Detecting AI-Generated Video: A Vision-Language Dual-View Survey

Dylan Xinming Hou[1] **Juntian Zhang**[2] **Xu Gu**[2] **Yichen Wu**[3]
Nils Lukas[1] **Gus Xia**[1] **Xiuying Chen**[1] **Yuhan Liu**[1]*
[1]MBZUAI
[2]Gaoling School of Artificial Intelligence, Renmin University of China
[3]Harvard University
dyxhou@gmail.com, {zhangjuntian, guxu}@ruc.edu.cn, yiwu6@mgh.harvard.edu
{nils.lukas, gus.xia, xiuying.chen, yuhan.liu}@mbzuai.ac.ae

Abstract

The evolving realism of AI-generated videos (AIGC-V) is rapidly rendering traditional artifact-centric detection insufficient, necessitating a paradigm shift from low-level inspection to high-level semantic verification. This paper presents a comprehensive survey of AIGC-V detection, reframing the task as Factual Fidelity Verification, which asks whether the events, entities, and physical processes depicted in a video are consistent with real-world facts. To systematize this rapidly evolving field, we propose a ***Vision-Language Dual-View*** taxonomy that organizes existing methods into a hierarchical, four-layer landscape spanning intrinsic cue analysis, spatiotemporal consistency modeling, cross-modal consistency reasoning, and language-guided world-level reasoning. This dual-view framing highlights a fundamental transition from artifact matching in traditional deepfake detection to evidence-based semantic verification enabled by vision-language models and agentic reasoning pipelines. Based on a systematic review of 221 works as of March 2026, we synthesize AIGC-V generation paradigms, survey the landscape of detection methods, and review evaluation metrics and benchmarks in line with the proposed views. Finally, we discuss current challenges and identify promising directions toward robust, explainable, and trustworthy detection.[1]

1 Introduction

The rapid evolution of video generation models, exemplified by Sora 2 (OpenAI, 2024a), Veo 3 (Google DeepMind, 2025), and Seedance 2.0 (ByteDance Seed, 2026), is fundamentally reshaping the information landscape. Unlike early deepfakes dominated by localized face swapping (Wang et al., 2024b), modern **AI-Generated Content-Videos (AIGC-V)** have achieved cinematic fidelity with coherent narratives, increasingly blurring the boundary between synthesized fiction and captured reality. This technological leap destabilizes the foundational trust in video evidence, traditionally relied upon to verify "who, when, where, and what" (Ho et al., 2022a).

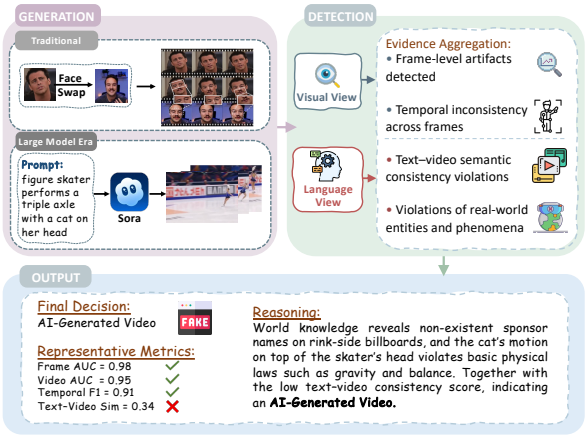

Figure 1: An example of the AIGC-V detection pipeline under our approach, illustrating AI-generated videos from traditional methods or text-to-video prompts, detection from the visual and language views, and outputs at different levels.

As generation paradigms shift from local manipulation to end-to-end synthesis, traditional detection methods (Wang et al., 2025d; Ma et al., 2025) face a critical bottleneck. Early detection systems primarily relied on low-level visual artifacts such as blending boundaries. However, advanced diffusion models and Transformers (Ho et al., 2022b; OpenAI, 2024a) can now produce visually high-fidelity videos. This paradigm shift necessitates

*Corresponding author: yuhan.liu@mbzuai.ac.ae
[1]Repository:github.com/dxhou/AI-Generated-Video-Detection. Project page: aigcvdetection.github.io.

Findings of the Association for Computational Linguistics: ACL 2026, pages 32221–32255
July 2-7, 2026

a new detection landscape, shifting from perceptual inspection to cognitive reasoning over semantic and factual violations. The surge of vision-language models(VLMs) (Zhang et al., 2025d,c; Wang et al., 2026) and agentic frameworks (Liu et al., 2024b, 2025g,f) offers a promising pathway, enabling detectors to perform world-level reasoning and verify cross-modal consistency. Although recent works (Fu et al., 2025; Park et al., 2025) have begun to explore these directions, the existing literature remains fragmented. Most prior surveys (Pei et al., 2024; Liu et al., 2025c) remain constrained to the early era of deepfakes or treat video merely as a sequence of images, failing to systematize the emerging class of methods that leverage language for semantic verification.

To address this gap, we propose a ***Vision-Language Dual-View*** taxonomy to organize the AIGC-V detection landscape, as illustrated in Figure 1. We reframe the problem as *Factual Fidelity Verification*, determining whether the video content aligns with real-world facts and physical laws, rather than simple binary classification. Our framework categorizes detection methods into a four-layer hierarchy, progressing from low-level perception to high-level cognition: (1) Intrinsic Cue Analysis, (2) Spatiotemporal Consistency, (3) Cross-Modal Consistency, and (4) Language-Guided World-Level Reasoning.

In this survey, we comprehensively review 221 works as of March 2026. We begin with AIGC-V paradigms (§2), then formulate AIGC-V detection from the perspective of *factual fidelity* and introduce our Vision-Language dual-view framing (§3). We systematically organize methods under this landscape (§4), revisit evaluation metrics (§5.1) from a dual-view, and review benchmarks in line with AIGC-V paradigms (§5.2). Finally, we identify the critical challenges (§6) in detecting increasingly sophisticated AIGC-V. By bridging the gap between traditional visual forensics and emerging multimodal reasoning, this paper aims to provide a structured roadmap for the next generation of explainable and trustworthy AIGC-V detection.

2 Paradigms of AI-Generated Video

Currently, diffusion-based and Transformer-based architectures are advancing rapidly in video generation, driving major progress in text-to-video and other AIGC-V pipelines (Ho et al., 2022b,a; Singer et al., 2022; OpenAI, 2024a; Google DeepMind, 2025; ByteDance Seed, 2026). We group mainstream methodologies into three categories according to the underlying generation paradigm: local manipulation, audio-visual editing, and generative video synthesis. Appendix B includes Figure 4 for illustration.

2.1 Local Manipulation

Methodologies categorized as *Local Manipulation* typically operate on authentic video sequences by modifying specific spatial regions or distinct semantic attributes (Brison et al., 2025; Heo and Woo, 2025). The resulting generation maintains a high degree of structural fidelity to the original footage. Another technical paradigm centers on video face swapping and facial element manipulation. Representative approaches typically utilize diffusion models (Wang et al., 2024b), 3D facial priors (Wang et al., 2025d), and controllable conditional encoding, integrating identity feature transfer with expression, pose, and illumination reconstruction from source videos into a unified generative framework. Given their ease of implementation and covert nature, these techniques represent a critical threat in real-world adversarial scenarios.

2.2 Audio-Visual Editing

Audio-Visual Editing uses speech as an explicit control signal to drive talking-head generation or to re-dub existing videos, requiring tight cross-modal alignment between audio and face dynamics while preserving identity and background (Cheng et al., 2022b). And recent diffusion-based or 3D-aware approaches (Ma et al., 2025; Hong et al., 2025) move toward unified conditional generation, where audio guidance and identity constraints are jointly modeled to produce lip-synced (Li et al., 2024a), visually consistent outputs. Due to their low operational barrier and strong perceptual plausibility, audio-visual edits can convincingly "bind" forged visual performances to plausible voice tracks, amplifying risks in impersonation and misinformation.

2.3 Generative Video Synthesis

Generative Video Synthesis targets end-to-end creation of complete video sequences from text (or noise), fabricating both appearance and dynamics without authentic carriers. Diffusion-based generators extend image diffusion with temporal modeling for coherence, exemplified by Video Diffusion Models (Ho et al., 2022b) and cascaded high-fidelity pipelines such as Imagen Video (Ho et al.,

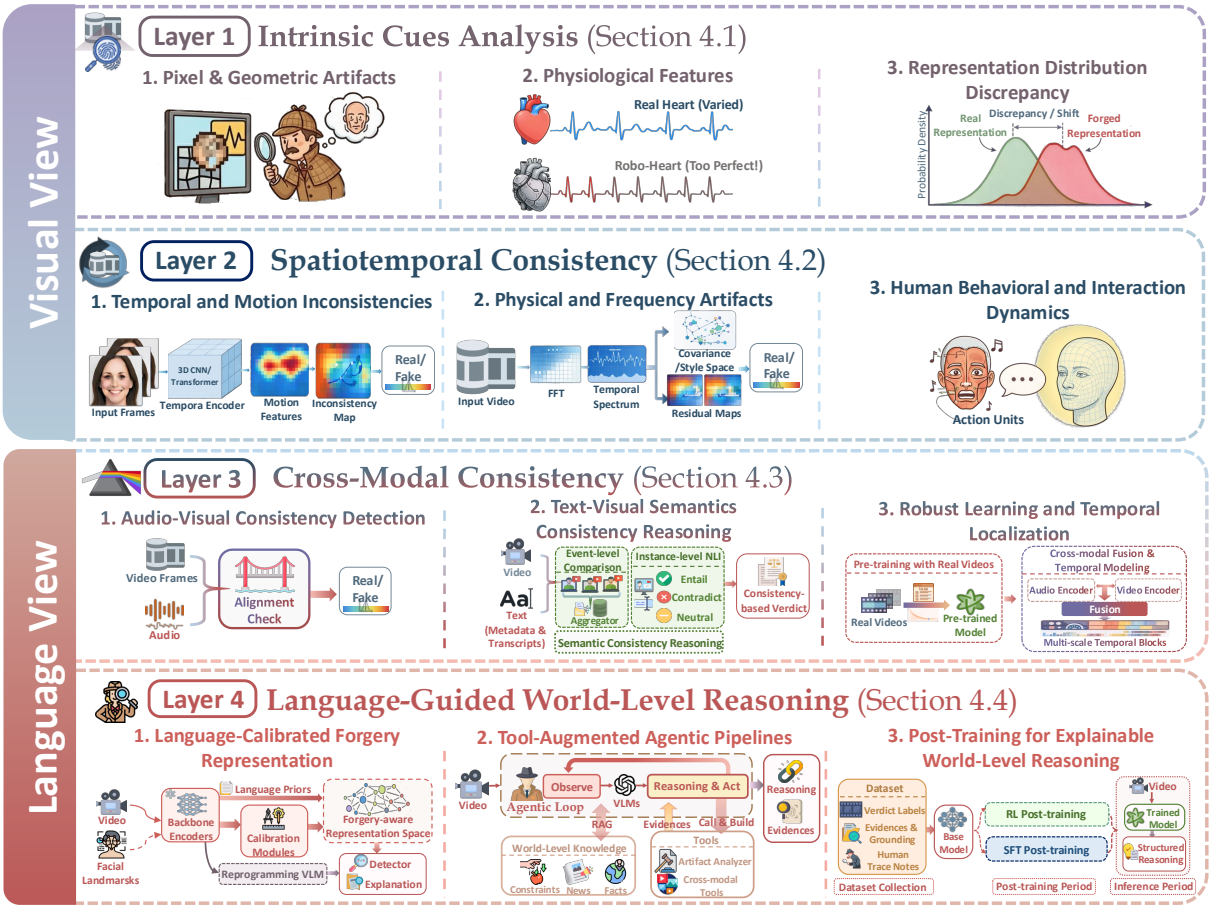

Figure 2: Dual-view, four-layer framework for AIGC-V detection. The visual view, spanning Layers 1-2, models intrinsic cues and spatiotemporal consistency, whereas the language view, spanning Layers 3-4, addresses cross-modal consistency and world-level reasoning, with each layer further decomposed into fine-grained subcategories.

2022a); Make-A-Video (Singer et al., 2022) further reduces reliance on paired text-video data by leveraging text-image priors and unpaired videos. Recent architectures further optimize representations: Show-1 (Zhang et al., 2025b) hybridizes pixel-based generation with latent refinement for efficiency, while Grid Diffusion (Lee et al., 2024) unifies spatiotemporal dimensions into a single 2D grid to simplify modeling. Transformer-style synthesis tokenizes videos and learns long-horizon structure via autoregressive or masked modeling. At the industrial frontier, large-scale text-to-video systems such as Sora 2 (OpenAI, 2024a), Seedance 2.0 (ByteDance Seed, 2026), and Veo 3 (Google DeepMind, 2025) demonstrate strong open-domain capability and raise urgent needs for provenance and verification (OpenAI, 2024b).

These three paradigms also imply different evidential emphases for detection. Local manipulation tends to preserve an authentic carrier and therefore leaves more localized forensic traces; audio-visual editing is more strongly constrained by synchrony and identity consistency across modalities; and generative video synthesis weakens explicit edit residue, shifting verification toward long-range coherence, factual plausibility, and provenance.

3 AI-Generated Video Detection

3.1 Task Scope

Visually high-fidelity AIGC-V with richer narratives make the reliability of methods that output only a "real/fake" probability increasingly low in security and privacy scenarios (Wang et al., 2024a; Kaur et al., 2024). The detection task should therefore shift from merely answering whether a video is AI generated to making a **fact-level** judgment about whether the depicted content objectively happened in the real world.

From a fact-level perspective, video content can be abstracted as a series of propositions of the form "at a certain time and place, which entities exist, and what happens to these entities." We define **Factual Fidelity** as the requirement that the fact-level

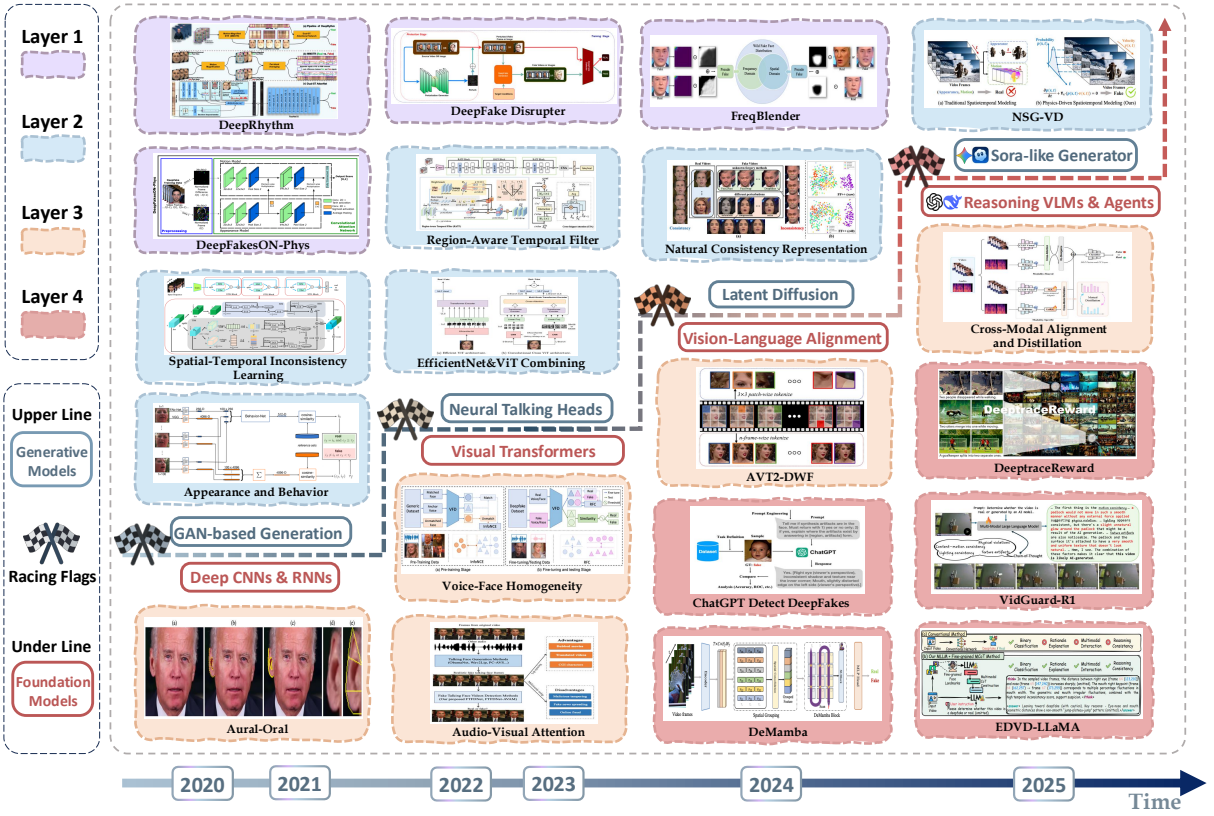

Figure 3: Landscape of representative AIGC-V detection methods aligned with the four-layer taxonomy in §4. Methods are grouped by the evolving capabilities of generative models that pose emerging threats and the corresponding advances in foundation models that enable detection, revealing an increasing share of language view methods.

propositions implied by the video and its metadata remain consistent with the real world, and treat *Factual Fidelity Verification* as the fundamental objective of AIGC-V detection. Different evidence pathways provided by detection methods can then be integrated to indicate at which levels the video violates factual fidelity (Zhang et al., 2024b; Yu et al., 2025; Shen et al., 2025), forming an interpretable and trustworthy evidence space.

3.2 Dual-view Four-Layer Taxonomy

The methodological perspectives of AIGC-V detection can be summarized as two complementary scientific views: visual and language. Out of this observation, as illustrated in Figure 2, we propose a **Vision-Language Dual-View** approach and organize existing methods into a four-layer methodological landscape from low-level perception to high-level cognition (§4).

Visual View. The visual view focuses on the visual modality and emphasizes the statistical differences between AIGC-V and real videos (Zhou et al., 2024; Gu et al., 2021). Its detection pathway extends from low-level intrinsic cue analysis at the frame level (§4.1) to spatiotemporal consistency across frames (§4.2), thereby forming the perceptual perspective of factual-fidelity verification.

Language View. The language view treats language as a grounded verification interface rather than restricting the view to spoken-language cues alone. It covers spoken-language signals (Bohacek and Farid, 2024; Cheng et al., 2022a), semantics internal to the video (Zhang et al., 2025f), and knowledge and facts about the world (Zhang et al., 2024b). Progressing from cross-modal consistency analysis (§4.3) to language-guided world-level reasoning at the factual and knowledge level (§4.4), it forms the cognitive perspective of factual-fidelity verification.

4 Landscape of Detection Methods

As illustrated in Figure 3, we organize AIGC-V detection methods into four layers, and provide a structured overview in this section.

4.1 Layer 1: Intrinsic Cues Analysis

Layer 1 asks whether low-level visual signals obey the statistical regularities of real videos, as well as

whether they match acquisition and post-processing effects (Le and Woo, 2023; Wang et al., 2022b), and whether AI generation or editing leaves intrinsic cues such as style patterns and model artifacts (Chen et al., 2021) or deviant physiological signals (Ciftci et al., 2020b; Qi et al., 2020; Ciftci et al., 2020a). Methods operate from the visual view by modeling, extracting, and amplifying these low-level signals (Vahdati et al., 2024).

Pixel and Geometric Artifacts. Video generation models introduce artifacts in frequency spectra and geometric structures. FreqBlender (Zhou et al., 2024) uses spectral blending and contrastive learning. Localized artifact detection enhances recognition of manipulated regions (Chen et al., 2021). These static-artifact methods are often low-cost and suitable for short clips, but remain susceptible to post-processing such as adaptive convolution and spectral regularization (Le and Woo, 2023; Wang et al., 2022b).

