indicating that even detectors within the same paradigm can make fundamentally different classification decisions. Among the AI-image detectors, pairwise agreement is also

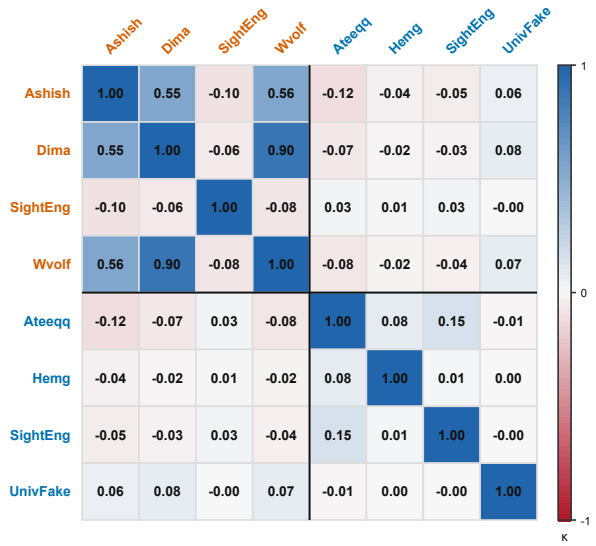


Figure 1: Pairwise Cohen's κ between all eight frame-level detectors, computed over frames with detected faces. Detector names are colored by paradigm (orange for deepfake, blue for AI-image). Cross-paradigm agreement is near zero, indicating that the two detection families make largely independent classification decisions.

low (κ ranging from -0.01 to 0.15), despite three of them being more effective than the deepfake detectors. This indicates that each AI-image detector exploits different artifact signatures and that their overall effectiveness does not imply redundancy. Importantly, low κ does not contradict individually high detection rates. Cohen's κ measures whether two detectors agree on the *same frames* both positively and negatively. Two detectors can each correctly classify the vast majority of generated frames yet misclassify two disjoint sets of frames, producing high per-detector accuracy but near-zero κ as they are not making errors on the same samples. This pattern is precisely what we observe among the AI-image detectors and confirms that their errors are distributed over different subsets of frames, making them complementary, not redundant.

Cross-paradigm agreement is near zero across all pairs (κ between -0.12 and 0.08), confirming that deepfake and AI-image detectors make largely independent classification decisions. Taken together, these results show that no single detector provides comprehensive coverage and suggest the potential for ensemble-based forensic detectors that integrate complementary detection approaches.

AI Video Detection. Table 8 reports AI video detection results for SightEngine and AI or Not across generators, sources, and talking settings. Both detectors correctly identify most prompt-based AI videos, but detectability varies substantially with the generator and content source, and the two detectors exhibit complementary failure modes.

The most consistent finding is that generation with seed images systematically evades both detectors. Across Ray 2, Gen 4.5, and Veo 3.1, image-conditioned videos achieve significantly higher misclassification rates than their prompt-based counterparts. Kling is the sole exception, with both detectors correctly identifying all or nearly all of its videos as AI-generated regardless of source modality or talking condition. The effect is most extreme for Gen 4.5: AI or Not classifies 100% of image-conditioned Gen 4.5 videos as real under both talking conditions, a complete detection failure despite SightEngine identifying most of the same content correctly, with a misclassification rate of 0–16%. Ray 2 image-conditioned videos similarly evade AI or Not at rates of 92–98%, far exceeding its 2–4% failure rate on Ray 2 prompt-based content.

TABLE 8: AI video detection results measured as the percentage of generated videos classified as real by two AI video detectors: SightEngine (SE) and AI or Not (AI/Not). Results are reported separately for generation type and talking condition, where No denotes no talking.

Model	Image / Cameo				Prompt			
	SE		AI/Not		SE		AI/Not	
	No	Talk	No	Talk	No	Talk	No	Talk
Doubao	–	–	–	–	0.0	0.0	16.0	10.0
Kling 2.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.0
Ray 2	38.0	64.0	92.0	98.0	8.0	6.0	4.0	2.0
Gen 4.5	16.0	0.0	100.0	100.0	0.0	0.0	4.0	0.0
Sora 2	8.0	4.0	0.0	0.0	0.0	4.0	0.0	28.0
Veo 3.1	16.0	10.0	74.0	34.0	2.0	2.0	0.0	0.0

The two detectors also diverge on specific generators in ways that reveal complementary blind spots. Sora 2 prompt-based talking videos trigger the clearest disagreement: SightEngine misclassifies only 4% of such videos, while AI or Not misclassifies 28%, the highest prompt-based evasion rate observed for any generator under either detector, suggesting that Sora 2's talking-mode generation produces artifacts that partially evade temporal forensic signals.

The talking condition further modulates detectability, especially for image-conditioned content, though not in a consistent direction. For Ray 2, misclassification rises from 38% to 64% under SightEngine and from 92% to 98% under AI or Not when the subject is talking, indicating that synchronized facial motion suppresses detectable artifacts. Veo image-conditioned content exhibits the opposite pattern: misclassification decreases from 16% to 10% under SightEngine and from 74% to 34% under AI or Not when talking is present, suggesting that for Veo, speech dynamics introduce rather than suppress forensic signals. This generator-dependent reversal indicates that the relationship between talking behavior and detectability is not uniform and likely depends on how each system's synthesis pipeline handles audio–visual coordination.

Overall, AI video detectability is determined primarily by the conditioning modality. Videos generated with seed images systematically challenge both detectors, with Ray 2 and Gen 4.5 exhibiting the strongest evasion. Videos generated from only text prompts are reliably detected by

both systems, with the exception of Sora 2 talking videos under AI or Not.

False Positives on Authentic Videos. We further evaluated forensic detectors on 100 authentic videos, measuring false-positive rates at the frame level for image detectors and at the video level for video detectors. Among deepfake image detectors, Wvolf, Dima, and Ashish showed relatively low false-positive rates (2.57–12.57%), while SightEngine reached 65.26% on non-talking videos. For video-level detectors, SightEngine achieved false-positive rates of 0% and 16% on talking and non-talking videos, respectively, whereas AI or Not reached 28% and 98%. These results reveal substantial differences in detector reliability on authentic content.

Robustness to Complex Prompts. To assess whether our findings generalize beyond simple prompts, we generated 150 additional videos using 10 complex prompts across the five available platforms, excluding Sora. These prompts describe motion, object interactions, lighting and viewpoint changes, dynamic backgrounds, and occlusion. SightEngine and AI or Not each correctly identified 99.3% of these videos as AI-generated. Lip-sync quality remained stable for Veo (LSE-D/LSE-C: 7.71/7.86), with some degradation for Doubao (7.69/5.70).

4.5. Summary of Findings by Research Question

We now consolidate our empirical findings and answer the five research questions posed in Section 1.

RQ1: Safety and privacy safeguards. Platform governance is highly inconsistent. Four of six platforms accept human-image uploads without consent controls, allowing non-consensual identity reuse. Modality-dependent copyright enforcement on Gen 4.5 and Ray 2 allows image-conditioned inputs to bypass prompt-level filters. Visible watermarks are absent on API-generated videos across most platforms, while visible watermarks and embedded C2PA provenance can be removed or made undetectable through basic post-processing. Only Sora 2 implemented structured identity-governance mechanisms (cameo permissions), but they are not fully retroactive (Table 1; §3.3).

RQ2: Facial and vocal identity consistency. All evaluated platforms preserve facial identity at high rates under favorable conditions. Veo 3.1 achieves near-perfect DPR across all conditions and face-recognition models, while Kling 2.6, Runway Gen-4.5, and Sora 2 maintain DPR above 93% in most settings. Luma Ray 2 exhibits condition-dependent variability, with no-talking DPR dropping to 66–72% (Table 2). For voice identity, Sora 2's generated speech closely matches real speakers on low-level acoustic (Log-Mel, MFCC) and content-level (wav2vec2) features, with a moderate but statistically significant gap on speaker-identity embeddings (Resemblyzer; Table 3).

RQ3: Lip-sync accuracy. All three platforms supporting synchronized speech generation achieve strong audio–visual alignment, although their LSE-D scores are higher than those of natural-video benchmarks. Veo 3.1 achieves the highest LSE-C (8.08), while Doubao achieves the lowest LSE-D (7.32). Sora 2 records an LSE-D of 7.76 and an LSE-C of 7.32. All three generators fall short of natural YouTube video on LSE-C, indicating room for further improvement in audio–visual confidence (Table 4).

TABLE 9: Optimal adversarial generator selection per forensic detector. For each detector, we report the generator–modality combination that achieves the highest evasion (maximum Real%). Values are drawn from Tables 6–8.