Physiological Features. Real videos display subtle yet stable physiological rhythms (heart rate, rPPG, blink) and identity-specific physiological signatures, which generative models struggle to replicate (Ciftci et al., 2020b; Qi et al., 2020; Ciftci et al., 2020a). Non-contact micro-expression analysis, notably blink pattern abnormalities, is also effective (Li et al., 2018). Other work analyzes spatial correlations and multi-region coupling in rPPG signals, and identity-specific muscle dynamics extend beyond simple blink cues (Wu et al., 2024; Stefanov et al., 2022; Mao and Yang, 2021; Cozzolino et al., 2021). These physiological methods depend on high frame-rate videos and unobstructed facial visibility, limiting effectiveness for short-duration or partially occluded scenarios.

Representation Distribution Discrepancy. Representation distribution discrepancy analysis models latent differences between real and fake videos to improve generalization. Latent space and style flow analysis reduce forgery specificity (Yan et al., 2024; Choi et al., 2024). Curriculum learning and soft discrepancy learning improve generalization (Lin et al., 2024b; Larue et al., 2023; Liu et al., 2024a). Test-time adaptation and quality-agnostic training stabilize robustness across shifts and compression (Chen et al., 2022; Le and Woo, 2023).

Overall, methods in the first layer detect by modeling, extracting, and amplifying low-level visual signals. Additional details and representative variants are summarized in Appendix C.1.

4.2 Layer 2: Spatiotemporal Consistency

Layer 2 asks whether spatiotemporal image flow satisfies the motion constraints of real videos. Real capture is constrained by continuous camera trajectories and physical scenes (Zhang et al., 2025e; Internò et al., 2025), so adjacent frames show continuous, predictable, physically feasible changes, while AIGC-V can exhibit long-range inconsistencies such as object or background distortion (Gu et al., 2022c; Zhang et al., 2024a) and sudden local blurring (Gu et al., 2022a). Spatiotemporal inconsistency learning (Gu et al., 2021) and motion-based detection (Ma et al., 2024b; Yin et al., 2023; Haliassos et al., 2021) exploit frame differences and flow residuals (Chen et al., 2025b).

Temporal and Motion Inconsistencies. Spatiotemporal methods model the evolution of facial content across frames and treat temporal inconsistency as the key signal (Gu et al., 2021, 2022c; Ge et al., 2021). Region-aware variants highlight facial regions where features change abnormally and can predict inconsistency maps for refinement (Gu et al., 2022c; Zhang et al., 2024a; Cai et al., 2025a). To reuse 2D backbones, some methods (Xu et al., 2023, 2024; Masi et al., 2020) reshape frame sequences into thumbnail layouts or use recurrent aggregation. To improve cross-generator transfer and interpretability, vulnerability-aware objectives (Nguyen et al., 2025), plug-and-play adapter tuning (Yan et al., 2025), and dynamic prototypes (Trinh et al., 2021) are explored.

Physical and Frequency Artifacts. Beyond raw RGB dynamics, many works design transformed descriptors that highlight spectral irregularities (Nie et al., 2024; Kim et al., 2025). Second-order or residual descriptors capture statistical mismatches (Zheng et al., 2025; Padhi et al., 2025; Xu et al., 2025d; Stamnas and Sanchez, 2025). These methods analyze temporal frequency responses to separate authentic motion spectra from synthetic components and improve robustness to post-processing (Kim et al., 2025). Forensic-oriented augmentation further emphasizes stable traces under perturbations (Corvi et al., 2025).

Human Behavioral and Interaction Dynamics. Some work moves from low-level artifacts to high-level human signals, asking whether expression (Anand et al., 2025), speech (Agarwal

and Farid, 2021), gaze (Kohler et al., 2025), identity (Dong et al., 2020), and 3D geometry (Yang et al., 2019; Tursman et al., 2020; Sun et al., 2021) behave like a real human, or whether cross-modal and interaction cues are broken. Given natural variation in behavior and recording setup, robustness can also benefit from self-supervised proxy tasks (Yang et al., 2025) and audiovisual anomaly objectives for video forensics (Feng et al., 2023).

Overall, methods in the second layer expose spatiotemporal inconsistency cues beyond frame-level artifacts. Additional details and representative variants are summarized in Appendix C.2.

4.3 Layer 3: Cross-Modal Consistency

Layer 3 focuses on whether the modalities in a video are well aligned in describing the same thing; real videos often accompany audio, text, and visuals that are highly aligned across modalities, whereas AIGC-V may present mismatches between lip movements and speech or between identity and voiceprint (Chugh et al., 2020; Yang et al., 2021; Zhang et al., 2021b; Cheng et al., 2022a). Mismatches can also appear between image content and associated text, motivating text-video consistency reasoning (Zhang et al., 2025f; Liu et al., 2025e). Symbolic and semantic lip-sync analyses further highlight this consistency space (Shahzad et al., 2022; Bohacek and Farid, 2024).

Audio-Visual Consistency Detection. Authentic content typically exhibits strong cross-modal consistency between audio and visual frames, whereas AIGC-V can break that alignment in subtle but revealing ways. Audio-visual consistency detection therefore seeks to identify fine-grained cross-modal inconsistencies in AIGC-V. Early work on low-level synchronization focuses on alignment between speech and mouth movements (Chugh et al., 2020; Yang et al., 2021). A more interpretable route maps audio and visual streams into symbolic sequences (Shahzad et al., 2022; Bohacek and Farid, 2024) and measures cross-modal discrepancies at the sequence level. Then, cross-modal identity consistency shifts the focus from "what is said" to "who is speaking" (Zhang et al., 2021b; Cheng et al., 2022a), using voiceprint-face mismatches as a clue.

Text-Video Semantic Consistency Reasoning. As AIGC-V becomes more semantically coherent (OpenAI, 2024a; Google DeepMind, 2025), detection methods that rely only on audio-visual artifacts become increasingly insufficient. Text-video semantic consistency reasoning evaluates whether visual content is aligned with associated textual information, often formulated as a Natural Language Inference problem where semantic inconsistency is the key signal (Zhang et al., 2025f; Liu et al., 2025e). CA-FVD (Zhang et al., 2025f) uses a VLM to generate pseudo-labels for video-text consistency, while CSCL (Liu et al., 2025e) unifies detection and grounding through cascaded decoders.

Robust Learning and Temporal Localization. Beyond modality-pair-specific cues, robustness to compression noise, language variations, and distribution shifts is a key concern. A significant trend is learning transferable representations from large-scale authentic data alongside cross-modal fusion (Liang et al., 2025; Zhou and Lim, 2021; Wang et al., 2022a), thereby improving robustness to compression and language shifts (Liang et al., 2025; Feng et al., 2023). Weakly supervised temporal localization frameworks further expand this line (Xu et al., 2025c; Wang et al., 2025b).

Overall, Layer 3 methods detect AIGC-V by measuring cross-modal inconsistencies across language and vision. Additional discussion and benchmark-specific details are provided in Appendix C.3.

4.4 Layer 4: Language-Guided World-Level Reasoning

Layer 4 elevates detection from internal consistency within the video to consistency with world-level rules and knowledge, and the research question shifts to whether the video content can truly exist in the real world and whether it is reasonable in semantic and factual dimensions. All content in real videos should be consistent with real-world facts, physical rules, and other domain knowledge and basic common sense. However, the content of AIGC-V is often difficult to fully align with the real world, which constitutes the detection space exploited by the fourth layer (Motamed et al., 2025b). Language serves as a natural interface that enables human-readable explanations, while baseline VLM prompting in zero- or few-shot settings provides a minimal approach to detecting inconsistencies and producing textual explanations (Jia et al., 2024; Shahzad et al., 2025).

Language-Calibrated Forgery Representations. Language-calibrated approaches inject textual priors such as class prompts or generated descriptions to reshape forgery-aware representations, often

without fully fine-tuning the backbone. The rPPG-based cues can be sensitive to post-processing and recording conditions (D'Amelio et al., 2023) and CPML (Lai et al., 2024) co-maps rPPG signals and facial landmark dynamics and aligns them with real/fake class prompts. RepDFD (Lin et al., 2025a) reprograms a pretrained VLM via learnable input perturbations and sample-level adaptive prompts. Knowledge-guided designs (Yu et al., 2025; Shen et al., 2025) further align visual features with textual embeddings or descriptions to improve generalization and interpretability.

Tool-Augmented Agentic Pipelines. Language-vision models can be embedded into agentic pipelines that make evidence gathering explicit. LAVID (Liu et al., 2025d) and FakeHunter (Chen et al., 2025a) frame detection as observe-tool-integrate pipelines that call external analyzers for evidence. Memory-Anchored reasoning (Chen et al., 2025a) and DAVID-XR1 (Gao et al., 2025) emphasize semantic anchors and interpretable evidence for explainable forensics. Related multi-agent settings are also reported (Zaman et al., 2025), but these pipelines improve accuracy while adding latency and orchestration overhead.

Post-Training for Explainable World-Level Reasoning. Post-training aligns world-level reasoning with explicit supervision or rewards so models internalize explainable behaviors. A common recipe is to start with supervised fine-tuning(SFT) on reasoning/explanation annotations to cold-start explainable behaviors, as in VidGuard-R1 (Park et al., 2025) and EDVD-LLaMA (Sun et al., 2025). Reinforcement learning can further align reasoning and improve generalization, e.g., VidGuard-R1 (Park et al., 2025) via DPO/GRPO with specialized reward models, and BusterX++ (Wen et al., 2025b) explores a cold-start-free RL post-training strategy. More explicitly, DeeptraceReward (Fu et al., 2025) constructs a perception-aligned benchmark with spatiotemporally grounded deepfake-trace annotations and fine-tunes a VLM as a reward model to improve "recognize-localize-explain" performance. Additionally, grounded artifact reasoning and modular explanation schemes can constrain explanations and mitigate hallucination (Li et al., 2025b; Chen et al., 2024b).

Overall, language-guided methods move beyond artifact matching by combining explicit evidence reasoning with world-level factual verification. Additional details are provided in Appendix C.4. Despite varying evaluation setups across layers, we also provide a compact layer-wise performance snapshot in Table 6 in Appendix C.5.

5 Evaluation of AIGC-V Detection

We provide a concise overview of a generic evaluation framework for AIGC-V detection, focusing on evaluation metrics under the dual-view setting and benchmarks organized in line with the three AIGC-V paradigms. We also highlight adjacent diagnostic resources for synthetic-video factual-fidelity evaluation, summarized in Appendix Table 8.

5.1 Evaluation Metrics

Shared Basic Metrics. Standard binary metrics remain the shared interface for real-versus-AIGC-V detection, including *Accuracy*, *AUC*, *Precision*, *Recall*, *F1*, *EER*, and, under class imbalance, *PR-AUC*. These metrics provide a necessary baseline, but they do not by themselves diagnose temporal coherence, physical plausibility, or semantic consistency. View-specific protocols are detailed in Appendix D.1.

Visual View Metrics. Metrics in the visual view emphasize robustness to intrinsic cues and distribution shifts, such as compression, codec, and resolution perturbations (Cheng et al., 2024; Zhou et al., 2024). Beyond in-domain *Acc*, *AUC*, and *EER*, it is common to report cross-dataset transfer and robustness under perturbation sweeps. In security-sensitive settings, fixed-operating-point measures such as $TPR@FPR{=}\alpha$ are also important. Visual evaluation also stresses video-level tests of spatiotemporal and physical consistency (Feng et al., 2023; Zhang et al., 2024a; Nie et al., 2024; Zhang et al., 2025e; Zheng et al., 2025; Kim et al., 2025). Accordingly, video-level aggregation and temporal perturbation or motion-ablation studies are used to verify whether temporal consistency modeling is actually being leveraged (Kundu et al., 2025; Yan et al., 2025).

Language View Metrics. Metrics in the language view stress cross-modal alignment via synchronization and retrieval measures (Katamneni and Rattani, 2024; Xu et al., 2025c), as well as factual-fidelity evaluation via question answering and human preference tests (Yu et al., 2025; Shen et al., 2025; Wen et al., 2025b). When the task includes temporal localization of mismatched segments, *Average Precision* is commonly reported

across multiple temporal *Intersection-over-Union* thresholds, together with the resulting *mAP* (Xu et al., 2025c; Anshul et al., 2025). Modality corruption further assesses reliance on cross-modal cues and robustness under degraded channels (Wang et al., 2025b). Under our factual-fidelity framing, explanation quality is often evaluated via *BLEU*, *ROUGE-L*, *METEOR*, *CIDEr*, and embedding-based semantic similarity (Sun et al., 2025; Gao et al., 2025; Hondru et al., 2025).

5.2 Benchmarks

Existing benchmarks are aligned with the three AIGC-V paradigms defined in Section 2: Local Manipulation Video (LMV), Audio-Visual Editing (AVE), and Generative Video Synthesis (GVS). Table 7 in the Appendix summarizes the broader benchmark landscape through March 2026, while Table 8 complements it with non-detector-first diagnostic resources.

LMV Related Benchmarks. LMV benchmarks preserve most of the source video and manipulate only local regions or attributes, so evaluation emphasizes subtle forensic traces, compression robustness, and cross-dataset transfer. Core resources include FaceForensics++ (Rossler et al., 2019), Celeb-DF (Li et al., 2020b), DFDC (Dolhansky et al., 2020), and DeeperForensics-1.0 (Jiang et al., 2020), with broader evaluations provided by ForgeryNet (He et al., 2021), DeepfakeBench (Yan et al., 2023), more recent explainability-oriented resources such as ExDDV (Hondru et al., 2025), and reasoning-oriented VLM benchmarks such as Beyond Static Artifacts (Gu et al., 2026).

AVE Related Benchmarks. AVE benchmarks focus on lip-audio alignment, speaker-content consistency, sensitivity to dubbing or splicing, robustness in real-world acoustic conditions, and the ability to localize narrative-level temporal manipulation. Representative resources include FakeAVCeleb (Khalid et al., 2021), LAV-DF (Cai et al., 2022), AV-Deepfake1M (Cai et al., 2024), multilingual or open-set extensions such as MAVOS-DD (Croitoru et al., 2025), and broader multimodal resources such as MMDF (Kim et al., 2026).

GVS Related Benchmarks. GVS benchmarks cover temporal consistency, cross-generator robustness, physical plausibility, and expanded scale and modality coverage (Ma et al., 2024a; Song et al., 2024b; Ji et al., 2024; Ni et al., 2025; Chen et al., 2024a; Ye et al., 2025). Beyond binary classification, benchmarks in this category increasingly incorporate richer supervision, including QA-style settings in LOKI (Ye et al., 2025), explanation-focused videos in Ivy-Fake (Jiang et al., 2025), defect-level spatiotemporal annotations in DAVID-X (Gao et al., 2025), and human-perceived fake traces in DeeptraceReward (Fu et al., 2025). The newest large-scale suites, including AIGVDBench (Ma et al., 2026), SynthForensics (Leotta et al., 2026), and MintVid (Tan et al., 2026), further emphasize generator diversity, stronger realism, and deployment-oriented stress testing.

Taken together, these benchmark families are structurally uneven: LMV has the deepest protocol history, AVE remains more costly because synchronized multimodal annotation is harder to obtain, and GVS changes fastest as generator turnover keeps resetting the task.

5.3 Evaluation and Future Trends

Recent work argues that classification metrics alone are not sufficiently evidential or interpretable for security-sensitive evaluation, and that strong clean-set scores can still mask brittleness under transfer-based attacks and newer synthesis pipelines (Chen et al., 2024b; Sun et al., 2025; Serrano et al., 2026; Hasan et al., 2026). Under our factual-fidelity framing, evaluation should go beyond binary decisions and incorporate evidence annotations, such as localized traces and rationale supervision, to make what is fake and where it is fake more traceable (Gao et al., 2025; Fu et al., 2025; Hondru et al., 2025). Robust evaluation should also emphasize generalization and avoid shortcut cues that do not reflect intrinsic inconsistencies (Cheng et al., 2024; Heo and Woo, 2025). Beyond detector-first benchmarks, adjacent diagnostic resources now span physical-rule stress tests such as VideoPhy-2 and Physion-Eval, world-dynamics probes such as StoryEval, VideoVerse, and T2VWorldBench, and explanation-oriented diagnosis such as SPOTLIGHT, VideoHallu, and PhyDetEx (Bansal et al., 2026; Zhang et al., 2026; Wang et al., 2024c, 2025f; Chen et al., 2025d; Chinchure et al., 2025; Li et al., 2025c; Wang et al., 2025g). Given fast-evolving generation models, a practical trend is a dynamic competitive arena with continuously updated evaluation sets and periodic regression under a unified protocol (Wang et al., 2025c). We provide an ex-

panded discussion of evaluation in Appendix D.3.

6 Challenges and Future

6.1 Unified Explainable Detection

Wang et al. (2025c) indicates that, for AIGC-V with sparse factual content but high perceptual fidelity and well-aligned modalities, VLMs can perform near random in perceptual-fidelity discrimination, far below humans. This highlights a core gap of language-only detection and the need for visual-view evidence. At the same time, stronger generators and anti-detection pipelines make it increasingly unsafe to assume that authenticity can be decided from perceptual traces alone. Commonsense reasoning (Zhang et al., 2024b), generalizable LVLM reasoning (Yu et al., 2025), language guidance (Shen et al., 2025), explainable video reasoning (Gao et al., 2025), and memory-anchored multimodal reasoning (Chen et al., 2025a) all point toward detectors that must reason over semantic and factual content as well.

Unified detection is therefore a two-pathway problem: perceptual evidence plus fact-level verification. In practice, this motivates cross-layer evidence fusion: frame-level intrinsic cues from Layer 1, sequence-level motion and physics coherence from Layer 2, multimodal verification from Layer 3, and external-knowledge reasoning from Layer 4 should be integrated into a unified evidence graph, where corroborating evidence boosts confidence and conflicts trigger abstention or human review. The goal is not to average heterogeneous scores, but to preserve evidential granularity so low-level traces, localized cross-modal conflicts, and claim-level checks remain individually inspectable. A practical system should output structured evidence objects, such as suspicious segments, localized cross-modal mismatches, and tested claims or entities (Chen et al., 2024b; Gao et al., 2025; Anshul et al., 2025), rather than only free-form rationales.

6.2 Evidence-First Trustworthy Detection

Trustworthy detection should follow an Evidence-First principle with an explicit reasoning path: Identify, Localize, and Explain (Chen et al., 2024b; Fu et al., 2025; Sun et al., 2025). Explanation should be downstream of evidence extraction, not a post-hoc gloss over a classifier score. This requires consistent evidence schemas (Gao et al., 2025; Chen et al., 2024b) together with calibrated uncertainty (Wang et al., 2024a), so conclusions remain traceable to inputs. Evaluation should therefore move beyond closed-set AUC toward claim-level grounding, shortcut-controlled stress tests, and dynamic benchmarks refreshed with new generators (Serrano et al., 2026; Hasan et al., 2026; Wang et al., 2025c).

Within this dual-pathway setup, each tool invocation or knowledge citation should map to a specific argumentation step (Liu et al., 2025d). Content-side analysis should also be cross-validated against source-side provenance and authentication signals when available, linking content analysis with "source-side" provenance tracing (Verdoliva, 2020; Deng et al., 2025; Coalition for Content Provenance and Authenticity, 2024; Tursman et al., 2020). This is especially important because detector outputs can otherwise become misleading priors in downstream claim verification without explicit evidence grounding (Sagar et al., 2026). Ultimately, trustworthy deployment depends on preserving both agreement and disagreement across evidence sources, rather than collapsing them too early into a single confidence score. We place the detailed discussion in the Appendix E.