Paradigm	Detector	Optimal Generator	Real%
Deepfake	SightEngine	Kling (Prompt)	99.1
	Wvolf	Sora, Ray, Doubao (Prompt)	100.0
	Dima	Sora, Ray (Prompt)	100.00
	Ashish	Veo (Prompt)	98.8
AI-Image	SightEngine	Ray (Image)	48.1
	Ateeqq	Ray (Image)	45.3
	Hemg	Kling (Prompt)	40.0
	UnivFD	Gen, Doubao (Prompt)	100.0
AI-Video	SightEngine	Ray (Image)	64.0
	AI or Not	Gen (Image)	100.0

RQ4: Forensic detectability. A clear hierarchy emerges across detection paradigms. Deepfake image detectors fail broadly, particularly on prompt-based videos (Table 6). AI-image detectors are more effective but exhibit generator-specific blind spots. Video-level detectors generally provide strong detection coverage of generated content, though image-conditioned videos systematically evade both evaluated video detectors for several generators, and some detectors exhibit high false-positive rates on authentic videos (Table 8). Cross-paradigm agreement is low (Figure 1), confirming that the two detection families capture complementary rather than redundant signals. For every detector tested, at least one commercially available generator achieves high evasion, enabling adversarial platform selection (Table 9).

RQ5: Human perception. Participants achieved an overall detection accuracy of 69.04% across 1,460 classification decisions. Synchronized audio exerted a generator-dependent effect: Veo 3.1's detection accuracy dropped by 15.60 percentage points when audio was present, while detection of authentic videos improved, although neither difference remained significant after Holm correction. Notably, participants misclassified authentic videos as AI-generated at a higher rate (35.27%) than they misclassified AI-generated videos as real (28.08%), suggesting that improving generative realism may shift the dominant failure mode toward erosion of trust in authentic media (Table 5; Table 11).

5. Discussions and Limitations

Implications for Governance and Forensics. Our results reveal a consistent theme across both governance and forensic dimensions: the heterogeneity of deployed platforms creates an attack surface that no single defense layer addresses. On the governance side, no two platforms enforce the same combination of identity, copyright, and provenance controls. As NIST guidance emphasizes that no single transparency approach is sufficient in isolation [48], our results

empirically confirm this. No evaluated platform provides effective consent controls, robust watermarks, and reliable provenance signaling. On forensics, the failure modes are paradigm-specific. Deepfake image detectors fail broadly because they target localized facial manipulations absent in diffusion-generated videos, a fundamental mismatch unlikely to close through incremental retraining. AI-image detectors achieve better coverage but are generator-dependent. Conditioning modality further complicates detection: image-conditioned videos are generally harder to detect.

We consider a practical scenario where only one deepfake/AI detector is deployed in a video-sharing platform. A stronger attacker, who has knowledge of the deployed detector, could easily identify the best AI video generator and generation modality to create videos that are highly likely to evade this detection, as shown in Table 9. Meanwhile, a dedicated attacker only needs to invoke a combination of two to three generations (e.g., generating two videos using Sora/Prompt and Ray/Image) to achieve near-optimal evasion performance against all detectors. These results demonstrate that platforms deploying only one detection paradigm leave systematic blind spots exploitable through informed generator selection. Effective mitigation requires coordinated defenses spanning platform governance, multi-paradigm detection, and provenance infrastructure.

Asymmetric Trust Erosion. An especially concerning finding from our user study is that participants misclassified authentic videos as AI-generated at a higher rate (35.27%) than they misclassified AI-generated videos as real (28.08%). This asymmetry suggests that the primary societal risk of improving generative video may not be deception through synthetic content alone, but also the erosion of trust in authentic media—a dynamic sometimes termed the "liar's dividend" [19], in which the mere existence of convincing synthetic media enables bad actors to dismiss authentic evidence as fabricated. Our results provide quantitative evidence that this dynamic is already observable: under muted conditions, only 62.33% of the authentic videos were classified correctly as real, the lowest accuracy of any category, including all six AI video generators.

This finding carries further implications: for legal proceedings, video evidence may face challenges to its admissibility; for journalism, authentic footage may be reflexively dismissed; and for interpersonal contexts, individuals may deny the authenticity of genuine recordings by claiming AI generation. Critically, existing forensic tools do not address this gap. They are designed to flag synthetic content, not to certify authentic content. Addressing trust erosion likely requires a shift toward provenance-first approaches, in which authentic media is cryptographically signed at the point of capture, enabling downstream verification that does not depend on the absence of synthetic artifacts.