7 Conclusion

This survey reframes AI-generated video detection as factual-fidelity verification and synthesizes prior work through a vision-language dual-view four-layer taxonomy with aligned benchmarks and metrics. Beyond summarizing methods, we connect these layers to deployment challenges, evaluation gaps, and emerging trends, and distill key requirements for trustworthy detection, including evidence-first and traceable decisions as well as robustness across generators and real-world scenarios. Future progress will depend on integrating perceptual forensics, multimodal reasoning, and provenance into a verifiable pipeline. We expect this survey to serve as a clear and actionable reference for future AIGC-V detection research, evaluation, and practice. More broadly, trustworthy AIGC-V detection should become a shared agenda for the computer vision, natural language processing, multimodal, and world model communities.

Limitations

In this survey, we acknowledge that several aspects may not be discussed in depth due to space limitations. First, our review is centered on AIGC-

V research, and complementary signals from the source side merit expanding our discussion. Second, substantial differences in evaluation protocols make systematic cross-layer comparisons infeasible. The field therefore calls for more unified and reproducible benchmarking/regression protocols. Despite these limitations, the proposed dual perspective and layered evidence space are expected to offer a precise reference for subsequent research in this critical yet still under-explored area.

Acknowledgments

This work was supported by funding from Mohamed bin Zayed University of Artificial Intelligence(MBZUAI).

References

Shruti Agarwal and Hany Farid. 2021. Detecting deepfake videos from aural and oral dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 981–989.

Shruti Agarwal, Hany Farid, Tarek El-Gaaly, and Ser-Nam Lim. 2020. Detecting deep-fake videos from appearance and behavior. In *2020 IEEE international workshop on information forensics and security (WIFS)*, pages 1–6. IEEE.

Tharun Anand, Siva Sankar, and Pravin Nair. 2025. Detecting localized deepfake manipulations using action unit-guided video representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4341–4351.

Ashutosh Anshul, Shreyas Gopal, Deepu Rajan, and Eng Siong Chng. 2025. Next-frame feature prediction for multimodal deepfake detection and temporal localization. *arXiv preprint arXiv:2511.10212*.

Marcella Astrid, Enjie Ghorbel, and Djamila Aouada. 2025. Audio-visual deepfake detection with local temporal inconsistencies. *Preprint*, arXiv:2501.08137.

Zechen Bai, Hai Ci, and Mike Zheng Shou. 2025. Impossible videos. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 2458–2483. PMLR.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. 2025. Videophy: Evaluating physical commonsense for video generation. In *The Thirteenth International Conference on Learning Representations*.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. 2026. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. In *The Fourteenth International Conference on Learning Representations*.

Arnesh Batra, Jashn Khemani, Arush Gumber, Anushk Kumar, Arhan Jain, and Somil Gupta. 2025. Socialdf: Benchmark dataset and detection model for mitigating harmful deepfake content on social media platforms. In *Proceedings of the 4th ACM International Workshop on Multimedia AI against Disinformation*, pages 81–89.

Matyas Bohacek and Hany Farid. 2024. Lost in translation: Lip-sync deepfake detection from audio-video mismatch. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4315–4323.

Gaetan Brison, Soobash Daiboo, Samy Aimeur, Awais Hussain Sani, Xi Wang, Gianni Franchi, and Vicky Kalogeiton. 2025. Fakeparts: a new family of ai-generated deepfakes. *arXiv preprint arXiv:2508.21052*.

ByteDance Seed. 2026. Official launch of seedance 2.0. `https://seed.bytedance.com/en/seedance2_0`. Accessed: 2026-02-24.

Yinqi Cai, Jichang Li, Zhaolun Li, Weikai Chen, Rushi Lan, Xi Xie, Xiaonan Luo, and Guanbin Li. 2025a. Deepshield: Fortifying deepfake video detection with local and global forgery analysis. *Preprint*, arXiv:2510.25237.

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

Zhixi Cai, Kartik Kuckreja, Shreya Ghosh, Akanksha Chuchra, Muhammad Haris Khan, Usman Tariq, Tom Gedeon, and Abhinav Dhall. 2025b. Av-deepfake1m++: A large-scale audio-visual deepfake benchmark with real-world perturbations. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 13686–13691.

Zhixi Cai, Kalin Stefanov, Abhinav Dhall, and Munawar Hayat. 2022. Do you really mean that? content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization. In *2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–10. IEEE.

Nuria Alina Chandra, Ryan Murtfeldt, Lin Qiu, Arnab Karmakar, Hannah Lee, Emmanuel Tanumihardja, Kevin Farhat, Ben Caffee, Sejin Paik, Changyeon Lee, and 1 others. 2025. Deepfake-eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. *arXiv preprint arXiv:2503.02857*.

Chen Chen and Dion Hoe-Lian Goh. 2026. Seeing, hearing, and knowing together: Multimodal strategies in deepfake videos detection. *CoRR*, abs/2602.01284.

Chen Chen, Runze Li, Zejun Zhang, Pukun Zhao, Fanqing Zhou, Longxiang Wang, and Haojian Huang. 2025a. Memory-anchored multimodal reasoning for explainable video forensics. *arXiv preprint arXiv:2508.14581*.

Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, and Huaxiong Li. 2024a. Demamba: Ai-generated video detection on million-scale genvideo benchmark. *arXiv preprint arXiv:2405.19707*.

Jiaxin Chen, Miao Hu, Dengyong Zhang, and Jingyang Meng. 2025b. Gc-consflow: Leveraging optical flow residuals and global context for robust deepfake detection. *arXiv preprint arXiv:2501.13435*.

Liang Chen, Yong Zhang, Yibing Song, Jue Wang, and Lingqiao Liu. 2022. Ost: Improving generalization of deepfake detection via one-shot test-time training. *Advances in Neural Information Processing Systems*, 35:24597–24610.

Peng Chen, Jin Liu, Tao Liang, Guangzhi Zhou, Hongchao Gao, Jiao Dai, and Jizhong Han. 2020. Fsspotter: Spotting face-swapped video by spatial and temporal clues. In *2020 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE Computer Society.

Weiliang Chen, Wenzhao Zheng, Yu Zheng, Lei Chen, Jie Zhou, Jiwen Lu, and Yueqi Duan. 2025c. Genworld: Towards detecting ai-generated real-world simulation videos. *Preprint*, arXiv:2506.10975.

Yize Chen, Zhiyuan Yan, Guangliang Cheng, Kangran Zhao, Siwei Lyu, and Baoyuan Wu. 2024b. X2-dfd: A framework for explainable and extendable deepfake detection. *Preprint*, arXiv:2410.06126.

Yubin Chen, Xuyang Guo, Zhenmei Shi, Zhao Song, and Jiahao Zhang. 2025d. T2vworldbench: A benchmark for evaluating world knowledge in text-to-video generation. *arXiv preprint arXiv:2507.18107*.

Zhikai Chen, Lingxi Xie, Shanmin Pang, Yong He, and Bo Zhang. 2021. Magdr: Mask-guided detection and reconstruction for defending deepfakes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9014–9023.

Harry Cheng, Yangyang Guo, Tianyi Wang, Qi Li, Xiaojun Chang, and Liqiang Nie. 2022a. Voice-face homogeneity tells deepfake. *Preprint*, arXiv:2203.02195.

Jikang Cheng, Zhiyuan Yan, Ying Zhang, Yuhao Luo, Zhongyuan Wang, and Chen Li. 2024. Can we leave deepfake data behind in training deepfake detector? *Advances in Neural Information Processing Systems*, 37:21979–21998.

Kun Cheng, Xiaodong Cun, Yong Zhang, Menghan Xia, Fei Yin, Mingrui Zhu, Xuan Wang, Jue Wang, and Nannan Wang. 2022b. Videoretalking: Audio-based lip synchronization for talking head video editing in the wild. In *SIGGRAPH Asia 2022 Conference Papers*.

Aditya Chinchure, Sahithya Ravi, Pushkar Shukla, Vered Shwartz, and Leonid Sigal. 2025. Spotlight: Identifying and localizing video generation errors using vlms. *arXiv preprint arXiv:2511.18102*.

Jongwook Choi, Taehoon Kim, Yonghyun Jeong, Seungryul Baek, and Jongwon Choi. 2024. Exploiting style latent flows for generalizing deepfake video detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1133–1143.

Komal Chugh, Parul Gupta, Abhinav Dhall, and Ramanathan Subramanian. 2020. Not made for each other- audio-visual dissonance-based deepfake detection and localization. In *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia*, MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia, pages 439–447, United States. Association for Computing Machinery (ACM). Publisher Copyright: © 2020 ACM.; 28th ACM International Conference on Multimedia, MM 2020, ACM MM ; Conference date: 12-10-2020 Through 16-10-2020.

Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. 2020a. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE transactions on pattern analysis and machine intelligence*.

Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. 2020b. How do the hearts of deep fakes beat? deep fake source detection via interpreting residuals with biological signals. In *2020 IEEE international joint conference on biometrics (IJCB)*, pages 1–10. IEEE.

Coalition for Content Provenance and Authenticity. 2024. C2pa technical specification. Accessed: 2026-01-01.

Davide Alessandro Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. 2022. Combining efficientnet and vision transformers for video deepfake detection. In *International conference on image analysis and processing*, pages 219–229. Springer.

Riccardo Corvi, Davide Cozzolino, Ekta Prashnani, Shalini De Mello, Koki Nagano, and Luisa Verdoliva. 2025. Seeing what matters: Generalizable ai-generated video detection with forensic-oriented augmentation. *arXiv preprint arXiv:2506.16802*.

Davide Cozzolino, Alessandro Pianese, Matthias Nießner, and Luisa Verdoliva. 2023. Audio-visual person-of-interest deepfake detection. *Preprint*, arXiv:2204.03083.

Davide Cozzolino, Justus Thies, Gernot Riegler, Oliver Wang, Matthias Niessner, and Luisa Verdoliva. 2021.

ID-Reveal: Identity-aware DeepFake Video Detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15088–15097. IEEE.

Florinel-Alin Croitoru, Andrei-Iulian Hîji, Vlad Hondru, Nicolae Cătălin Ristea, Paul Irofti, Marius Popescu, Cristian Rusu, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. 2024. Deepfake media generation and detection in the generative ai era: A survey and outlook. *arXiv preprint*.

Florinel-Alin Croitoru, Vlad Hondru, Marius Popescu, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. 2025. Mavos-dd: Multilingual audio-video open-set deepfake detection benchmark. *arXiv preprint arXiv:2505.11109*.

Alessandro D'Amelio, Raffaella Lanzarotti, Sabrina Patania, Giuliano Grossi, Vittorio Cuculo, Andrea Valota, and Giuseppe Boccignone. 2023. On using rppg signals for deepfake detection: A cautionary note. In *Image Analysis and Processing – ICIAP 2023*, volume 14234 of *Lecture Notes in Computer Science*, pages 235–246. Springer.

Soumyya Kanti Datta, Tanvi Ranga, Chengzhe Sun, and Siwei Lyu. 2025. Pia: Deepfake detection using phoneme-temporal and identity-dynamic analysis. *Preprint*, arXiv:2510.14241.

Jingyi Deng, Chenhao Lin, Zhengyu Zhao, Shuai Liu, Zhe Peng, Qian Wang, and Chao Shen. 2025. A survey of defenses against ai-generated visual media: Detection, disruption, and authentication. *ACM Computing Surveys*, 58(5):1–35.

Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. 2020. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*.

Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Dong Chen, Fang Wen, and Baining Guo. 2020. Identity-driven deepfake detection. *arXiv preprint arXiv:2012.03930*.

Mengnan Du, Shiva Pentyala, Yuening Li, and Xia Hu. 2020. Towards generalizable deepfake detection with locality-aware autoencoder. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 325–334.

Yuxuan Du, Zhendong Wang, Yuhao Luo, Caiyong Piao, Zhiyuan Yan, Hao Li, and Li Yuan. 2025. Cad: A general multimodal framework for video deepfake detection via cross-modal alignment and distillation. *Preprint*, arXiv:2505.15233.

Chao Feng, Ziyang Chen, and Andrew Owens. 2023. Self-supervised video forensics by audio-visual anomaly detection. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10491–10503.

Steven Lawrence Fernandes, Raghavender Reddy O, Satoshi Oishi, Niusha Vosoughi, Subrahmanyam Mupparaju, and Utkarsh Mittal. 2019. Predicting heart rate variations of deepfake videos using neural ode. In *2019 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1721–1729. IEEE.

Xingyu Fu, Siyi Liu, Yinuo Xu, Pan Lu, Guangqiuse Hu, Tianbo Yang, Taran Anantasagar, Christopher Shen, Yikai Mao, Yuanzhe Liu, Keyush Shah, Chung Un Lee, Yejin Choi, James Zou, Dan Roth, and Chris Callison-Burch. 2025. Learning human-perceived fakeness in ai-generated videos via multimodal llms. *arXiv preprint arXiv:2509.22646*. Project page: `https://deeptracereward.github.io/`.

Yifeng Gao, Yifan Ding, Hongyu Su, Juncheng Li, Yunhan Zhao, Lin Luo, Zixing Chen, Li Wang, Xin Wang, Yixu Wang, and 1 others. 2025. David-xr1: Detecting ai-generated videos with explainable reasoning. *arXiv preprint arXiv:2506.14827*.

Shiming Ge, Fanzhao Lin, Chenyu Li, Daichi Zhang, Jiyong Tan, Weiping Wang, and Dan Zeng. 2021. Latent pattern sensing: Deepfake video detection via predictive representation learning. In *MMAsia '21: ACM Multimedia Asia*, pages 6:1–6:7. ACM.

Google DeepMind. 2025. Veo 3. `https://aistudio.google.com/models/veo-3`. Accessed: 2025-12-15.

Jing Gu, Xin Wang, Hang Xu, Dong Wu, Yijin Chen, Xianghe Cong, Shaoyi Duan, Yuan-Fang Li, Chongyang Gao, Fukun Yin, Xun Li, Ziwei Liu, Jiaxing Huang, Yu Qiao, and Wei Gao. 2025. Phyworldbench: A physical realism benchmark for text-to-video generation. *arXiv preprint arXiv:2507.13428*.

Zheyuan Gu, Qingsong Zhao, Yusong Wang, Zhaohong Huang, Xinqi Li, Cheng Yuan, Jiaowei Shao, Chi Zhang, and Xuelong Li. 2026. Beyond static artifacts: A forensic benchmark for video deepfake reasoning in vision language models. In *2026 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Submitted for publication.

Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. 2021. Spatiotemporal inconsistency learning for deepfake video detection. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3473–3481.

Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, and Lizhuang Ma. 2022a. Delving into the local: Dynamic inconsistency learning for deepfake video detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 744–752.

Zhihao Gu, Taiping Yao, Yang Chen, Shouhong Ding, and Lizhuang Ma. 2022b. Hierarchical contrastive inconsistency learning for deepfake video detection. In *European conference on computer vision*, pages 596–613. Springer.

Zhihao Gu, Taiping Yao, Yang Chen, Ran Yi, Shouhong Ding, and Lizhuang Ma. 2022c. Region-aware temporal inconsistency learning for deepfake video detection. In *IJCAI*, volume 1, pages 920–926.

Xuyang Guo, Haoyu Zhou, Vivek Ramanujan, Chunhui Wang, Kang Li, Pengchuan Zhang, Kate Saenko, Kalyan Veeramachaneni, and X. Edward Zhou. 2025. T2vphysbench: A first-principles benchmark for physical consistency in text-to-video generation. *arXiv preprint arXiv:2505.00337*.

Fahima Hajjej, Muhammad Hamid, and Ala Saleh Alluhaidan. 2026. An integrated framework for proactive deepfake mitigation via attention-driven watermarking and blockchain-based authenticity verification. *Scientific Reports*. Early access article.

Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 2021. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5039–5049.

Md. Tarek Hasan, Sanjay Saha, Shaojing Fan, Swakkhar Shatabda, and Terence Sim. 2026. Deepfake synthesis vs. detection: An uneven contest. *CoRR*, abs/2602.07986.

Ammarah Hashmi, Sahibzada Adil Shahzad, Chia-Wen Lin, Yu Tsao, and Hsin-Min Wang. 2024. Understanding audiovisual deepfake detection: Techniques, challenges, human factors and perceptual insights. *arXiv preprint*.

Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu. 2021. Forgerynet: A versatile benchmark for comprehensive forgery analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4360–4369.

Minji Heo and Simon S Woo. 2025. Fakechain: Exposing shallow cues in multi-step deepfake detection. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 855–866.

Javier Hernandez-Ortega, Ruben Tolosana, Julian Fierrez, and Aythami Morales. 2020. Deepfakeson-phys: Deepfakes detection based on heart rate estimation. *arXiv preprint arXiv:2010.00400*.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. 2022a. Imagen video: High definition video generation with diffusion models. *Preprint*, arXiv:2210.02303.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. 2022b. Video diffusion models. *Preprint*, arXiv:2204.03458.

Vlad Hondru, Eduard Hogea, Darian Onchis, and Radu Tudor Ionescu. 2025. Exddv: A new dataset for explainable deepfake detection in video. *arXiv preprint arXiv:2503.14421*.

Fa-Ting Hong, Zunnan Xu, Zixiang Zhou, Jun Zhou, Xiu Li, Qin Lin, Qinglin Lu, and Dan Xu. 2025. Audio-visual controlled video diffusion with masked selective state spaces modeling for natural talking head generation. *Preprint*, arXiv:2504.02542.

Ziheng Hu, Hongtao Xie, Yuxin Wang, Jiahong Li, Zhongyuan Wang, and Yongdong Zhang. 2021. Dynamic inconsistency-aware deepfake video detection. In *IJCAI*, pages 736–742.

Christian Internò, Robert Geirhos, Markus Olhofer, Sunny Liu, Barbara Hammer, and David Klindt. 2025. Ai-generated video detection via perceptual straightening. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Lichuan Ji, Yingqi Lin, Zhenhua Huang, Yan Han, Xiaogang Xu, Jiafei Wu, Chong Wang, and Zhe Liu. 2024. Distinguish any fake videos: Unleashing the power of large-scale data and motion features. *arXiv preprint arXiv:2405.15343*.

Shan Jia, Reilin Lyu, Kangran Zhao, Yize Chen, Zhiyuan Yan, Yan Ju, Chuanbo Hu, Xin Li, Baoyuan Wu, and Siwei Lyu. 2024. Can chatgpt detect deepfakes? a study of using multimodal large language models for media forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4324–4333.

Changjiang Jiang, Wenhui Dong, Zhonghao Zhang, Chenyang Si, Fengchang Yu, Wei Peng, Xinbin Yuan, Yifei Bi, Ming Zhao, Zian Zhou, and 1 others. 2025. Ivy-fake: A unified explainable framework and benchmark for image and video aigc detection. *arXiv preprint arXiv:2506.00979*.

Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. 2020. 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 (CVPR)*.

Juho Jung, Sangyoun Lee, Jooeon Kang, and Yunjin Na. 2024. Www: Where, which and whatever enhancing interpretability in multimodal deepfake detection. *arXiv preprint arXiv:2408.02954*.

Vinaya Sree Katamneni and Ajita Rattani. 2024. Contextual cross-modal attention for audio-visual deepfake detection and localization. *Preprint*, arXiv:2408.01532.

Achhardeep Kaur, Azadeh Noori Hoshyar, Vidya Saikrishna, Selena Firmin, and Feng Xia. 2024. Deepfake video detection: challenges and opportunities. *Artificial Intelligence Review*, 57.

Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. 2021. Fakeavceleb: A novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*.