Limitations and Future Directions. This study has a few limitations: (1) Our video dataset size is modest due to API costs and rate limits. (2) Our analysis reflects only publicly observable behavior, not internal moderation pipelines. Voice similarity and lip-sync evaluations are limited to platforms supporting synchronized speech. (3) Platform policies evolve rapidly, so that our findings only reflect behavior during the measurement time. In particular, although Sora 2's interfaces were discontinued after our measurement study, the issues identified in this study remain relevant to future systems that expose similar capabilities, such as cameo-style personalization and user-facing consent controls. Hence, we interpret the Sora 2 results not only as a retrospective audit of a specific platform, but also as evidence of governance and forensic challenges likely to recur in successor or functionally similar AI video systems.

Finally, our methodological choices are designed to prioritize ecological validity over controlled comprehensiveness. We use simple prompts that provide a conservative forensic baseline but do not represent adversarial prompt strategies (e.g., jailbreaks). Future work should complement this approach with adversarial prompts, additional research-grade detectors, longitudinal monitoring of platform evolution, larger-scale multimodal benchmarks, and evaluations of emerging provenance technologies [49], [48].

Ethical Considerations. The objective of this study is to enhance the understanding of the security, privacy, and forensic risks posed by generative video platforms and to stimulate public awareness. All evaluations were conducted through publicly accessible interfaces and APIs using standard user accounts. We did not exploit any platform vulnerabilities or attack any real-world system. Seed images were drawn from the publicly available FFHQ dataset. Generated videos were used solely for research analysis. The user study presented in this paper has been reviewed and approved by the Institutional Review Board at our institution.

6. Conclusion

In this paper, we present an external, black-box security and forensic audit of six deployed generative video platforms. We find that platform governance is fragmented, with modality-dependent enforcement gaps that allow malicious users to bypass prompt-level restrictions. Deepfake image detectors fail broadly, AI-image detectors are more effective but generator-dependent, and no single detector reliably covers all generators and conditioning modalities. Generated videos preserve strong facial identity and achieve strong audio–visual alignment. Human observers achieve only 69.04% detection accuracy and misclassify authentic videos as AI-generated more often than the reverse. Together, these findings demonstrate a need for coordinated defenses spanning platform governance, multi-paradigm detection, and provenance infrastructure.

Acknowledgment

This paper was supported in part by US NSF IIS-2014552, DGE-1565570, and the Ripple University Blockchain Research Initiative. We thank the anonymous reviewers for their valuable comments and suggestions.

References

[1] OpenAI, "Sora 2 is here," 2025. [Online]. Available: https://openai.com/index/sora-2

[2] Google, "Introducing veo 3.1 and advanced capabilities in flow," https://blog.google/technology/ai/veo-updates-flow/, 2025.

[3] I. Amerini, M. Barni, S. Battiato, P. Bestagini, G. Boato, T. S. Bonaventura, V. Bruni, R. Caldelli, F. De Natale, R. De Nicola *et al.*, "Deepfake media forensics: State of the art and challenges ahead," in *International Conference on Advances in Social Networks Analysis and Mining*. Springer, 2024, pp. 33–48.

[4] A. A. Khan, A. A. Laghari, S. A. Inam, S. Ullah, M. Shahzad, and D. Syed, "A survey on multimedia-enabled deepfake detection: state-of-the-art tools and techniques, emerging trends, current challenges & limitations, and future directions," *Discover Computing*, vol. 28, p. 48, 2025.

[5] YouTube, "Disclosing use of altered or synthetic content," https://support.google.com/youtube/answer/14328491.

[6] TikTok, "Ai-generated content," https://support.tiktok.com/en/using-tiktok/creating-videos/ai-generated-content.

[7] L. AI, "Ray2: Next-generation video ai model." [Online]. Available: https://lumalabs.ai/ray2

[8] K. Technology, "Kling ai: Text-to-video generation system." [Online]. Available: https://kling.ai/

[9] I. Runway AI, "Gen-4: Video generation model." [Online]. Available: https://runwayml.com/research/introducing-runway-gen-4.5

[10] ByteDance Seed Team, "Seedance 1.5 Pro," https://seed.bytedance.com/en/seedance1_5_pro, 2025.

[11] W. Hong, M. Ding, W. Zheng, X. Liu, and J. Tang, "Cogvideo: Large-scale pretraining for text-to-video generation via transformers," 2022. [Online]. Available: https://arxiv.org/abs/2205.15868

[12] R. Villegas, M. Babaeizadeh, P.-J. Kindermans, H. Moraldo, H. Zhang, M. T. Saffar, S. Castro, J. Kunze, and D. Erhan, "Phenaki: Variable length video generation from open domain textual descriptions," in *International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=vOEXS39nOF

[13] J. Wang, H. Yuan, D. Chen, Y. Zhang, X. Wang, and S. Zhang, "Modelscope text-to-video technical report," 2023. [Online]. Available: https://arxiv.org/abs/2308.06571

[14] Z. Yuan, Y. Liu, Y. Cao, W. Sun, H. Jia, R. Chen, Z. Li, B. Lin, L. Yuan, L. He, C. Wang, Y. Ye, and L. Sun, "Mora: Enabling generalist video generation via a multi-agent framework," 2024. [Online]. Available: https://arxiv.org/abs/2403.13248

[15] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards accurate generative models of video: A new metric & challenges," 2019. [Online]. Available: https://arxiv.org/abs/1812.01717

[16] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," in *Proceedings of the 2021 conference on empirical methods in natural language processing*, 2021, pp. 7514–7528.

[17] Y. Miao, Y. Zhu, L. Yu, J. Zhu, X.-S. Gao, and Y. Dong, "T2vsafetybench: Evaluating the safety of text-to-video generative models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 63 858–63 872, 2024.

[18] S. Agarwal, H. Farid, T. El-Gaaly, and S.-N. Lim, "Detecting deepfake videos from appearance and behavior," in *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2020, pp. 1–6.

[19] B. Chesney and D. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," *Calif. L. Rev.*, vol. 107, p. 1753, 2019.

[20] U. Ojha, Y. Li, and Y. J. 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. 24 480–24 489.

[21] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li, "Dire for diffusion-generated image detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 445–22 455.

[22] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "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.

[23] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)*. IEEE, 2018, pp. 1–6.

[24] R. Corvi, D. Cozzolino, E. Prashnani, S. D. Mello, K. Nagano, and L. Verdoliva, "Seeing what matters: Generalizable AI-generated video detection with forensic-oriented augmentation," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. [Online]. Available: https://openreview.net/forum?id=dOGXKBL7IE

[25] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "Fakeavceleb: A novel audio-video multimodal deepfake dataset," *arXiv preprint arXiv:2108.05080*, 2021.

[26] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Asian conference on computer vision*. Springer, 2016, pp. 251–263.

[27] C. Villa, S. Mirza, and C. Pöpper, "Exposing the guardrails:{Reverse-Engineering} and jailbreaking safety filters in {DALL· E}{Text-to-Image} pipelines," in *34th USENIX Security Symposium (USENIX Security 25)*, 2025, pp. 897–916.

[28] T. Karras, S. Laine, and T. 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.

[29] OpenAI, "What to know about the sora discontinuation," https://help.openai.com/en/articles/20001152-what-to-know-about-the-sora-discontinuation, 2026.

[30] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, "Closing the ai accountability gap: defining an end-to-end framework for internal algorithmic auditing," in *Conference on Fairness, Accountability, and Transparency*, 2020.

[31] M. A. Madaio, L. Stark, J. Wortman Vaughan, and H. Wallach, "Co-designing checklists to understand organizational challenges and opportunities around fairness in ai," in *CHI*, 2020.

[32] S. Serengil and A. Ozpinar, "A benchmark of facial recognition pipelines and co-usability performances of modules," *Journal of Information Technologies*, vol. 17, no. 2, pp. 95–107, 2024.

[33] J. Deng, J. Guo, X. Niannan, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *CVPR*, 2019.

[34] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, Jun. 2015.

[35] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[36] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[37] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[38] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *ACM Multimedia*, ser. MM '20, 2020.

[39] Sightengine, "Deepfake detection," https://sightengine.com/docs/deepfake-detection.

[40] Ashish, "Deepfake image detection using fine-tuned vision transformer (vit) (ashish-001/deepfake-detection-using-ViT)," https://huggingface.co/ashish-001/deepfake-detection-using-ViT, 2024.