Taehoon Kim, Jongwook Choi, Yonghyun Jeong, Haeun Noh, Jaejun Yoo, Seungryul Baek, and Jongwon Choi. 2025. Beyond spatial frequency: Pixel-wise temporal frequency-based deepfake video detection. *arXiv preprint arXiv:2507.02398*.

Youngseo Kim, Kwan Yun, Seokhyeon Hong, Sihun Cha, Colette Suhjung Koo, and Junyong Noh. 2026. X-AVDT: Audio-visual cross-attention for robust deepfake detection. In *2026 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Accepted as conference paper.

Odin Kohler, Rahul Vijaykumar, and Masudul H Imtiaz. 2025. Deepfake detection in dyadic video calls using point of gaze tracking. *arXiv preprint arXiv:2509.25503*.

Christos Koutlis and Symeon Papadopoulos. 2025. Dimodif: Discourse modality-information differentiation for audio-visual deepfake detection and localization. *Preprint*, arXiv:2411.10193.

Kuaishou Technology. 2025. Kling o1: Unified multimodal video model. `https://klingai.com/`. Accessed: 2025-12-15.

Kartik Kuckreja, Parul Gupta, Injy Hamed, Thamar Solorio, Muhammad Haris Khan, and Abhinav Dhall. 2025. Tell me habibi, is it real or fake? *arXiv preprint arXiv:2505.22581*.

Ivan Kukanov and Jun Wah Ng. 2025. Klassify to verify: Audio-visual deepfake detection using ssl-based audio and handcrafted visual features. *Preprint*, arXiv:2508.07337.

Rohit Kundu, Hao Xiong, Vishal Mohanty, Athula Balachandran, and Amit K Roy-Chowdhury. 2025. Towards a universal synthetic video detector: From face or background manipulations to fully ai-generated content. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28050–28060.

Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, and Gyeongsu Chae. 2021. Kodf: A large-scale korean deepfake detection dataset. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10744–10753.

Ching-Yi Lai, Chiou ting Hsu, Chih-Chung Hsu, and Chia-Wen Lin. 2024. Prompt-guided multi-modal contrastive learning for cross-compression-rate deepfake detection. In *35th British Machine Vision Conference 2024, BMVC 2024, Glasgow, UK, November 25–28, 2024*. BMVA.

Nicolas Larue, Ngoc-Son Vu, Vitomir Struc, Peter Peer, and Vassilis Christophides. 2023. Seeable: Soft discrepancies and bounded contrastive learning for exposing deepfakes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21011–21021.

Binh M Le and Simon S Woo. 2023. Quality-agnostic deepfake detection with intra-model collaborative learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22378–22389.

Soo-Hyun Lee, Gyung-Eun Yun, Min Young Lim, and Youn Kyu Lee. 2021. A study on effective use of bpm information in deepfake detection. In *2021 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 425–427. IEEE.

Taegyeong Lee, Soyeong Kwon, and Taehwan Kim. 2024. Grid diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8734–8743.

Roberto Leotta, Salvatore Alfio Sambataro, Claudio Vittorio Ragaglia, Mirko Casu, Yuri Petralia, Francesco Guarnera, Luca Guarnera, and Sebastiano Battiato. 2026. Synthforensics: A multi-generator benchmark for detecting synthetic video deepfakes. *CoRR*, abs/2602.04939.

Chuqiao Li, Zhiwu Huang, Danda Pani Paudel, Yabin Wang, Mohamad Shahbazi, Xiaopeng Hong, and Luc Van Gool. 2023. A continual deepfake detection benchmark: Dataset, methods, and essentials. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1339–1349.

Jieyu Li, Xin Zhang, and Joey Tianyi Zhou. 2025a. Aegis: Authenticity evaluation benchmark for ai-generated video sequences. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pages 13346–13353, New York, NY, USA. Association for Computing Machinery.

Tianqi Li, Ruobing Zheng, Minghui Yang, Jingdong Chen, and Ming Yang. 2024a. Ditto: Motion-space diffusion for controllable realtime talking head synthesis. *Preprint*, arXiv:2411.19509.

Xiaodan Li, Yining Lang, Yuefeng Chen, Xiaofeng Mao, Yuan He, Shuhui Wang, Hui Xue, and Quan Lu. 2020a. Sharp multiple instance learning for deepfake video detection. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1864–1872.

Xiaolou Li, Zehua Liu, Chen Chen, Lantian Li, Li Guo, and Dong Wang. 2024b. Zero-shot fake video detection by audio-visual consistency. *Preprint*, arXiv:2406.07854.

Yifei Li, Wenzhao Zheng, Yanran Zhang, Runze Sun, Yu Zheng, Lei Chen, Jie Zhou, and Jiwen Lu. 2025b. Skyra: Ai-generated video detection via grounded artifact reasoning. *Preprint*, arXiv:2512.15693.

Yuezun Li, Ming-Ching Chang, and Siwei Lyu. 2018. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International workshop on information forensics and security (WIFS)*, pages 1–7. Ieee.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2020b. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3207–3216.

Zongxia Li, Xiyang Wu, Guangyao Shi, Yubin Qin, Hongyang Du, Tianyi Zhou, Dinesh Manocha, and Jordan Lee Boyd-Graber. 2025c. Videohallu: Evaluating and mitigating multi-modal hallucinations on synthetic video understanding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yachao Liang, Min Yu, Gang Li, Jianguo Jiang, Boquan Li, Feng Yu, Ning Zhang, Xiang Meng, and Weiqing Huang. 2025. Speechforensics: Audio-visual speech representation learning for face forgery detection. *Preprint*, arXiv:2508.09913.

Kaiqing Lin, Yuzhen Lin, Weixiang Li, Taiping Yao, and Bin Li. 2025a. Standing on the shoulders of giants: Reprogramming visual-language model for general deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5262–5270.

Li Lin, Neeraj Gupta, Yue Zhang, Hainan Ren, Chun-Hao Liu, Feng Ding, Xin Wang, Xin Li, Luisa Verdoliva, and Shu Hu. 2024a. Detecting multimedia generated by large ai models: A survey. *arXiv preprint*.

Li Lin, Santosh Santosh, Mingyang Wu, Xin Wang, and Shu Hu. 2025b. Ai-face: A million-scale demographically annotated ai-generated face dataset and fairness benchmark. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3503–3515.

Yuzhen Lin, Wentang Song, Bin Li, Yuezun Li, Jiangqun Ni, Han Chen, and Qiushi Li. 2024b. Fake it till you make it: Curricular dynamic forgery augmentations towards general deepfake detection. In *European Conference on Computer Vision*, pages 104–122.

Haoyu Liu, Chaoyu Gong, Mengke He, Jiate Li, Kai Han, and Siqiang Luo. 2025a. When deepfake detection meets graph neural network: a unified and lightweight learning framework. *arXiv preprint arXiv:2508.05526*.

Jiaxin Liu, Jia Wang, Saihui Hou, Min Ren, Huijia Wu, Long Ma, Renwang Pei, and Zhaofeng He. 2025b. Beyond face swapping: A diffusion-based digital human benchmark for multimodal deepfake detection. *arXiv preprint arXiv:2505.16512*.

Ping Liu, Qiqi Tao, and Joey Tianyi Zhou. 2025c. Evolving from single-modal to multi-modal facial deepfake detection: Progress and challenges. *arXiv preprint*.

Qingyuan Liu, Pengyuan Shi, Yun-Yun Tsai, Chengzhi Mao, and Junfeng Yang. 2024a. Turns out i'm not real: Towards robust detection of ai-generated videos. *arXiv preprint arXiv:2406.09601*.

Qingyuan Liu, Yun-Yun Tsai, Ruijian Zha, Victoria Li, Pengyuan Shi, Chengzhi Mao, and Junfeng Yang. 2025d. Lavid: An agentic lvlm framework for diffusion-generated video detection. *arXiv preprint arXiv:2502.14994*.

X. Liu and 1 others. 2025e. Unleashing the potential of consistency learning for detecting and grounding multi-modal media manipulation. *arXiv preprint arXiv:2506.05890*.

Yuhan Liu, Xiuying Chen, Xiaoqing Zhang, Xing Gao, Ji Zhang, and Rui Yan. 2024b. From skepticism to acceptance: simulating the attitude dynamics toward fake news. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 7886–7894.

Yuhan Liu, Yuxuan Liu, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. 2025f. The truth becomes clearer through debate! multi-agent systems with large language models unmask fake news. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 504–514.

Yuhan Liu, Zirui Song, Juntian Zhang, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. 2025g. The stepwise deception: Simulating the evolution from true news to fake news with llm agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26187–26203.

Luma AI. 2024. Dream machine. `https://lumalabs.ai/dream-machine`. Accessed: 2025-12-15.

Junxian Ma, Shiwen Wang, Jian Yang, Junyi Hu, Jian Liang, Guosheng Lin, Jingbo Chen, Kai Li, and Yu Meng. 2025. Sayanything: Audio-driven lip synchronization with conditional video diffusion. *Preprint*, arXiv:2502.11515.

Long Ma, Zihao Xue, Yan Wang, Zhiyuan Yan, Jin Xu, Xiaorui Jiang, Haiyang Yu, Yong Liao, and Zhen Bi. 2026. Your one-stop solution for ai-generated video detection. *CoRR*, abs/2601.11035.

Long Ma, Zhiyuan Yan, Qinglang Guo, Yong Liao, Haiyang Yu, and Pengyuan Zhou. 2024a. Detecting ai-generated video via frame consistency. *arXiv preprint arXiv:2402.02085*.

Long Ma, Jiajia Zhang, Hongping Deng, Ningyu Zhang, Yong Liao, and Haiyang Yu. 2024b. Decof: Generated video detection via frame consistency. *CoRR*.

Maoyu Mao and Jun Yang. 2021. Exposing deepfake with pixel-wise ar and ppg correlation from faint signals. *arXiv preprint arXiv:2110.15561*.

Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. 2020. Two-branch recurrent network for isolating deepfakes in videos. In *European conference on computer vision*, pages 667–684. Springer.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Quanfeng Lu, Wenqi Shao, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. 2025. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In *Proceedings of the 42nd International Conference on Machine Learning*.

Hui Miao, Yuanfang Guo, Zeming Liu, and Yunhong Wang. 2025. Multi-modal deepfake detection via multi-task audio-visual prompt learning. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Yisroel Mirsky and Wenke Lee. 2021. The creation and detection of deepfakes: A survey. *ACM Computing Surveys*, 54(1):1–41.

Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. Emotions don't lie: An audio-visual deepfake detection method using affective cues. *Preprint*, arXiv:2003.06711.

Saman Motamed, Minghao Chen, Luc Van Gool, and Iro Laina. 2025a. Travl: A recipe for making video-language models better judges of physics implausibility. *arXiv preprint arXiv:2510.07550*.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. 2025b. Do generative video models understand physical principles? *arXiv preprint arXiv:2501.09038*.

Sneha Muppalla, Shan Jia, and Siwei Lyu. 2023. Integrating audio-visual features for multimodal deepfake detection. *Preprint*, arXiv:2310.03827.

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

Dat Nguyen, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. 2025. Vulnerability-aware spatio-temporal learning for generalizable and interpretable deepfake video detection. *arXiv preprint arXiv:2501.01184*.

Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, and Cuong M. Nguyen. 2022. Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding*, 223:103525.

Hong-Hanh Nguyen-Le, Van-Tuan Tran, Dinh-Thuc Nguyen, and Nhien-An Le-Khac. 2025. Passive deepfake detection across multi-modalities: A comprehensive survey. *arXiv preprint*.

Zhenliang Ni, Qiangyu Yan, Mouxiao Huang, Tianning Yuan, Yehui Tang, Hailin Hu, Xinghao Chen, and Yunhe Wang. 2025. Genvidbench: A challenging benchmark for detecting ai-generated video. *arXiv preprint arXiv:2501.11340*.

Fan Nie, Jiangqun Ni, Jian Zhang, Bin Zhang, and Weizhe Zhang. 2024. Dip: diffusion learning of inconsistency pattern for general deepfake detection. *IEEE Transactions on Multimedia*.

Trevine Oorloff, Surya Koppisetti, Nicolò Bonettini, Divyaraj Solanki, Ben Colman, Yaser Yacoob, Ali Shahriyari, and Gaurav Bharaj. 2024. Avff: Audio-visual feature fusion for video deepfake detection. *Preprint*, arXiv:2406.02951.

OpenAI. 2024a. Sora 2. `https://sora.chatgpt.com`. Accessed: 2025-12-15.

OpenAI. 2024b. Video generation models as world simulators. `https://openai.com/index/video-generation-models-as-world-simulators/`. Accessed: 2025-12-29.

Sudev Kumar Padhi, Harshit Kumar, Umesh Kashyap, and Sk Subidh Ali. 2025. De-fake: Style based anomaly deepfake detection. *arXiv preprint arXiv:2507.03334*.

Kyoungjun Park, Yifan Yang, Juheon Yi, Shicheng Zheng, Yifei Shen, Dongqi Han, Caihua Shan, Muhammad Muaz, and Lili Qiu. 2025. Vidguard-r1: Ai-generated video detection and explanation via reasoning mllms and rl. *arXiv preprint arXiv:2510.02282*.

Gan Pei, Jiangning Zhang, Menghan Hu, Zhenyu Zhang, Chengjie Wang, Yunsheng Wu, Guangtao Zhai, Jian Yang, Chunhua Shen, and Dacheng Tao. 2024. Deepfake generation and detection: A benchmark and survey. *arXiv preprint*.

Wenshuo Peng, Gongxuan Wang, Tianmeng Yang, Chuanhao Li, Xiaojie Xu, Hui He, and Kaipeng Zhang. 2025. Svbench: Evaluation of video generation models on social reasoning. *arXiv preprint arXiv:2512.21507*.

Hua Qi, Qing Guo, Felix Juefei-Xu, Xiaofei Xie, Lei Ma, Wei Feng, Yang Liu, and Jianjun Zhao. 2020. Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms. In *Proceedings of the 28th ACM international conference on multimedia*, pages 4318–4327.

Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, Lei Bai, Wanli Ouyang, and Ruimao Zhang. 2025. Worldsimbench: Towards video generation models as world simulators. In *Forty-Second International Conference on Machine Learning*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. *Preprint*, arXiv:2112.10752.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11.

Runway Research. 2024. Introducing gen-3 alpha: A new frontier for video generation. https://runwayml.com/research/introducing-gen-3-alpha. Accessed: 2025-12-15.

A. S. M. Sharifuzzaman Sagar, Mohammed Bennamoun, Farid Boussaid, Naeha Sharif, Lian Xu, Shaaban Sahmoud, and Ali Kishk. 2026. Fact or fake? assessing the role of deepfake detectors in multimodal misinformation detection. *CoRR*, abs/2602.01854.

Adrian Serrano, Erwan Umlil, and Ronan Thomas. 2026. Deepfake detectors are DUMB: A benchmark to assess adversarial training robustness under transferability constraints. *CoRR*, abs/2601.05986.

Sahibzada Adil Shahzad, Ammarah Hashmi, Sarwar Khan, Yan-Tsung Peng, Yu Tsao, and Hsin-Min Wang. 2022. Lip sync matters: A novel multimodal forgery detector. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1885–1892.

Sahibzada Adil Shahzad, Ammarah Hashmi, Yan-Tsung Peng, Yu Tsao, and Hsin-Min Wang. 2025. How good is chatgpt at audiovisual deepfake detection: A comparative study of chatgpt, ai models and human perception. *APSIPA Transactions on Signal and Information Processing*, 14:e11.

Rui Shao, Tianxing Wu, Liqiang Nie, and Ziwei Liu. 2025. Deepfake-adapter: Dual-level adapter for deepfake detection. *International Journal of Computer Vision*, 133(6):3613–3628.

Guangyu Shen, Zhihua Li, Xiang Xu, Tianchen Zhao, Zheng Zhang, Dongsheng An, Zhuowen Tu, Yifan Xing, and Qin Zhang. 2025. Authguard: Generalizable deepfake detection via language guidance. *arXiv preprint arXiv:2506.04501*.

Danqing Shi, Lan Jiang, Katherine M. Collins, Shangzhe Wu, Ayush Tewari, and Miri Zilka. 2026. How do people watch AI-generated videos of physical scenes? *CoRR*, abs/2602.03374.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. 2022. Make-a-video: Text-to-video generation without text-video data. *Preprint*, arXiv:2209.14792.

Stefan Smeu, Dragos-Alexandru Boldisor, Dan Oneata, and Elisabeta Oneata. 2025. Circumventing shortcuts in audio-visual deepfake detection datasets with unsupervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18815–18825.

Wentang Song, Zhiyuan Yan, Yuzhen Lin, Taiping Yao, Changsheng Chen, Shen Chen, Yandan Zhao, Shouhong Ding, and Bin Li. 2024a. A quality-centric framework for generic deepfake detection. *arXiv preprint arXiv:2411.05335*.

Xiufeng Song, Xiao Guo, Jiache Zhang, Qirui Li, Lei Bai, Xiaoming Liu, Guangtao Zhai, and Xiaohong Liu. 2024b. On learning multi-modal forgery representation for diffusion generated video detection. In *Advances in Neural Information Processing Systems*, volume 37, pages 122054–122077. Curran Associates, Inc.

Sotirios Stamnas and Victor Sanchez. 2025. Difffake: Exposing deepfakes using differential anomaly detection. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 695–705.

Kalin Stefanov, Bhawna Paliwal, and Abhinav Dhall. 2022. Visual representations of physiological signals for fake video detection. *arXiv preprint arXiv:2207.08380*.

Haoran Sun, Chen Cai, Huiping Zhuang, Kong Aik Lee, Lap-Pui Chau, and Yi Wang. 2025. Edvd-llama: Explainable deepfake video detection via multimodal large language model reasoning. *arXiv preprint arXiv:2510.16442*.

Zekun Sun, Yujie Han, Zeyu Hua, Na Ruan, and Weijia Jia. 2021. Improving the efficiency and robustness of deepfakes detection through precise geometric features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3609–3618.

Hao Tan, Jun Lan, Senyuan Shi, Zichang Tan, Zijian Yu, Huijia Zhu, Weiqiang Wang, Jun Wan, and Zhen Lei. 2026. Videoveritas: AI-generated video detection via perception pretext reinforcement learning. *CoRR*, abs/2602.08828.

Hao Tan, Jun Lan, Zichang Tan, Ajian Liu, Chuanbiao Song, Senyuan Shi, Huijia Zhu, Weiqiang Wang, Jun Wan, and Zhen Lei. 2025. Veritas: Generalizable deepfake detection via pattern-aware reasoning. *arXiv preprint arXiv:2508.21048*.

Shahroz Tariq, Sangyup Lee, and Simon Woo. 2021. One detector to rule them all: Towards a general deepfake attack detection framework. In *Proceedings of the web conference 2021*, pages 3625–3637.

Jiahe Tian, Cai Yu, Xi Wang, Peng Chen, Zihao Xiao, Jiao Dai, Jizhong Han, and Yesheng Chai. 2024. Real appearance modeling for more general deepfake detection. In *European Conference on Computer Vision*, pages 402–419. Springer.

Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. 2020. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148.

Timothy Tong, David Anastasiu, and Yuhong Liu. 2025. Deepfake detection using spatiotemporal methods and vision-language models. *Proceedings of the KDD Undergraduate and Master's Consortium (KDD-UMC'25)*, 3.

Loc Trinh, Michael Tsang, Sirisha Rambhatla, and Yan Liu. 2021. Interpretable and trustworthy deepfake detection via dynamic prototypes. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1973–1983.

Eleanor Tursman, Marilyn George, Seny Kamara, and James Tompkin. 2020. Towards untrusted social video verification to combat deepfakes via face geometry consistency. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 654–655.

Danial Samadi Vahdati, Tai D Nguyen, Aref Azizpour, and Matthew C Stamm. 2024. Beyond deepfake images: Detecting ai-generated videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4397–4408.

Keerthi Veeramachaneni, Praveen Tirupattur, Amrit Singh Bedi, and Mubarak Shah. 2025. Leveraging pre-trained visual models for ai-generated video detection. *arXiv preprint arXiv:2507.13224*.

Luisa Verdoliva. 2020. Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):910–932.

Ganglai Wang, Peng Zhang, Lei Xie, Wei Huang, Yufei Zha, and Yanning Zhang. 2022a. An audio-visual attention based multimodal network for fake talking face videos detection. *Preprint*, arXiv:2203.05178.

H. Wang and 1 others. 2025a. T^3svfnd: Towards an evolving fake news detector for emergencies with test-time training on short video platforms. *arXiv preprint arXiv:2507.20286*.

Hanyi Wang, Zihan Liu, and Shilin Wang. 2023. Exploiting complementary dynamic incoherence for deepfake video detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(8):4027–4040.

Jian Wang, Baoyuan Wu, Li Liu, and Qingshan Liu. 2025b. Fauforensics: Boosting audio-visual deepfake detection with facial action units. *Preprint*, arXiv:2505.08294.

Jiaqi Wang, Weijia Wu, Yi Zhan, Rui Zhao, Ming Hu, James Cheng, Wei Liu, Philip Torr, and Kevin Qinghong Lin. 2025c. Video reality test: Can ai-generated asmr videos fool vlms and humans? *Preprint*, arXiv:2512.13281.

Runqi Wang, Yang Chen, Sijie Xu, Tianyao He, Wei Zhu, Dejia Song, Nemo Chen, Xu Tang, and Yao Hu. 2025d. Dynamicface: High-quality and consistent face swapping for image and video using composable 3d facial priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13438–13447.

Tianyi Wang and Kam Pui Chow. 2023. Noise based deepfake detection via multi-head relative-interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14548–14556.

Tianyi Wang, Xin Liao, Kam Pui Chow, Xiaodong Lin, and Yinglong Wang. 2024a. Deepfake detection: A comprehensive survey from the reliability perspective. *ACM Computing Surveys*.

Xueyu Wang, Jiajun Huang, Siqi Ma, Surya Nepal, and Chang Xu. 2022b. Deepfake disrupter: The detector of deepfake is my friend. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14920–14929.

Yaohua Wang, Siying Cui, Aixi Zhang, Wei-Long Zheng, Senzhang Wang, and 1 others. 2024b. Fuseanypart: Diffusion-driven facial parts swapping via multiple reference images. *Advances in Neural Information Processing Systems*, 37:80864–80884.

Yiping Wang, Xuehai He, Kuan Wang, Luyao Ma, Jianwei Yang, Shuohang Wang, Simon Shaolei Du, and Yelong Shen. 2024c. Is your world simulator a good story presenter? a consecutive events-based benchmark for future long video generation. *arXiv preprint arXiv:2412.16211*.

Yubo Wang, Juntian Zhang, Yichen Wu, Yankai Lin, Nils Lukas, and Yuhan Liu. 2026. Forest before trees: Latent superposition for efficient visual reasoning. *arXiv preprint arXiv:2601.06803*.

Yuxi Wang, Yikang Wang, Qishan Zhang, Hiromitsu Nishizaki, and Ming Li. 2025e. Vcapav: A video-caption based audio-visual deepfake detection dataset. In *Proc. Interspeech 2025*, pages 3908–3912.

Zeqing Wang, Xinyu Wei, Bairui Li, Zhen Guo, Jinrui Zhang, Hongyang Wei, Keze Wang, and Lei Zhang. 2025f. Videoverse: How far is your t2v generator from a world model? *arXiv preprint arXiv:2510.08398*.

Zeqing Wang, Xiaohui Yuan, Shiwei Ding, Debing Ye, Zizheng Chen, and Lin Chen. 2025g. Phydetex: A benchmark dataset and method for detecting and explaining physical plausibility in text-to-video models. *arXiv preprint arXiv:2512.01843*.

Haiquan Wen, Yiwei He, Zhenglin Huang, Tianxiao Li, Zihan Yu, Xingru Huang, Lu Qi, Baoyuan Wu, Xiangtai Li, and Guangliang Cheng. 2025a. Busterx: Mllm-powered ai-generated video forgery detection and explanation. *arXiv preprint arXiv:2505.12620*.

Haiquan Wen, Tianxiao Li, Zhenglin Huang, Yiwei He, and Guangliang Cheng. 2025b. Busterx++: Towards unified cross-modal ai-generated content detection and explanation with mllm. *arXiv preprint arXiv:2507.14632*.

Jiahui Wu, Yu Zhu, Xiaoben Jiang, Yatong Liu, and Jiajun Lin. 2024. Local attention and long-distance interaction of rppg for deepfake detection. *The Visual Computer*, 40(2):1083–1094.

Xuecheng Wu, Danlei Huang, Heli Sun, Xinyi Yin, Yifan Wang, Hao Wang, Jia Zhang, Fei Wang, Peihao Guo, Suyu Xing, Junxiao Xue, and Liang He. 2025. Hola: Enhancing audio-visual deepfake detection via hierarchical contextual aggregations and efficient pre-training. *Preprint*, arXiv:2507.22781.

Qiang Xu, Wenpeng Mu, Jianing Li, Tanfeng Sun, and Xinghao Jiang. 2025a. Advancements in ai-generated content forensics: A systematic literature review. *ACM Computing Surveys*, 58(3):1–36.

Wenbo Xu, Wei Lu, and Xiangyang Luo. 2025b. Weakly supervised multimodal temporal forgery localization via multitask learning. *Preprint*, arXiv:2508.02179.

Wenbo Xu, Junyan Wu, Wei Lu, Xiangyang Luo, and Qian Wang. 2025c. A multimodal deviation perceiving framework for weakly-supervised temporal forgery localization. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pages 11581–11589. ACM.

Ying Xu, Marius Pedersen, and Kiran Raja. 2025d. Vod: Learning volume of differences for video-based deepfake detection. *arXiv preprint arXiv:2503.07607*.

Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. 2023. Tall: Thumbnail layout for deepfake video detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22658–22668.

Yuting Xu, Jian Liang, Lijun Sheng, and Xiao-Yu Zhang. 2024. Learning spatiotemporal inconsistency via thumbnail layout for face deepfake detection. *International Journal of Computer Vision*, 132(12):5663–5680.

Zhiyuan Yan, Tao Liu, Mei-Lin Zhang, Wenbin Li, Jianmin Zhang, and Jun Lu. 2024. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8984–8994.

Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. 2023. Deepfakebench: A comprehensive benchmark of deepfake detection. In *Advances in Neural Information Processing Systems*, volume 36, pages 4534–4565. Curran Associates, Inc.

Zhiyuan Yan, Yandan Zhao, Shen Chen, Mingyi Guo, Xinghe Fu, Taiping Yao, Shouhong Ding, Yunsheng Wu, and Li Yuan. 2025. Generalizing deepfake video detection with plug-and-play: Video-level blending and spatiotemporal adapter tuning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12615–12625.

Chenzhao Yang, Jun Ma, Shilin Wang, and Alan Wee-Chung Liew. 2021. Preventing deepfake attacks on speaker authentication by dynamic lip movement analysis. *IEEE Transactions on Information Forensics and Security*, 16:1841–1854.

Qingyang Yang, Chuanxu Wang, Peng Liu, Zitai Jiang, and Jiajiong Li. 2025. Video anomaly detection via self-supervised and spatio-temporal proxy tasks learning. *Pattern Recognition*, 158:111021.

Xin Yang, Yuezun Li, and Siwei Lyu. 2019. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 8261–8265. IEEE.

Junyan Ye, Baichuan Zhou, Zilong Huang, Junan Zhang, Tianyi Bai, Hengrui Kang, Jun He, Honglin Lin, Zihao Wang, Tong Wu, and 1 others. 2025. Loki: A comprehensive synthetic data detection benchmark using large multimodal models. *ICLR*.

Zhenhui Ye, Ziyue Jiang, Yi Ren, Jinglin Liu, JinZheng He, and Zhou Zhao. 2023. Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. *Preprint*, arXiv:2301.13430.

Qilin Yin, Wei Lu, Bin Li, and Jiwu Huang. 2023. Dynamic difference learning with spatio–temporal correlation for deepfake video detection. *IEEE Transactions on Information Forensics and Security*, 18:4046–4058.

Peipeng Yu, Jianwei Fei, Hui Gao, Xuan Feng, Zhihua Xia, and Chip Hong Chang. 2025. Unlocking the capabilities of large vision-language models for generalizable and explainable deepfake detection. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 72925–72943. PMLR.

Sayeem Been Zaman, Wasimul Karim, Arefin Ittesafun Abian, Reem E. Mohamed, Md Rafiqul Islam, Asif Karim, and Sami Azam. 2025. Deepagent: A dual stream multi agent fusion for robust multimodal deepfake detection. *Preprint*, arXiv:2512.07351.

Chenyu Zhang, Xunyu Wang, Seunghoon Hur, Yung-Hsu Yang, Shih-Lun Wu, Fanjie Wu, and Junsong Yuan. 2025a. Morpheus: Benchmarking physical reasoning of video generative models with real physical experiments. *arXiv preprint arXiv:2504.02918*.

Daichi Zhang, Chenyu Li, Fanzhao Lin, Dan Zeng, and Shiming Ge. 2021a. Detecting deepfake videos with temporal dropout 3dcnn. In *IJCAI*, pages 1288–1294.

Daichi Zhang, Zihao Xiao, Shikun Li, Fanzhao Lin, Jianmin Li, and Shiming Ge. 2024a. Learning natural consistency representation for face forgery video detection. In *European Conference on Computer Vision*, pages 407–424. Springer.

David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. 2025b. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *International Journal of Computer Vision*, 133(4):1879–1893.

Jian Zhang, Jiangqun Ni, and Hao Xie. 2021b. Deepfake videos detection using self-supervised decoupling network. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

Juntian Zhang, Chuanqi Cheng, Yuhan Liu, Wei Liu, Jian Luan, and Rui Yan. 2025c. Weaving context across images: Improving vision-language models through focus-centric visual chains. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27782–27798.

Juntian Zhang, Song Jin, Chuanqi Cheng, Yuhan Liu, Yankai Lin, Xun Zhang, Yufei Zhang, Fei Jiang, Guojun Yin, Wei Lin, and 1 others. 2025d. Viper: Empowering the self-evolution of visual perception abilities in vision-language model. *arXiv preprint arXiv:2510.24285*.

Qin Zhang, Peiyu Jing, Hong-Xing Yu, Fangqiang Ding, Fan Nie, Weimin Wang, Yilun Du, James Zou, Jiajun Wu, and Bing Shuai. 2026. Physion-eval: Evaluating physical realism in generated video via human reasoning. *arXiv preprint arXiv:2603.19607*.

Shuhai Zhang, ZiHao Lian, Jiahao Yang, Daiyuan Li, Guoxuan Pang, Feng Liu, Bo Han, Shutao Li, and Mingkui Tan. 2025e. Physics-driven spatiotemporal modeling for ai-generated video detection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Y. Zhang and 1 others. 2025f. Consistency-aware fake videos detection on short video platforms. *arXiv preprint arXiv:2504.21495*.

Yue Zhang, Ben Colman, Xiao Guo, Ali Shahriyari, and Gaurav Bharaj. 2024b. Common sense reasoning for deepfake detection. In *European conference on computer vision*, pages 399–415. Springer.

Chende Zheng, Ruiqi Suo, Chenhao Lin, Zhengyu Zhao, Le Yang, Shuai Liu, Minghui Yang, Cong Wang, and Chao Shen. 2025. D3: Training-free ai-generated video detection using second-order features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12852–12862.

Jiaran Zhou, Yuezun Li, Baoyuan Wu, Bin Li, Junyu Dong, and 1 others. 2024. Freqblender: Enhancing deepfake detection by blending frequency knowledge. *Advances in Neural Information Processing Systems*, 37:44965–44988.

Yipin Zhou and Ser-Nam Lim. 2021. Joint audio-visual deepfake detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14780–14789.

Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. 2020. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, pages 2382–2390, New York, NY, USA. Association for Computing Machinery.

Heqing Zou, Meng Shen, Yuchen Hu, Chen Chen, Eng Siong Chng, and Deepu Rajan. 2024. Cross-modality and within-modality regularization for audio-visual deepfake detection. *Preprint*, arXiv:2401.05746.

Yueying Zou, Peipei Li, Zekun Li, Huaibo Huang, Xing Cui, Xuannan Liu, Chenghanyu Zhang, and Ran He. 2025. Survey on ai-generated media detection: From non-mllm to mllm. *arXiv preprint*.

A Detailed Related Work

Table 1 systematically compares representative surveys along dimensions of scope and framing, as well as their treatment of the language view, to delineate our position relative to prior work.

A.1 Scope and Framing

Video-targeted scope. Prior surveys take a face-manipulation or broad *deepfake* scope and therefore organize the field around generation pipelines, detector paradigms, and perceptual traces. This breadth is valuable for coverage, but when video is not the primary target it tends to leave video-specific threat models and analyses underdeveloped (e.g., temporal dynamics, long-horizon artifacts, and the diversity of video manipulation/editing/-generation types) can be discussed only briefly rather than surveyed systematically. For example, Tolosana et al. (2020), Mirsky and Lee (2021), and Nguyen et al. (2022) provide broad generation-detection overviews organized by generation pipelines and deep learning paradigms. From a media-forensics viewpoint, Verdoliva (2020) frames deepfakes as part of a larger integrity-verification pipeline (detection, localization, attribution, and authentication), bringing in forensic goals beyond binary classification.

Trustworthiness framing. Reliability- and defense-oriented surveys treat deployment risk and adversarial evolution as first-class concerns, emphasizing robustness, uncertainty, failure modes,

Survey	Scope & Framing			Language View		
	Video-Targeted Scope	Trustworthiness Framing	Factual-Fidelity Verification	Explainability & Localization	Semantic Cues	World-level Reasoning
Wang et al. (Wang et al., 2024a)	✗	✓	✗	✗	✗	✗
Xu et al. (Xu et al., 2025a)	✗	✓	✗	✗	✗	✗
Croitoru et al. (Croitoru et al., 2024)	✗	✓	✗	✗	✗	✗
Hashmi et al. (Hashmi et al., 2024)	✓	✗	✗	✗	✗	✗
Kaur et al. (Kaur et al., 2024)	✓	✓	✗	✗	✗	✗
Deng et al. (Deng et al., 2025)	✗	✓	✗	✓	✗	✗
Nguyen-Le et al. (Nguyen-Le et al., 2025)	✗	✓	✗	✓	✗	✗
Lin et al. (Lin et al., 2024a)	✗	✓	✗	✓	✓	✗
Zou et al. (Zou et al., 2025)	✗	✓	✗	✓	✓	✗
Ours	✓	✓	✓	✓	✓	✓

Table 1: Comparison of representative surveys and ours across (i) **Scope & Framing** (video-targeted scope; trustworthiness framing; factual-fidelity verification) and (ii) **Language View** (explainability/localization; semantic cues; world-level reasoning). ✓indicates substantial discussion; ✗indicates absent/brief mention.

and complementary defenses such as authentication and disruption signals (Wang et al., 2024a; Deng et al., 2025). Beyond the deepfake-only scope, Lin et al. (2024a) surveys detection for AI-generated multimedia across modalities and explicitly separates accuracy-driven methods from "beyond detection" goals such as generalizability, robustness, and interpretability, while Zou et al. (2025) highlights the shift from domain-specific detectors to VLMs-based general-purpose detection with explainability/localization.

Factual-fidelity verification. Across prior surveys, "verification" is typically discussed within broader media-forensics pipelines that encompass detection, localization, attribution, and authentication (Verdoliva, 2020), or under deployment reliability with emphasis on robustness, uncertainty, and countermeasures (Wang et al., 2024a; Deng et al., 2025). VLM-centric surveys bring in semantic cues and interpretability/localization (Lin et al., 2024a; Zou et al., 2025), but they often stop at recognizing/explaining manipulations and do not systematize *fact-level* checking of the claims implied by a video. As summarized in Table 1, this gap leaves verification at the fact-level in the prior survey landscape.

A.2 Language View

Explainability & localization. Recent surveys increasingly treat interpretability/localization as part of trustworthy deployment, with VLM-centric views and AIGC-defense/forensics surveys including dedicated discussion of explainability and localization (Lin et al., 2024a; Zou et al., 2025; Deng et al., 2025; Nguyen-Le et al., 2025).

Semantic cues. As deepfakes move from static artifacts to temporally coherent synthesis, surveys increasingly emphasize *video-level* evidence and multimodal cues beyond pixels. Kaur et al. (2024) highlights video-specific challenges and robustness under real-world degradations, while multimodal surveys center on audio-visual correspondence and identity/synchrony cues (Hashmi et al., 2024). Complementary to cross-modal cues, a large portion of the literature still treats video authenticity as a *visual-forensics* problem: exploiting intrinsic traces (e.g., frequency-domain irregularities, camera/codec footprints) and spatiotemporal inconsistencies (e.g., motion flicker, geometry/-physics violations) that emerge when generators fail to maintain coherent dynamics. Recent VLM-centric surveys (Lin et al., 2024a; Zou et al., 2025) further incorporate semantic cues as language-level evidence. Complementary multimodal surveys consolidate passive detection cues across modalities and systematize emerging AIGC forensics directions (Nguyen-Le et al., 2025; Xu et al., 2025a).

World-level reasoning. Explicit world-level reasoning about events and plausibility is still rarely treated as a first-class evidence pathway; most surveys stop short of using language-guided world knowledge as primary evidence for video authenticity. Table 1 highlights that language-guided world-level reasoning and, especially, fact-level factual-fidelity verification remain under-emphasized compared with semantic cues. Our survey fills this gap by adopting a video-targeted scope, foregrounding factual-fidelity verification, and organizing methods, evaluation, and benchmarks under a vision-language dual-view taxonomy.

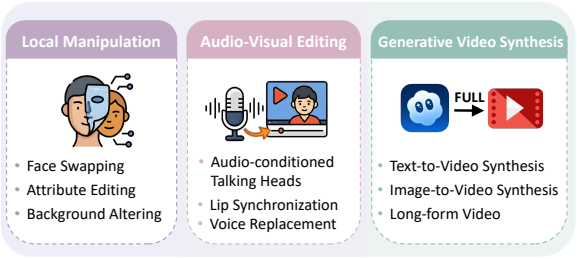

Figure 4: Overview of the three AIGC-V paradigms defined in this survey, highlighting their typical forms.

B Detailed AIGC-V Paradigms

This appendix complements the three AIGC-V paradigms defined in §2 by clarifying practical boundaries and representative generation pipelines. We use these clarifications when aligning datasets/benchmarks and discussing which evidence sources a detector is expected to rely on.

B.1 Local Manipulation

Beyond §2, we treat a sample as *Local Manipulation* when it preserves an authentic-video carrier (background, camera trajectory, and most pixels) and only re-renders *localized* regions/attributes (e.g., face parts, local objects, short temporal segments). This boundary is useful because forensic evidence is often spatially concentrated but can propagate temporally (e.g., blending halos, texture mismatch, or boundary inconsistencies).