[41] D. Iakubovskyi, "Deepfake vs. Real Faces Detection with ViT," https://huggingface.co/dima806/deepfake_vs_real_image_detection, 2023.

[42] R. K. Enyimba, "ViT Deepfake Detection," https://huggingface.co/Wvolf/ViT_Deepfake_Detection, 2023.

[43] Sightengine, "Ai-generated image detection model," https://sightengine.com/docs/ai-generated-image-detection.

[44] A. Azam, "Ai and human image classification model - v1(Ateeqq/ai-vs-human-image-detector)," https://huggingface.co/Ateeqq/ai-vs-human-image-detector, 2025.

[45] H. B. Gurung, "Ai-vs-real-image-detection (Hemg/AI-VS-REAL-IMAGE-DETECTION)," https://huggingface.co/Hemg/AI-VS-REAL-IMAGE-DETECTION, 2024.

[46] Sightengine, "Ai-generated video detection," https://sightengine.com/docs/ai-generated-video-detection.

[47] AI or Not, "AI or Not: AI-Generated Video Detection," https://www.aiornot.com, 2024, accessed: 2025.

[48] B. Chandra, J. Dunietz, K. Roberts, Y. Lee, P. Fontana, and G. Awad, "Reducing risks posed by synthetic content an overview of technical approaches to digital content transparency," 2024-11-20 05:11:00 2024. [Online]. Available: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=959123

[49] Coalition for Content Provenance and Authenticity, "C2PA technical specification," https://c2pa.org/specifications/specifications/2.1/specs/C2PA_Specification.html, 2024.

Appendix

TABLE 10: Safety and access-control observations in Sora 2.

Control / Mechanism	Expected Behavior	Observed Behavior	Security Implication
Human image upload restriction	Human likeness inputs are blocked to reduce impersonation risk.	Uploads of human faces, cropped faces, and bodies were rejected.	Reduces direct deepfake creation from real imagery; shifts attackers toward prompt-only synthesis.
Cameo permission enforcement	Generation requires explicit cameo permission.	Cameo selection is disabled when permission is absent.	Prevents unauthorized or confused deputy attempts to generate likenesses.
Consent revocation (post-generation)	Withdrawing consent should prevent continued access to generated media.	Previously published videos remained viewable until manually deleted.	Manual takedown introduces a temporal window for exposure or misuse.
Screen recording protection	Prevent capture of generated media during playback.	Display blackened during recording attempts, but audio remained recordable.	Partial exfiltration channel remains via audio leakage.
Copyrighted content filtering	Restrict copyrighted-character generation.	Most prompts/uploads were blocked; rare bypasses occurred.	Reduces IP replication, but bypasses leave residual risk.
Violence guardrails	Disallow generation of violent content.	Violent prompts were blocked.	Baseline enforcement for high-risk prompt categories.

1. Watermark Illustration

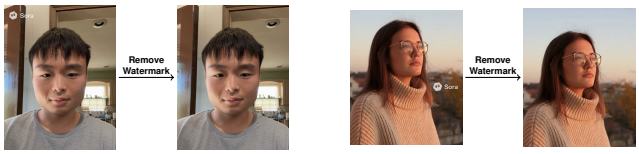


Figure 2: Sequence illustrating watermark handling in Sora 2. The sequence shows a visible watermark and how it can be removed through basic post-processing.

2. Confidence and Familiarity

Table 11 reports classification accuracy stratified by self-reported confidence and AI-video familiarity. Both factors exhibit a generally positive relationship with detection accuracy. At the lowest confidence level, participants performed below chance (48.84%), while the most confident responses reached 78.32%. A similar pattern holds for familiarity: the least familiar participants scored near chance (57.69%), and even the most familiar group misclassified roughly one-tenth of videos.

TABLE 11: Classification accuracy by (a) per-decision confidence level and (b) participant-level AI-video familiarity. Both factors show a positive relationship with accuracy.

(a) Confidence			(b) Familiarity		
Level	n	**Acc. (%)**	**Level**	n	**Acc. (%)**
1	86	48.84	1 (Low)	13	57.69
2	153	52.29	2	35	67.14
3	373	66.49	3	46	66.09
4	419	72.08	4	40	71.75
5	429	78.32	5 (High)	12	89.17
n = responses			n = participants		