Typical local-editing pipelines include "edit a part, keep the rest" designs such as FakeParts (Brison et al., 2025), and multi-stage pipelines that compose a local edit and then refine it with an additional generator (e.g., FakeChain (Heo and Woo, 2025)). For identity-centric attacks, DynamicFace (Wang et al., 2025d) combines 3D priors with temporal modeling for consistent swapping, while FuseAnyPart (Wang et al., 2024b) performs mask-guided latent fusion for fine-grained element editing.

B.2 Audio-Visual Editing

Audio-visual editing (§2.2) emphasizes manipulations where the *primary vulnerability* is cross-modal alignment (speech-mouth motion, speaker identity vs. voiceprint), often while keeping the scene/background largely intact. This type covers both (i) visually edited talking-head generation driven by audio and (ii) A/V mismatch via dubbing/splicing or timeline re-composition, where pixel-level artifacts can be weak but alignment cues are violated.

VideoReTalking (Cheng et al., 2022b) is a multi-stage "canonicalize-then-drive" paradigm that injects audio-driven motion followed by refinement, while 3D-aware routes such as GeneFace (Ye et al., 2023) explicitly model geometry/appearance to stabilize identity and motion under pose/view changes.

B.3 Generative Video Synthesis

Generative video synthesis (§2.3) refers to end-to-end generation without an authentic-video carrier: the model synthesizes the scene, camera motion, and dynamics jointly from text/image/noise conditions. Compared with LMV/AVE, evidence shifts from localized edit traces to global spatiotemporal coherence, physical plausibility, and provenance signals. Practical systems include large-scale text-to-video services such as Sora 2 (OpenAI, 2024a), Seedance 2.0 (ByteDance Seed, 2026), Veo 3 (Google DeepMind, 2025), Kling (Kuaishou Technology, 2025), Runway Gen-3 (Runway Research, 2024), and Dream Machine (Luma AI, 2024), which increasingly blur the line between "generation" and "editing" by supporting multi-condition control and iterative refinement.

C Detailed Method Landscape

We extend §4 by providing layer-wise details in §4.1–§4.4.

C.1 Layer 1: Intrinsic Cues

Table 2 summarizes representative Layer 1 methods. From the paradigm perspective, Layer 1 is most suitable for LMV, where localized edits on an authentic carrier often leave intrinsic residues (texture, frequency, geometry) around manipulated regions. For AVE and GVS, Layer 1 cues can still be helpful when synthesis and post-processing introduce stable fingerprints, but they are less reliable under strong compression and model adaptation.

Pixel and Geometric Artifacts. This line of work targets intrinsic traces introduced by AIGC-V synthesis and post-processing, including frequency fingerprints, texture residues, and geometric inconsistencies. FreqBlender amplifies spectral cues via spectral blending and contrastive learning (Zhou et al., 2024). Spectral cues relate to systematic frequency artifacts introduced by synthesis pipelines. Localized designs reduce reliance on global spectra by focusing on region-level residues, such as

Method	Cue	Input	Mechanism	Output	Date
A. Pixel and geometric artifacts					
Inconsistent Head Poses (Yang et al., 2019)	3D head-pose inconsistency	V	Head-pose estimation + geometry checks	Score	05/2019
MagDR (Chen et al., 2021)	Localized artifacts around manipulated regions	F+V	Mask-guided localization + reconstruction loss	Score+loc.	06/2021
DeepFake Disrupter (Wang et al., 2022b)	Detector-sensitive artifacts	F+V	Detector-in-the-loop optimization for perturbations	Protection	06/2022
HCIL (Gu et al., 2022b)	Region-level inconsistency	V	Hierarchical contrastive learning over facial regions	Score	10/2022
NoiseDF (Wang and Chow, 2023)	Forensic noise traces (face vs. bg)	F	Enhanced denoiser noise extraction + multi-head interaction	Score	06/2023
FreqBlender (Zhou et al., 2024)	Frequency-domain fingerprints	F	Spectral blending augmentation for frequency cues	Score	12/2024
B. Physiological features					
In Ictu Oculi (Li et al., 2018)	Eye-blink irregularities	V	Blink detection + temporal pattern modeling	Score	12/2018
FakeCatcher (Ciftci et al., 2020a)	PPG-like biological signal maps	V	Biological signal maps + detector	Score	07/2020
Hearts (Ciftci et al., 2020b)	Heart-related signals in residuals	V	Residual maps → rPPG features	Score	09/2020
DeepRhythm (Qi et al., 2020)	Visual heartbeat rhythms (rPPG)	V	Spatiotemporal attention over rPPG patterns	Score	10/2020
DeepFakesON-Phys (Hernandez-Ortega et al., 2020)	Heart-rate estimation inconsistency	V	rPPG heart-rate estimation + anomaly cues	Score	10/2020
Local rPPG Interaction (Wu et al., 2024)	Cross-region rPPG coupling	V	Local attention + long-range rPPG interaction	Score	02/2024
C. Distribution discrepancy and robustness					
OST (Chen et al., 2022)	Domain shift / compression sensitivity	F	One-shot test-time adaptation	Score	12/2022
SeeABLE (Larue et al., 2023)	Soft discrepancies under shifts	F	Real-only bounded contrastive learning	Score	10/2023
QAD (Le and Woo, 2023)	Compression-robust representations	F	Quality-aware regularization	Score	10/2023
LSDA (Yan et al., 2024)	Cross-generator transfer	V	Latent-space augmentation for transfer	Score	06/2024
Style Latent Flows (Choi et al., 2024)	Abnormal style-latent trajectories	V	Style-latent flow modeling + contrastive learning	Score	06/2024
Fake It Till You Make It (Lin et al., 2024b)	Curriculum-based forgery augmentation	V	Difficulty scheduling to improve generalization	Score	10/2024

Table 2: **Layer 1: Intrinsic cues analysis.** Representative methods grouped by cue family and ordered by time. *Cue* = evidence source; *Mechanism* = modeling design; Input: F=frames, V=video.

mask-guided localization in MagDR and texture-consistency learning in HCIL (Chen et al., 2021; Gu et al., 2022b). Geometric cues complement texture and frequency by testing structural plausibility, as instantiated by LRNet, RAM, and head-pose inconsistency tests (Sun et al., 2021; Tian et al., 2024; Yang et al., 2019). These cues are lightweight but degrade under compression and adaptive filtering, so quality-aware training and shift-aware settings are required in low-quality pipelines (Le and Woo, 2023; Wang et al., 2022b).

Physiological Features. Physiological cues treat faces as carriers of biological rhythms and micro-dynamics that are difficult to synthesize consistently across space and time. Heart-related signals are extracted either by mapping residual representations to heart-rate domains (Ciftci et al., 2020b) or by learning spatiotemporal attention over rPPG-related patterns (Qi et al., 2020). Higher-order heart rate variability features and cross-region coupling improve discriminative power and stability under noise (Hernandez-Ortega et al., 2020; Fernandes et al., 2019; Lee et al., 2021; Wu et al., 2024; Stefanov et al., 2022; Mao and Yang, 2021; Ciftci et al., 2020a). Complementary physiological evidence includes blink irregularities and identity-specific muscle dynamics (Li et al., 2018; Cozzolino et al., 2021). These cues require stable facial visibility and sufficient temporal support, and they remain sensitive to occlusion and aggressive re-encoding.

Representation Distribution Discrepancy. This line of work targets cross-generator transfer by modeling distribution gaps between authentic and synthetic videos rather than committing to a fixed artifact family. Latent and style-space perturbations diversify synthetic styles and reduce forgery specificity (Yan et al., 2024). Latent-flow modeling captures abnormal style trajectories (Choi et al., 2024). Curriculum learning and synthetic Blend-Fake data expose models to controlled difficulty and reduce shortcut learning (Lin et al., 2024b; Cheng et al., 2024). Discrepancy objectives separate overlapping real/fake manifolds with bounded contrastive learning and explicitly modeled appearance shifts (Larue et al., 2023; Liu et al., 2024a). At inference time, one-shot adaptation and quality-aware training improve stability under domain shifts and compression (Chen et al., 2022; Le and Woo, 2023), while locality-aware reconstruction cues provide region-level evidence for review (Du et al., 2020).

C.2 Layer 2: Spatiotemporal Consistency

Table 3 summarizes representative methods in this layer. Layer 2 provides a cross-paradigm bridge: it complements LMV by capturing temporal leakage beyond edited regions, and it is especially important for GVS where end-to-end synthesis must maintain coherent motion and physically plausible dynamics over long horizons. In AVE, Layer 2 often serves as auxiliary evidence (e.g., mouth/head motion realism), while the primary vulnerability typically lies in cross-modal alignment (Layer 3).

Temporal and Motion Inconsistencies. Layer 2 complements §4.2 by making explicit how temporal coherence is operationalized as a modeling prior. Instead of classifying frames independently, spatiotemporal methods model how facial appearance and geometry evolve across time and treat incoherence as evidence for forgery (Gu et al., 2021, 2022c). Clip encoders treat short sequences as space-time volumes and learn joint patterns with 3D CNNs or attention (Gu et al., 2021, 2022c; Hu et al., 2021; Zhang et al., 2024a; Chen et al., 2020; Zhang et al., 2021a). By jointly observing multiple frames, they reveal abnormal changes that are weak or ambiguous at the single-frame level. Region-aware variants highlight regions (e.g., eyes and mouth) where features change abnormally, and auxiliary objectives can predict inconsistency maps to force localized temporal evidence (Gu et al., 2022c; Hu et al., 2021; Zhang et al., 2024a). Temporal Dropout and rhythm perturbation further reduce overfitting to specific frame rates or speaking rhythms (Chen et al., 2020; Zhang et al., 2021a). To reuse 2D backbones, some methods reshape frame sequences into grid-like layouts to enable efficient 2D processing and leverage image pre-training (Xu et al., 2024, 2023). Hybrid pipelines aggregate per-frame features with recurrent branches to fuse static appearance and dynamic motion cues at the decision stage (Masi et al., 2020). Motion-centric designs construct intermediate signals such as frame differences or optical flow to emphasize temporal transitions (Ma et al., 2024b; Yin et al., 2023). Flow-residual cues test whether local facial motion is compatible with global head motion, remaining informative under blur and compression (Chen et al., 2025b; Wang et al., 2023). Overall, spatiotemporal inconsistency methods frame deepfake detection as coherence analysis rather than static appearance classification.

Human Behavioral and Interaction Dynamics. Behavioral cues move from artifact matching to higher-level human realism, asking whether expressions, gaze, and identity evolve plausibly over time. Action-unit guided and appearance-behavior joint representations model muscle activations and expression trajectories to expose abnormal combinations or unnatural dynamics (Anand et al., 2025; Agarwal et al., 2020). Gaze-based cues probe attention and interaction patterns that are difficult to reproduce consistently in face swapping or reenactment (Kohler et al., 2025). Identity-driven methods test whether temporal dynamics remain compatible with the claimed person under varying pose, illumination, and compression (Dong et al., 2020). Geometry-consistency checks enforce compatibility between 2D observations and 3D face/head motion, exposing pose- or shape-inconsistent sequences (Sun et al., 2021; Yang et al., 2019; Tursman et al., 2020). These semantics- and geometry-aware cues can be strong against visually polished fakes but are sensitive to natural behavioral diversity and recording conditions.

Physical and Frequency Artifacts. Transform-domain designs compute descriptors where non-physical dynamics become separable from content variation, yielding traces that can be more stable than raw RGB. Compared with purely spatial or spatiotemporal backbones trained on RGB, these descriptors can be more robust to appearance changes and sometimes more transferable across generators (Kim et al., 2025). Temporal frequency responses expose abnormal periodic components and energy distributions along time, helping separate authentic motion spectra from synthetic artifacts (Nie et al., 2024; Kim et al., 2025). Second-order statistics capture covariance mismatches and subtle distribution shifts in spatiotemporal features that survive mild post-processing (Zheng et al., 2025; Padhi et al., 2025). Residual- and difference-based features amplify deviations from real-video manifolds by emphasizing temporal changes rather than absolute appearance (Xu et al., 2025d; Stamnas and Sanchez, 2025). Physics-inspired priors provide complementary constraints on motion feasibility and perceptual plausibility, linking detection to camera trajectories and physical scenes (Zhang et al., 2025e; Internò et al., 2025).

Generalization and Robustness. Beyond cue design, robustness strategies aim to maintain performance under unseen generators and post-

Method	Cue	Mechanism	Output	Date
A. Temporal and motion inconsistencies				
FS-Spotter (Chen et al., 2020)	Spatiotemporal swap artifacts	Fuse spatial+temporal cues	Score	07/2020
Dynamic Prototypes (Trinh et al., 2021)	Transferable spatiotemporal evidence	Dynamic prototypes + predictive learning	Score	01/2021
Temporal Dropout 3DCNN (Zhang et al., 2021a)	Frame-rate dependent temporal artifacts	3D CNN + temporal dropout	Score	08/2021
STIL (Gu et al., 2021)	Short-clip spatiotemporal inconsistency	Clip encoder for ST inconsistency	Score	10/2021
EfficientNet-ViT Ensemble (Coccomini et al., 2022)	Robust temporal features	CNN/ViT ensemble	Score	05/2022
Region-Aware Temporal Inconsistency (Gu et al., 2022c)	Region-wise dynamics anomalies	Region-aware learning + maps	Score+loc.	07/2022
TALL (Xu et al., 2023)	Cross-frame inconsistency	Thumbnail layout + 2D backbone	Score	10/2023
Natural Consistency Representation (Zhang et al., 2024a)	Temporal naturalness deviations	SSL consistency + fine-tune	Score	11/2024
Vulnerability-Aware Learning (Nguyen et al., 2025)	Cross-generator vulnerability patterns	Vulnerability-aware transfer	Score	01/2025
GC-ConsFlow (Chen et al., 2025b)	Optical-flow residual inconsistency	Flow residuals + global context	Score	06/2025
Plug-and-Play Adapters (Yan et al., 2025)	Generalization under generator shift	Video blending + ST adapters	Score	06/2025
B. Physical and frequency artifacts				
DIP (Nie et al., 2024)	Temporal frequency response anomalies	Transform-domain temporal spectra	Score	12/2024
Perceptual Straightening (Internò et al., 2025)	Non-physical motion/trajectory irregularities	Perceptual straightening	Score	07/2025
Beyond RGB (Kim et al., 2025)	Stable descriptors beyond raw RGB	Transformed descriptors	Score	07/2025
Physics-Driven ST Modeling (Zhang et al., 2025e)	Physics constraints on spatiotemporal flow	Physics-driven ST modeling	Score	10/2025
C. Human behavioral and interaction dynamics				
Emotions Don't Lie (Mittal et al., 2020)	Affective cue coherence (A/V)	Siamese A/V + triplet loss	Score	03/2020
Appearance and Behavior (Agarwal et al., 2020)	Appearance/behavior realism over time	Behavior-aware temporal features	Score	12/2020
Identity-Driven Detection (Dong et al., 2020)	Identity dynamics consistency	Track identity dynamics	Score	12/2020
Gaze Tracking (Kohler et al., 2025)	Gaze dynamics	Gaze features for dyadic calls	Score	09/2025

Table 3: **Layer 2: Spatiotemporal consistency.** Representative methods grouped by (A) temporal/motion inconsistencies, (B) physical/frequency artifacts, and (C) human behavioral/interaction dynamics (§4.2). *Cue* = evidence source; *Mechanism* = modeling design. All methods take video as input.

processing by changing objectives, architectures, and training protocols (Nguyen et al., 2025; Yan et al., 2025; Trinh et al., 2021; Coccomini et al., 2022; Tong et al., 2025). Vulnerability-aware objectives and plug-and-play adaptation modules steer backbones toward transferable spatiotemporal evidence (Nguyen et al., 2025; Yan et al., 2025). Unified or cross-model detectors reuse pretrained backbones and expand the hypothesis space to handle diverse manipulations and media types (Trinh et al., 2021; Veeramachaneni et al., 2025; Coccomini et al., 2022; Liu et al., 2025a; Tariq et al., 2021). Forensic augmentation and quality-centric curricula expand coverage of perturbations and hard cases (Corvi et al., 2025; Song et al., 2024a). Self-supervised video forensics learns normal dynamics via synchronization or anomaly objectives and then treats deviations as manipulations (Yang et al., 2025; Feng et al., 2023). Multiple-instance formulations aggregate clip-level evidence to reduce reliance on exhaustive forged labels and handle mixed-quality videos (Li et al., 2020a).

C.3 Layer 3: Cross-Modal Consistency

Table 4 summarizes representative methods in this layer. Layer 3 is most aligned with AVE, since the attack surface of AVE is cross-modal synchronization and identity coherence (speech–lip, voice–face, and caption/video semantics). It also applies to LMV and GVS whenever audio or textual context (e.g., speech, ASR, captions, titles) is present and can be verified against visual evidence.

Audio-Visual Consistency Detection. Audio-visual consistency methods decompose evidence into synchrony, symbolic alignment, and identity coherence. Synchrony cues compare speech dynamics with mouth motion and learn cross-modal embeddings that penalize misalignment (Chugh et al., 2020; Yang et al., 2021). Cues extend beyond the mouth to additional regions whose micro-motions correlate with speech production and head dynamics (Agarwal and Farid, 2021). Symbolic alignment improves interpretability by mapping audio and visual streams to token or phoneme sequences and comparing them at the sequence level (Shahzad et al., 2022; Bohacek and Farid, 2024; Li et al., 2024b; Koutlis and Papadopoulos, 2025). Identity coherence shifts the question from "what is said" to "who is speaking" by testing whether voice and face embeddings belong to the same person, including person-of-interest settings trained with authentic talking-head data only (Zhang et al., 2021b; Cheng et al., 2022a; Cozzolino et al., 2023; Muppalla et al., 2023).

Method	Cue	Input	Mechanism	Output	Date
A. Audio-visual consistency detection					
Not Made for Each Other (Chugh et al., 2020)	Audio-visual dissonance	Speech+Face	Cross-modal mismatch + localization	Score+loc.	10/2020
Dynamic Lip Movement (Yang et al., 2021)	Lip motion vs. speech mismatch	Speech+Lip	Lip-motion dynamics analysis	Score	12/2020
Aural-Oral Dynamics (Agarwal and Farid, 2021)	Aural-oral dynamics coherence	Speech+Lip	Audio+mouth dynamics features	Score	06/2021
Joint Audio-Visual Detection (Zhou and Lim, 2021)	A/V synchrony and semantics	Speech+Face	Joint synchrony+semantic modeling	Score	10/2021
Audio-Visual Attention (Wang et al., 2022a)	Cross-modal alignment	Speech+Face	Cross-attention fusion	Score	03/2022
Lip Sync Matters (Shahzad et al., 2022)	Symbolic lip-sync mismatch	Speech+Lip	Token/phoneme sequence comparison	Score	11/2022
Voice-Face Homogeneity (Cheng et al., 2022a)	Voice-face identity coherence	Voice+Face-ID	Compare voice/face-ID embeddings	Score	11/2023
Lost in Translation (Bohacek and Farid, 2024)	Language-aware A/V mismatch	Speech+Lip	Language-conditioned lip-sync	Score	06/2024
AVFF (Oorloff et al., 2024)	Intrinsic A/V correspondences	Speech+Face	Real-only SSL pretrain + classifier	Score	06/2024
Multi-task A/V Prompt Learning (Miao et al., 2025)	Fine-grained A/V consistency	Speech+Face	Prompting + matching loss + fusion	Score	04/2025
CAD (Du et al., 2025)	Semantic A/V misalignment + cues	Speech+Face	Alignment + distillation	Score	05/2025
PIA (Datta et al., 2025)	Phoneme-timing mismatch + ID dynamics	Speech+Lip	Phoneme timing + ID dynamics	Score	10/2025
KLASSify to Verify (Kukanov and Ng, 2025)	Robust A/V cues (unseen attacks)	Speech+Face	Audio SSL + graph attn + handcrafted	Score+loc.	10/2025
B. Text-video semantic consistency reasoning					
CA-FVD (Zhang et al., 2025f)	Video-text semantic mismatch	Video+Text	VLM pseudo-labels + consistency loss	Score	04/2025
CSCL (Liu et al., 2025e)	Text-image inconsistency grounding	Frame+Text	Cascaded consistency decoders	Score+loc.	06/2025
T^{3}SVFND (Wang et al., 2025a)	Event-shifted semantics	Video+Text	Test-time MLM reconstruction	Score	07/2025
C. Robust learning and temporal localization					
Cross/Within-Modality Regularization (Zou et al., 2024)	Modality separation under perturb.	Speech+Face	A/V Transformer + cross/within regs	Score	04/2024
Audio-Visual Local Inconsistencies (Astrid et al., 2025)	Local temporal A/V inconsistency	Speech+Face	Local inconsistency + localization	Score+loc.	04/2025
Circumventing Shortcuts (Smeu et al., 2025)	Shortcut-robust A/V reps (silence)	Speech+Face	Real-only SSL A/V alignment	Score	06/2025
SpeechForensics (Liang et al., 2025)	Real-only A/V repr. learning	Speech+Lip	SSL pretrain on real + fine-tune	Score	08/2025
WMMT (Xu et al., 2025b)	Weakly-sup. temporal loc.	Speech+Face	Multitask + MoE + deviation loss	Score+loc.	08/2025
HOLA (Wu et al., 2025)	Hierarchical A/V aggregation	Speech+Face	SSL pretrain + gated aggregation	Score	10/2025
Weakly-Supervised Temporal Localization (Xu et al., 2025c)	Sparse multimodal deviations	Speech+Face	Weak supervision + deviation modeling	Score+loc.	10/2025

Table 4: **Layer 3: Cross-modal consistency.** Representative methods grouped by (A) audio-visual alignment, (B) text-video semantics, and (C) robust learning/localization.

Text-Video Semantic Consistency Reasoning. Text-video semantic consistency reasoning reframes detection as claim verification: given a video and its associated text (e.g., titles, captions, OCR/ASR transcripts), the model tests whether the textual statements are supported by visual evidence, with semantic contradictions serving as a key signal. CA-FVD supervises cross-modal consistency by prompting a VLM to obtain pairwise modality-consistency pseudo labels (visual-text-audio), then training with cosine-based consistency losses; co-attention fusion and collaborative diagnosis aggregate modality-level confidence into a final score (Zhang et al., 2025f). CSCL moves from global scoring to grounded evidence by building patch-token consistency matrices and using cascaded decoders to separately model within-modality context and cross-modality semantics; forgery-aware aggregation reduces confusing but locally consistent content and enables localization (Liu et al., 2025e). To handle event shifts, T^3SVFND introduces test-time training with an MLM-style reconstruction objective conditioned on multimodal context, suggesting event-robust text-video reasoning for detection settings (Wang et al., 2025a).

Robust Learning and Temporal Localization. In realistic settings, cross-modal inconsistencies are temporally sparse, so localization is needed for review and diagnosis. Two-stream fusion and attention-based alignment provide a common backbone for joint modeling and interpretation (Zhou and Lim, 2021; Wang et al., 2022a). Transferable representations learned from large-scale authentic data improve stability under compression and language shifts (Liang et al., 2025; Feng et al., 2023). Weak supervision with video-level labels and bias-aware objectives enables fine-grained temporal localization without dense annotations (Xu et al., 2025c; Astrid et al., 2025).

C.4 Layer 4: World-Level Reasoning

Table 5 summarizes representative methods in this layer. Layer 4 is most relevant for GVS and high-fidelity edits where low-level artifacts are weak: language-guided world-level reasoning tests semantic plausibility, commonsense/knowledge consistency, and causal/physical narratives. These checks can also support LMV and AVE in claim-centric settings by exposing contradictions that are not captured by purely perceptual cues.

Language-Calibrated Forgery Representations. Layer 4 complements §4.4 by specifying how language participates in evidence formation and calibration. Language-calibrated designs inject prompts or textual priors into multimodal representations to reshape forgery-sensitive features without rebuilding the backbone. CPML aligns physiological pulse (rPPG) signals and facial landmark dynamics with prompts and enforces cross-quality and cross-modal consistency to stabilize physiological evidence under compression (Lai et al., 2024; D'Amelio et al., 2023). RepDFD reprograms pre-trained vision-language models by freezing the backbone and learning input-side perturbations and adaptive prompts, while knowledge-guided variants build textual prototypes and uncertainty modeling for improved transfer (Lin et al., 2025a; Yu et al., 2025; Shen et al., 2025). Adapter-style reprogramming provides a lightweight alternative when full fine-tuning is infeasible (Shao et al., 2025).

Tool-Augmented Agentic Pipelines. Tool-augmented pipelines cast AIGC-V detection as explicit evidence gathering, where language guides what to inspect and which analyzer to call next. LAVID uses an observe-tool-integrate loop: it forms a coarse hypothesis, calls external tools, updates prompts based on intermediate outputs, and fuses multi-step evidence (Liu et al., 2025d). FakeHunter adds memory-anchored observe-think-act routines, including retrieval of semantic anchors and reuse of intermediate results across steps (Chen et al., 2025a). DAVID-XR1 and related pipelines emphasize semantic anchors and modular analyzers to support interpretable evidence composition (Gao et al., 2025). Multi-agent coordination further decomposes modality-specific evidence before fusion (Zaman et al., 2025).

Post-Training for Explainable World-Level Reasoning. Post-training internalizes evidence selection and reasoning style so that inference depends less on external controllers. Reinforcement learning and reward shaping encode preferences over how to gather evidence and how to structure explanations, as explored by VidGuard-R1, Veritas, and BusterX++ (Park et al., 2025; Tan et al., 2025; Wen et al., 2025b). Reward and evaluation modeling provides complementary supervision for explanation quality and localization granularity, as in DeeptraceReward (Fu et al., 2025). Related efforts extend post-training to multi-stage reasoning and modular explanation under complex AIGC-V settings (Wen

Method	Base Model / System	Training	Output	Date
A. Prompts/adapters for representation calibration				
ChatGPT Detect (Jia et al., 2024; Shahzad et al., 2025)	ChatGPT	Prompt	Verdict+exp.	06/2024
CPML (Lai et al., 2024; D'Amelio et al., 2023)	rPPG+landmark encoder	Prompt-guided	Score	11/2024
DeepFake-Adapter (Shao et al., 2025)	ViT (frozen)	Adapter tuning	Frame label (agg.)	01/2025
RepDFD (Lin et al., 2025a)	Frozen VLM	Prompt tuning	Label	04/2025
AuthGuard (Shen et al., 2025)	Vision encoder + LLM	Cls+ITC; uncertainty	Verdict+exp.	06/2025
LVLMDFD (Yu et al., 2025)	LVLM + LLM	Detector+prompt learner	Verdict+loc.+exp.	07/2025
B. Tool-augmented agents for evidence gathering				
LAVID (Liu et al., 2025d)	LVLM agent + tools	Tool loop	Verdict+evidence	02/2025
DAVID-XR1 (Gao et al., 2025)	Video VLM	SFT	Loc.+defect reasoning	06/2025
FakeHunter (Chen et al., 2025a)	Vision+audio encoders + VLM agent	Agent+retrieval	Verdict+evidence	08/2025
DeepAgent (Zaman et al., 2025)	Multi-agent pipeline	Not stated	MM verdict	12/2025
C. Post-training, preferences and rewards				
X2-DFD (Chen et al., 2024b)	LLaVA + aux detectors	SFT (explainable data)	Verdict+exp.	10/2024
BusterX (Wen et al., 2025a)	Qwen2.5-VL	SFT → RL	Verdict+exp.	05/2025
BusterX++ (Wen et al., 2025b)	Qwen2.5-VL	RL → SFT → RL	Verdict+struct. exp.	07/2025
VERITAS (Tan et al., 2025)	InternVL3 / Qwen2.5-VL	SFT → MiPO → P-GRPO	Verdict+exp.	08/2025
DeeptraceReward (Fu et al., 2025)	VideoLLaMA3 / Qwen2.5-VL (reward)	Reward-model train	Reward	09/2025
VidGuard-R1 (Park et al., 2025)	Qwen2.5-VL	SFT → DPO → GRPO	Verdict+exp.	10/2025
EDVD-LLaMA (Sun et al., 2025)	Qwen2.5-7B + video encoder	SFT	Verdict+exp.	10/2025
Skyra / Skyra-RL (Li et al., 2025b)	Qwen2.5-VL	SFT → RL	Verdict+grounded reasoning	12/2025

Table 5: **Layer 4: Language-guided world-level reasoning.** Representative methods grouped by (A) prompts/adapters, (B) tool-augmented agents, and (C) post-training with preferences/rewards. *Base Model / System* reports backbone names when available.

et al., 2025a; Sun et al., 2025; Li et al., 2025b; Chen et al., 2024b).

C.5 Layer-wise Performance Snapshot

To complement the qualitative taxonomy with a compact quantitative reference, Table 6 summarizes reported AUC (%) for representative methods across our four layers. Given inconsistent datasets and protocols across papers (as discussed in Limitations), these numbers are a best-effort snapshot, not a controlled comparison. For L1/L2/L4 rows, we follow the standard cross-dataset protocol used in deepfake detection: train on FaceForensics++ (FF++) and evaluate on Celeb-DF-v2 (CDFv2), DFDC Preview (DFDCP), and DFDC. Entries not reported in the original paper are marked as "–". † denotes an in-domain DFDC result reported for the audiovisual (Layer 3) method.

Method	Layer	CDFv2	DFDCP	DFDC
FreqBlender (Zhou et al., 2024)	L1	94.6	87.6	74.6
SeeABLE (Larue et al., 2023)	L1	87.3	86.3	75.9
LSDA (Yan et al., 2024)	L1	83.0	81.5	73.6
Style Latent Flows (Choi et al., 2024)	L1	89.0	–	–
TALL-Swin (Xu et al., 2023)	L2	90.8	–	76.8
LipForensics (Haliassos et al., 2021)	L2	82.4	–	73.5
Two-branch (Masi et al., 2020)	L2	76.6	–	–
MDS (Chugh et al., 2020)	L3	–	–	90.6†
RepDFD (Lin et al., 2025a)	L4	89.9	95.0	81.0
LVLMDFD (Yu et al., 2025)	L4	94.3	92.4	77.0

Table 6: AUC (%) snapshot across the four-layer taxonomy. Unless otherwise noted, results follow the cross-dataset protocol (train on FF++ and test on CDFv2/DFDCP/DFDC) as reported in the original papers; † reports an in-domain DFDC AUC for the audiovisual method. "–" denotes not reported.

D Detailed Evaluation

D.1 Evaluation Metrics

We expand the dual-view metrics summarized in Section 5.1, covering shared basic metrics, visual-view protocols, and language-view protocols.

Shared Basic Metrics. We report standard binary classification metrics for real vs. AIGC-V detection, including *Acc*, *AUC*, *Precision/Recall/F1*, and *EER* (optionally *PR-AUC* under heavy class imbalance). Unless stated otherwise, AIGC-V is treated as the positive class. Scores may be computed at the frame level or video level, with temporal aggregation (e.g., mean pooling, voting). These metrics provide a shared baseline but are insufficient to diagnose temporal coherence, physical plausibility, or semantic consistency on their own.

Visual View Metrics. The visual view evaluates whether a detector captures perceptual evidence that a video *looks and moves* like a real capture. This typically involves two intertwined aspects. **(i)**

Intrinsic-cue robustness: evaluation focuses on artifacts introduced during acquisition or generation (e.g., sensor/ISP traces, resampling and coding footprints, diffusion/AR sampling artifacts), often tested via cross-dataset generalization and robustness under compression or resolution perturbations. Performance is still commonly reported with *Acc*, *AUC*, and *EER* (Cheng et al., 2024; Zhou et al., 2024; Gu et al., 2022b), but the key is *how* they are reported: (a) *cross-dataset* scores (train on one source, test on another) and (b) *stress-test* scores under perturbation sweeps (e.g., codec/bitrate, downsampling), ideally including the average and worst-case *AUC/EER* across conditions. When detectors are deployed in security-sensitive settings, it is also common to report a fixed-operating-point metric such as *TPR@FPR*=α (true positive rate at a prescribed false positive rate). **(ii) Spatiotemporal and physical consistency:** evaluation targets whether motion, 3D layout, and interactions obey plausible temporal and physical constraints (Feng et al., 2023; Zhang et al., 2024a; Anand et al., 2025; Zhang et al., 2025e; Zheng et al., 2025; Agarwal and Farid, 2021; Kim et al., 2025; Xu et al., 2024). Here, video-level aggregation (e.g., *Video-Acc* or *Video-AUC*) is usually more informative than frame-level reporting (Kundu et al., 2025; Nie et al., 2024; Yan et al., 2025). To make "temporal reasoning" measurable rather than implicit, many works accompany *Video-AUC/EER* with *AUC/EER* drops under temporal perturbations (e.g., shuffling, re-timing, clip truncation), or sequence-level ablations that remove motion/interaction cues, and report the resulting performance change. Overall, strong frame-level scores can be misleading if temporal reasoning is weak; thus, visual-view evaluation should prioritize video-level analyses and robustness-to-perturbation protocols.

Language View Metrics. The language view evaluates whether the video, together with audio and accompanying text/metadata, supports a coherent and plausible *facts about the world*. It also naturally decomposes into two aspects. **(i) Cross-modal alignment:** metrics probe whether vision, audio, and text describe the same content, such as lip-audio synchrony, speaker identity consistency, and caption-video alignment. Evaluation typically combines basic classification metrics (e.g., *Acc/AUC*) with synchronization- or retrieval-oriented measures. When the output is a ranked list of candidate matches or suspicious segments, *Average Precision (AP)* (area under the *Precision-Recall* curve for the ranked predictions) and *Average Recall (AR)* (recall averaged over thresholds or over a fixed number of proposals) are commonly reported (Katamneni and Rattani, 2024; Xu et al., 2025c); in retrieval-style settings, *Recall@K* and related rank statistics are also used. Modality corruption experiments further reveal reliance on cross-modal cues and robustness under degraded channels (Wang et al., 2025b). When the task includes temporal localization of mismatched segments, it is common to report *AP* at multiple *Intersection-over-Union (IoU)* thresholds (where temporal *IoU* is the overlap duration divided by the union duration between predicted and ground-truth segments), as well as the mean *mAP* across thresholds (Anshul et al., 2025). **(ii) World knowledge and reasoning:** for fact-level and narrative-level verification, standard classification scores are often insufficient. Evaluation therefore, incorporates human judgments, pairwise preference tests, and question-answering style tasks (Yu et al., 2025; Shen et al., 2025; Wen et al., 2025b), as well as explanation-driven metrics to assess the quality and usefulness of rationales. Common automatic measures include *BLEU* (n-gram precision), *ROUGE-L* (LCS-based overlap), *METEOR* (alignment-based F-score), *CIDEr* (TF-IDF weighted n-gram consensus), and embedding-based similarity such as *Cosine Semantic Similarity (CSS)* (Sun et al., 2025; Gao et al., 2025; Fu et al., 2025; Hondru et al., 2025). Recent studies further show that observers rely on mixed visual, vocal, and knowledge cues, and that detector outputs can mislead downstream claim verification when they are not tied to explicit evidence (Chen and Goh, 2026; Shi et al., 2026; Sagar et al., 2026). In summary, language-view evaluation shifts the goal from "does it look real" to "does it state a plausible, well-grounded story" and requires metrics that explicitly score alignment, reasoning, and explanation.

D.2 Benchmarks

We further clarify the distinctions among the three paradigms of benchmarks in the following discussion and summarize a richer collection of benchmarks and datasets in Table 7.

LMV Related Benchmarks. Localized modification is the earliest and most established form of AIGC-V forgery, where only specific regions are replaced, altered, or enhanced while the rest of the

Benchmark & Dataset	Description	Paradigms	Date
FaceForensics++ (Rossler et al., 2019)	A forensics dataset consisting of 1000 original video sequences.	*LMV*	01/2019
Celeb-DF (Li et al., 2020b)	A large-scale challenging dataset for deepfake forensics.	*LMV*	09/2019
DeeperForensics-1.0 (Jiang et al., 2020)	A large-scale dataset for real-world face forgery detection.	*LMV*	05/2020
DFDC (Dolhansky et al., 2020)	A large-scale face-swapping videos, mainly involving local forgeries.	*LMV*	06/2020
WildDeepfake (Zi et al., 2020)	A real-world dataset for AIGC-V detection.	*LMV*	10/2020
ForgeryNet (He et al., 2021)	A mega-scale benchmark for both image- and video-level face forgery analysis.	*LMV*	06/2021
KoDF (Kwon et al., 2021)	A large-scale korean AIGC-V detection dataset.	*LMV*	10/2021
CDDB (Li et al., 2023)	Multi-level evaluations for easy, hard, and long-sequence AIGC-V detection.	*LMV*	01/2023
DF-Platter (Narayan et al., 2023)	Multi-face heterogeneous deepfake dataset.	*LMV*	06/2023
DeepfakeBench (Yan et al., 2023)	A comprehensive benchmark for AIGC-V detection.	*LMV*	07/2023
AI-Face (Lin et al., 2025b)	A million-scale demographically annotated AI-generated face dataset.	*LMV&GVS*	06/2024
DD-VQA (Zhang et al., 2024b)	AIGC-V detection VQA: Triplets of images, questions and answers.	*LMV*	09/2024
ExDDV (Hondru et al., 2025)	A dataset and benchmark for Explainable AIGC-V detection in Video.	*LMV*	11/2025
FAQ / Beyond Static Artifacts (Gu et al., 2026)	Forensic benchmark and instruction-tuning resource for temporal video deepfake reasoning in VLMs.	*LMV*	02/2026
FakeAVCeleb (Khalid et al., 2021)	A novel A/V multimodal deepfake dataset.	*AVE*	08/2021
LAV-DF (Cai et al., 2022)	Localized Audio Visual DeepFake: A content driven A/V dataset.	*AVE*	11/2022
FakeMix (Jung et al., 2024)	Clip-level benchmark detecting manipulated video and audio segments.	*AVE*	08/2024
AV-Deepfake1M (Cai et al., 2024)	A large-scale LLM-Driven A/V deepfake dataset.	*AVE*	10/2024
ArEnAV (Kuckreja et al., 2025)	An A/V deepfake dataset focuses specifically on Arabic-English CSW.	*AVE*	05/2025
MAVOS-DD (Croitoru et al., 2025)	A comprehensive multilingual open-set benchmark for A/V AIGC-V detection.	*AVE*	05/2025
DigiFakeAV (Liu et al., 2025b)	Large-scale benchmark for diffusion-based digital human A/V forgeries.	*AVE*	05/2025
SocialDF (Batra et al., 2025)	A dataset of 2,126 deepfake and real social media videos with sota manipulations.	*AVE*	07/2025
VCapAV (Wang et al., 2025e)	A video-caption based A/V AIGC-V detection dataset.	*AVE*	08/2025
X-AVFake (Chen et al., 2025a)	Dual-modality manipulations paired with grounded natural language reasoning.	*AVE*	08/2025
AV-Deepfake1M++ (Cai et al., 2025b)	2M clips extending A/V-Deepfake1M with diverse audio-visual manipulations.	*AVE*	10/2025
MMDF (Kim et al., 2026)	Multimodal deepfake dataset spanning GAN, diffusion, and flow-matching manipulations.	*AVE*	03/2026
GVF (Ma et al., 2024a)	Generated Video Forensics: evaluate AI generated video detectors.	*GVS*	02/2024
GenVidDet (Ji et al., 2024)	Real and AIGC-V from 8 generation models.	*GVS*	05/2024
GenVideo (Chen et al., 2024a)	An AIGC-V detection dataset.	*GVS*	05/2024
DVF (Song et al., 2024b)	Diffusion Video Forensics: a comprehensive diffusion video dataset.	*GVS*	12/2024
Physics-IQ (Motamed et al., 2025b)	Benchmarking physical understanding in generative video models.	*GVS*	01/2025
GenVidBench (Ni et al., 2025)	An AIGC-V detection dataset.	*GVS*	01/2025
IPV-Bench (Bai et al., 2025)	A well-structured benchmark comprising a diverse prompt suite and video dataset.	*GVS*	03/2025
Deepfake-Eval-2024 (Chandra et al., 2025)	A multi-modal in-the-wild benchmark of deepfakes circulated in 2024.	*L&A&G*	03/2025
LOKI (Ye et al., 2025)	A multimodal synthetic data detection benchmark (video, image, 3D, text, audio).	*GVS*	04/2025
GenBuster-200K (Wen et al., 2025a)	Two parts: real videos and synthetic videos that simulate real-world conditions.	*GVS*	05/2025
GenWorld (Chen et al., 2025c)	A large-scale real-world simulation dataset for AI-generated video detection.	*GVS*	06/2025
Ivy-Fake (Jiang et al., 2025)	A large-scale multimodal benchmark for explainable AIGC detection.	*GVS*	06/2025
DAVID-X (Gao et al., 2025)	AIGC-V with defect-level spatiotemporal annotations and rationales.	*GVS*	06/2025
GenBuster++ (Wen et al., 2025b)	A cross-modal benchmark for VLM evaluation.	*GVS*	07/2025
DeeptraceReward (Fu et al., 2025)	Spatiotemporal benchmark with human-perceived fake traces for generative video synthesis.	*GVS*	09/2025
ER-FF++set (Sun et al., 2025)	Comprehensive supervision for both detection and explanation.	*GVS*	10/2025
AEGIS (Li et al., 2025a)	Large-scale benchmark for sophisticated AIGC-V authenticity.	*GVS*	10/2025
ViFBench (Li et al., 2025b)	A benchmark comprising 3K generated by over ten SOTA generators.	*GVS*	12/2025
Video Reality Test (Wang et al., 2025c)	A ASMR-sourced benchmark.	*GVS*	12/2025
AIGVDBench (Ma et al., 2026)	Large-scale benchmark with 440K videos from 31 generation models and 33 evaluated detectors.	*GVS*	01/2026
SynthForensics (Leotta et al., 2026)	Human-centric synthetic-video benchmark with 6,815 videos from five open-source T2V generators.	*GVS*	02/2026
MintVid (Tan et al., 2026)	Lightweight high-quality benchmark with 3K videos from nine state-of-the-art generators.	*GVS*	02/2026

Table 7: Overview of existing datasets and benchmarks through March 2026, and their alignment with the AIGC-V paradigms introduced in our survey.

video remains intact. Accordingly, detection typically concentrates on microscopic traces around manipulated areas. Common cases include face swapping, lip-sync/face reenactment, video inpainting or removal, and localized object replacement. Representative datasets include FaceForensics++ (FF++) (Rossler et al., 2019), Celeb-DF (Li et al., 2020b), DFDC (Dolhansky et al., 2020), DeeperForensics (Jiang et al., 2020), and ForgeryNet (He et al., 2021). Despite extensive prior work, LMV detection remains challenging because (i) GAN-based patching and re-rendering can make local textures highly realistic and suppress traditional artifacts; (ii) compression at different levels may wash out subtle boundary (and depth-related) cues; and (iii) modern reenactment often applies cross-frame smoothing, reducing obvious temporal inconsistencies. Evaluation therefore emphasizes fine-grained artifact sensitivity (e.g., texture-level cues), robustness across compression levels and capture devices, and the ability to reveal subtle temporal anomalies. Recent work also reframes LMV evaluation as temporal deepfake reasoning for VLMs rather than only artifact capture, as in Beyond Static Artifacts (Gu et al., 2026).

AVE Related Benchmarks. Audio-visual inconsistency manipulation has become one of the fastest-growing multimodal forgery types, characterized by disrupting semantic or temporal alignment between real video and real or synthesized audio without modifying visual pixels. Typical forms include voice over or speech synthesis, lip sync manipulation, Audio or Video mismatch through reused or concatenated audio, clip-level timeline misalignment, and narrative-level recomposition. Existing datasets remain limited, with LAV-DF (Cai et al., 2022), FakeAVCeleb (Khalid et al., 2021), lip-sync deepfake datasets, and VCapAV (Wang et al., 2025e) A/V mismatch clips as the main resources. Detection is challenging because (a) pixel-level artifacts are absent, requiring explicit modeling of lip-phoneme synchronization; (b) diffusion/Transformer TTS produces highly realistic synthetic speech; (c) semantic consistency depends jointly on speech content, context, and speaker identity; and (d) robust frame-level alignment is needed, often via cross-modal attention. Evaluation therefore focuses on A/V synchronization, sensitivity to dubbing and splicing, robustness under real-world acoustic conditions, and the ability to detect narrative-level temporal manipulation. More recent resources such as X-AVFake and MMDF broaden grounded language reasoning and manipulation diversity in this branch (Chen et al., 2025a; Kim et al., 2026).

GVS Related Benchmarks. Fully synthetic videos produced by Satble Diffusion (Rombach et al., 2022), and T2V models (e.g., Sora2 (OpenAI, 2024a), Kling (Kuaishou Technology, 2025), Runway (Runway Research, 2024)) represent a rapidly emerging and highly challenging forgery type: videos are generated end-to-end with minimal pixel artifacts, realistic cinematography, and coherent physical motion. Benchmarks for this setting remain limited but are expanding, including GenVideo (Chen et al., 2024a), GenVidBench (Ni et al., 2025), GenWorld (Chen et al., 2025c), LOKI (Ye et al., 2025), DVF (Song et al., 2024b), and public text-to-video test suites released with commercial models. More recent large-scale resources such as AIGVDBench (Ma et al., 2026), SynthForensics (Leotta et al., 2026), and MintVid (Tan et al., 2026) further emphasize cross-generator transfer, stronger realism, and deployment-oriented stress testing. These datasets aim to evaluate cross-model generalization, physical and temporal consistency, and the detection of semantically fabricated events under large-scale open-world generation. AIGVDBench is particularly consequential because its scale makes cross-generator transfer patterns visible at ecosystem level.

Diagnostic Resources Beyond Detection Benchmarks. Adjacent diagnostic resources extend evaluation along a clear progression: from basic rule violations, to coherent world dynamics, and finally to explanation-oriented diagnosis. Physical-rule resources such as VideoPhy (Bansal et al., 2025), Physics-IQ (Motamed et al., 2025b), IPV-Bench (Bai et al., 2025), Morpheus (Zhang et al., 2025a), T2VPhysBench (Guo et al., 2025), PhyWorldBench (Gu et al., 2025), VideoPhy-2 (Bansal et al., 2026), and Physion-Eval (Zhang et al., 2026) ask whether generated videos obey basic physical constraints. World-dynamics resources such as WorldSimBench (Qin et al., 2025), Towards World Simulator (PhyGenBench) (Meng et al., 2025), StoryEval (Wang et al., 2024c), VideoVerse (Wang et al., 2025f), T2VWorldBench (Chen et al., 2025d), and SVBench (Peng et al., 2025) probe temporally coherent events and world knowledge over time. Explanation-oriented resources such as SPOTLIGHT (Chinchure et al., 2025),

VideoHallu (Li et al., 2025c), TRAVL (Motamed et al., 2025a), and PhyDetEx (Wang et al., 2025g) ask whether systems can localize and explain failures rather than only flag them. Table 8 summarizes this adjacent landscape.

D.3 Discussion of Evaluation

This subsection expands Section 5.3.

Recent work suggests that artifact-centric evaluation is increasingly misaligned with our *factual-fidelity* objective (§3.1): a high *AUC* on a closed set does not necessarily imply that a detector has evidence that the video violates real-world propositions. Under the dual-view, four-layer framing (§3), evaluation should therefore be *diagnostic*, probing which evidence pathway (visual vs. language) is used and at which layer factual fidelity breaks (intrinsic cues, spatiotemporal/physical consistency, cross-modal consistency, and world-level reasoning).

On the visual view side, improved diffusion and large video models suppress noticeable per-frame artifacts, making robustness-to-shifts and video-level consistency tests essential (Sun et al., 2025; Fu et al., 2025; Ni et al., 2025; Motamed et al., 2025b). Recent robustness studies further show that strong clean-set scores can mask severe brittleness under transfer-based attacks and modern synthesis pipelines (Serrano et al., 2026; Hasan et al., 2026). Beyond in-domain *AUC/EER*, protocols should emphasize cross-dataset generalization, perturbation sweeps (codec/bitrate, resolution), and fixed-operating-point reporting (e.g., $TPR@FPR{=}\alpha$) to reflect deployment risk.

On the language view side, evaluation must go beyond synchronization and treat the video as making checkable claims: cross-modal alignment metrics (sync/retrieval and temporal localization) should be complemented with fact- and knowledge-level verification (QA, preference tests) and rationale quality measures (Zhang et al., 2024b; Yu et al., 2025; Shen et al., 2025; Wen et al., 2025b). This aligns benchmarks with the shift from perceptual fidelity discrimination to factual-fidelity verification.

Finally, to make results auditable and to avoid shortcut learning, benchmarks increasingly add evidence supervision (tampered locations, spatiotemporal traces, grounded explanations) and explicitly control external cues such as watermarks, codec signatures, or metadata artifacts (Sun et al., 2025; Fu et al., 2025; Hondru et al., 2025; Gao et al., 2025; Wang et al., 2025c; Cheng et al., 2024; Heo and Woo, 2025). Where available, cross-validating content-side decisions with provenance/authentication signals can further improve trustworthiness in security-sensitive settings.

Beyond static test sets, we expect evaluation to move toward a closed loop between a *real-time detection platform* and a *dynamic benchmark*. In an arena/leaderboard-style regime, the benchmark is continuously refreshed by streaming outputs from newly released generators and by modeling platform ingestion/transcoding pipelines (e.g., codec and resolution changes); detectors are then periodically re-evaluated under a unified protocol to obtain regression trajectories rather than one-off scores. Hard cases that repeatedly fool strong detectors can then be promoted into subsequent refreshes with failure-type annotations. This requires coupling evaluation with deployment in a traceable, reproducible pipeline, so that performance remains comparable as detectors become public and adversaries iterate.

E Detailed Challenges

This appendix complements §6 with concise, implementation-oriented details for the evaluation and trustworthiness directions discussed there.

E.1 Robust Diagnostic Evaluation

Chen et al. (2024b) and Sun et al. (2025) argue that classification-centric metrics such as clip-level *AUC/EER* are often not sufficiently evidential or interpretable for security-sensitive settings. Under our *factual-fidelity* objective in §3.1, strong closed-set discrimination does not necessarily mean the detector can substantiate which real-world proposition is violated, where the violation occurs, or why the judgment is reliable. Evaluation should therefore be *diagnostic*: it should probe the detector's reliance on complementary pathways (perceptual or temporal cues versus claim-level semantic verification), and reveal the failure modes that lead to factual-fidelity breakdown.

One critical axis is robustness of perceptual and temporal evidence under distribution shift. As diffusion and large video models increasingly suppress noticeable per-frame artifacts, evaluation must stress-test video-level consistency and robustness to shift rather than only in-domain performance. This pressure is already visible in explainable detection settings such as EDVD-

Work and Resource	Description	Type	Date
A. Physical Rule Violations			
VideoPhy (Bansal et al., 2025)	Benchmarks whether generated videos satisfy physical commonsense about objects, attributes, and interactions.	*Eval*	06/2024
Physics-IQ (Motamed et al., 2025b)	Probes whether text-to-video models obey basic physical principles under controlled prompts.	*Eval*	01/2025
IPV-Bench (Impossible Videos) (Bai et al., 2025)	Designs impossible scenarios across physical, geographical, biological, and social domains for stress testing generated videos.	*Eval*	03/2025
Morpheus (Zhang et al., 2025a)	Benchmarks physical reasoning of video generative models with real physical experiments and measurable conservation-law violations.	*Eval*	04/2025
T2VPhysBench (Guo et al., 2025)	Tests first-principles physical consistency in text-to-video generation, including counterfactual robustness and state-transition fidelity.	*Eval*	05/2025
PhyWorldBench (Gu et al., 2025)	Evaluates physical realism of text-to-video models with physics-grounded prompts and anti-physics stress cases.	*Eval*	07/2025
VideoPhy-2 (Bansal et al., 2026)	Extends physical-commonsense evaluation to more challenging action-centric and interaction-heavy generated videos.	*Eval*	01/2026
Physion-Eval (Zhang et al., 2026)	Provides expert reasoning traces, localized physical glitches, and explanations for physical-realism failures in generated videos.	*Eval*	03/2026
B. World Dynamics and Causality			
WorldSimBench (Qin et al., 2025)	Evaluates whether video generators behave like world simulators by combining perceptual quality and embodied action consistency.	*Eval*	10/2024
Towards World Simulator (PhyGenBench) (Meng et al., 2025)	Crafts a physical-commonsense benchmark for video generation with multi-step dynamics and simulator-style diagnostics.	*Eval*	10/2024
StoryEval (Wang et al., 2024c)	Benchmarks whether T2V models can present short stories composed of consecutive events for future long-video generation.	*Eval*	12/2024
T2VWorldBench (Chen et al., 2025d)	Evaluates world-knowledge generation across physics, nature, activity, culture, causality, and object domains.	*Eval*	07/2025
VideoVerse (Wang et al., 2025f)	World-model-oriented benchmark with hidden semantics, event-level temporal causality, and world-knowledge questions.	*Eval*	10/2025
SVBench (Peng et al., 2025)	Evaluates social reasoning in video generation, including intention, joint attention, norms, and prosocial behavior.	*Eval*	12/2025
C. Explanation-Oriented Diagnosis			
TRAVL (Motamed et al., 2025a)	Adds intra-frame spatial and trajectory-guided temporal attention and fine-tunes on ImplausiBench to improve VLM judgments of physically implausible videos.	*Method*	10/2025
SPOTLIGHT (Chinchure et al., 2025)	Fine-grained identification and localization of generation errors using VLMs.	*Eval*	11/2025
VideoHallu (Li et al., 2025c)	Evaluates multimodal hallucination on synthetic videos and studies mitigation via targeted post-training.	*Eval*	12/2025
PhyDetEx (Wang et al., 2025g)	Introduces a benchmark and trains models to detect and explain violated physical rules in text-to-video outputs.	*Method*	12/2025

Table 8: Adjacent factual-fidelity diagnostic resources relevant to AIGC-V detector design and world-model-oriented evaluation.

LLaMA (Sun et al., 2025) and DeeptraceReward (Fu et al., 2025), in generator-diverse benchmarks such as GenVidBench (Ni et al., 2025), and even in physics-oriented generative evaluation such as Physics-IQ (Motamed et al., 2025b). Serrano et al. (2026) and Hasan et al. (2026) further show that strong clean-set scores can mask severe brittleness under transfer-based adversarial attacks and newer synthesis pipelines, leaving a gap between benchmark discrimination and deployment robustness. This motivates cross-dataset generalization protocols, systematic perturbation sweeps over codec, bitrate, and resolution, and deployment-oriented reporting at fixed operating points such as $TPR@FPR{=}\alpha$.

E.2 Claim-Level and Dynamic Evaluation

A natural next step is to make benchmarks *evidence-first* by moving from clip-level labels to claim-level supervision. In this setting, each clip is decomposed into a compact set of propositions specifying who, when, where, and what happens, and each proposition is paired with timestamped in-video evidence. Evaluation then scores both claim correctness and evidence grounding, making "what is fake" and "where it is fake" traceable rather than implicit. This direction is consistent with recent explainable AIGC detection resources that augment binary supervision with rationales, traces, and defect-level annotations (Chen et al., 2024b; Sun et al., 2025; Jiang et al., 2025; Gao et al., 2025; Fu et al., 2025; Hondru et al., 2025).

To prevent shortcut learning and isolate genuine reasoning, evidence-first benchmarks should also incorporate targeted stress tests. One complementary approach is claim-level stress testing for event logic: construct counterfactual or adversarial cases by rewriting event scripts or state transitions while controlling visual-quality confounds, so performance differences primarily reflect relational and event reasoning rather than superficial appearance cues (Cheng et al., 2024; Heo and Woo, 2025). Benchmarks increasingly add tampered locations, spatiotemporal traces, and grounded explanations, while high-fidelity resources such as Video Reality Test expose cases where obvious artifacts are weak (Sun et al., 2025; Fu et al., 2025; Hondru et al., 2025; Gao et al., 2025; Wang et al., 2025c). Given fast-evolving generators, we likewise expect evaluation to move beyond static test sets toward a closed loop between a real-time detection platform and a dynamic benchmark refreshed with newly released models and realistic ingestion effects (Wang et al., 2025c).

E.3 Unified Explainable Detection

When perceptual fidelity is high and modalities align, language-view cues may still be insufficient (Wang et al., 2025c), and the decisive evidence can come from subtle visual statistics (Zhou et al., 2024) or long-range temporal inconsistency (Gu et al., 2022a; Zhang et al., 2024a; Yan et al., 2025). When generators suppress artifacts, visual-view cues alone can also be insufficient if the video is visually coherent yet factually implausible (Zhang et al., 2024b; Yu et al., 2025; Shen et al., 2025; Gao et al., 2025). Unified detection is therefore a *two-pathway* problem: (i) perceptual evidence and (ii) fact-level verification.

The key challenge is to connect low-level cues to high-level conclusions in an auditable way (Chen et al., 2024b; Sun et al., 2025). A unified detector should output structured evidence (Chen et al., 2024b; Gao et al., 2025) such as suspicious segments (Anshul et al., 2025), entities, cross-modal mismatches (Katamneni and Rattani, 2024), and the claims being tested (Zhang et al., 2024b), rather than only free-form rationales (Sun et al., 2025). Tool-augmented pipelines (Liu et al., 2025d) help only when each tool invocation is tied to a specific sub-claim and produces reusable evidence objects (Chen et al., 2025a).

E.4 Evidence-First Trustworthy Detection

Trustworthy detection needs an explicit reasoning path (Chen et al., 2024b; Fu et al., 2025; Sun et al., 2025) that separates (i) identifying candidate issues, (ii) localizing where and when they occur, and (iii) explaining why they violate perceptual, temporal, cross-modal, or factual constraints. This requires consistent evidence formats (Gao et al., 2025; Chen et al., 2024b) and calibrated uncertainty (Wang et al., 2024a), so conclusions remain traceable to inputs. Calibration matters because trustworthy deployment depends on knowing when evidence is weak, conflicting, or incomplete rather than forcing a binary answer.

Content-side analysis should be cross-validated against source-side signals (Verdoliva, 2020; Deng et al., 2025; Coalition for Content Provenance and Authenticity, 2024; Hajjej et al., 2026) when available, rather than treated as unrelated defenses. This matters beyond standalone detection accuracy: detector outputs can become misleading priors when

they are injected into downstream claim verification without explicit evidence grounding (Sagar et al., 2026). The practical challenge is to reconcile conflicts and present them within a unified evidence space that supports auditing and abstention under uncertainty (Wang et al., 2024a).